UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN
INGENIERÍA DE REDES Y SERVICIOS TELEMÁTICOS

TRABAJO FIN DE MÁSTER

COMPARATIVE ANALYSIS OF CURRENT TRENDS FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

LENIN SANTIAGO RIVADENERIA PUETATE

2020

TRABAJO DE FIN DE MÁSTER

Título:	Análisis Comparativo De Las Tendencias Actuales De Inteligencia Artificial Explicable.
Título (inglés):	Comparative Analysis Of Current Trends For Explainable Artificial Intelligence .
Autor:	Lenin Santiago Rivadeneira Puetate
Tutor:	Álvaro Carrera Barroso
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	——
Vocal:	——
Secretario:	——
Suplente:	——

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO DE FIN DE MÁSTER

Comparative Analysis Of Current Trends For Explainable Artificial Intelligence .

Julio 2020

Resumen

El estudio que se realizó muestra algunas de las diversas técnicas que se están utilizando en la actualidad para explicar cómo se comportan y como toman las decisiones los diferentes modelos de Inteligencia Artificial (IA) o Machine Learning (ML), haciendo uso de algunos de estos modelos, ya su vez comparando su rendimiento. Hoy se puede encontrar inteligencia artificial en muchos aspectos de nuestra vida diaria. Desde el aspecto más simple, por decir un ejemplo el negocio de la venta de cosméticos o ya sea en algo más relevante como la ingeniería, la medicina, las leyes, la electrónica, la administración de las empresas, en fin, en todos los músculos que el ser humano cree relevante, es por esto que eventualmente se necesitará que los sistemas impulsados por IA sean lo suficientemente seguros, confiables, accesibles y que sobre todo cumplan con las expectativas que la sociedad demanda, y sobre sobre todo se tenga un entendimiento relativo del porque es que un modelo hace determinada acción o toma determinada decisión. Para ello en este trabajo primero, se han elegido dos conjuntos de datos, del depósito de aprendizaje automático de UCI, que son frecuentes usados por toda la comunidad que se dedica al aprendizaje automático, como segundo paso una vez que se eligió con que conjuntos de datos trabajar, tratar los datos de los mismos, se procedió a utilizar varias técnicas de IA / ML de tal manera que se pudo ver como uno variaba muchas veces con respecto a otro y otras veces muy poco entre ellos, obteniendo resultados interesantes que se más adelante, en la tercera parte del trabajo nos sirve como punto de partida para poder comenzar con el proceso de explicación de cada modelo de IA, para esto lo que se hizo fue aplicar varias técnicas que nos ayude con estas explicaciones, primero se verá explicaciones que por su naturaleza son compatibles con cada uno de los modelos que se implementan, siendo técnicas bastante gráficas de como un elemento por lo general “el más relevante” tiene mucho que ver con la toma de decisión de cada modelo, y luego se encuentran otras técnicas por separado para ciertos modelos con los que mantienen una mejor compatibilidad con otras técnicas de igual manera se explica como el modo lo tomo alguna decisión o hizo un discernimiento.

Palabras clave: Dataset, Modelos, Inteligencia Artificial, Machine Learning, Técnicas, Explicación

Abstract

The study that was carried out shows some of the various techniques that are currently being used to explain how the different models of Artificial Intelligence (AI) or Machine Learning (ML) behave and how decisions are made, making use of some of these models, and in turn comparing their performance. Today artificial intelligence can be found in many aspects of our daily life. From the simplest aspect, to say an example, the business of selling cosmetics or whether it is in something more relevant such as engineering, medicine, law, electronics, business administration, in short, in all muscles that humans believe is relevant, which is why it will eventually be necessary for AI-powered systems to be sufficiently safe, reliable, accessible and above all to meet the expectations that society demands, and above all to have a relative understanding of why a model does a certain action or makes a certain decision. To do this, in this work first, two data sets have been chosen, from the UCI machine learning repository, which is frequently used by the entire machine learning community, As a second step, once it was chosen with which data sets to work, to treat the data from them, we proceeded to use various AI / ML techniques in such a way that it could be seen how one varied many times with respect to another and others. Sometimes very little between them, obtaining interesting results that are later, in the third part of the work it serves as a starting point to start with the process of explaining each AI model, for this what was done was to apply various techniques to help us with these explanations, first, we will see explanations that by their nature are compatible with each of the models that are implemented, being quite graphic techniques of how an element in general "the most relevant" has a lot to do with the making of decision of each model, and then other techniques are found separately for certain models with which they maintain better compatibility with other techniques, in the same way, it is explained as the way I made a decision or made a discernment

Keywords: Dataset, Models, Artificial Intelligence, Machine Learning, Techniques, Explanation

Agradecimientos

Antes que nada quiero agradecer primero a Dios que me ha dado la oportunidad de llegar hasta esta instancia. ademas quiero exaltar a mi padre, que ha sabido ser más que un padre, un amigo, dándome lecciones de vida, consejos y ánimos en todo momento. He sido firme, pero justo, gracias a su sacrificio estoy aquí, mis hermanas y sobrinas que sin su amor incondicional han sido más duras aún. Y agradecer sobre manera a mi bella madre que con su ternura, paciencia y complicidad ha logrado que llegue a ser el hombre que soy ahora, aún me falta mucho por mejorar pero juntos iremos mejorando gracias por todo Maricita. Y por ultimo pero no menos importante agradecerte a ti amolsito, que desde siempre has sido mi amiga mi compañera mi consejera, sin juzgarme, queriéndome y confiando en mi. Este logro es de todos nosotros. El futuro comienza hoy.

Contents

List of Figures

Glossary

AI Artificial Intelligence

ANN Artificial Neural Networks

AUC Area Under the ROC Curve

BN Bayesian Networks

BPDP Bivariate Partial Dependence Plots

DNN Deep Neural Networks

EDA Exploration Data Analysis

FPR False Positive Rate

GIRP Global Model Through Recursive Partitions

HCI Human Computer Interaction

ICE Individual Conditional Expectatives

ICI Individual Condition Importance

KNN k-Nearest Neighbor

LIME Local Interpretable Model-agnostic Explanations

LOCO Leave One Covariate Out

LRP Layerwise Relevance Propagation

MCR Model Class Dependency

MDD Model Driven Development

MDP Markov Decision Processes

ML Machine Learning

NLG Natural Language Generation

OSRE Orthogonal Search-based Rule Extraction

PDP Partial Dependence Plot

ROC Receiver Operating characteristic Curve

SHAP Shapley Additive exPlanations

SFIMP Permutation based Shapley Feature

TPR True Positive Rate

WDBC Wisconsin Diagnostic Breast Cancer

XAID eXplainable Airtificial Intelligence for Designers

XAI eXplainable Artificial Intelligence

CHAPTER 1

Introduction

In this chapter, we will describe the context of the project. For this purpose, we will enumerate each part that composes the project, underlining its role in the final goal of this project. Moreover, we will describe the motivation for the development of this project and the selection of each one of the technologies used in its development. We will also detail the goals of this project and the structure of this thesis.

1.1 Context

eXplainable Artificial Intelligence (XAI) is a term that was first implemented in 2004 by Van Lent et al. [5] in this way, it was intended to describe the ability of your system to explain how AI controlled entities behave in the simulation video game application. Certainly, the term XAI is correspondingly new in our society.

However, the need to explain this skill dates back to the mid-1970s as we will see later when discussing the evolution of explanation in machine learning systems. Unfortunately, the improvement to solve this problem was affected and slowed down when the AI reached a turning point with the advances in machine learning. Since then, the focus of AI research has shifted towards the implementation of models and algorithms that emphasize predictive power while the ability to explain decision processes has been left in the background.

Recently, the XAI idea has again received the attention of academics and practitioners. The rebirth of this research topic is the continuing result of AI's unstoppable penetration of all industries and its impact on the critical decision-making process, all without being able to provide a detailed argument about the reasoning process that the system has had to get to make such decisions, recommendations, predictions or actions. Due to this, in the face of clear pressure, not only socially, but ethically and legally, new AI techniques capable of making decisions that are not only explicable but must also be understandable are required.

The term XAI usually refers to initiatives and constant efforts that are made seeking that response to transparency and lack of trust, rather than a technical and formal concept, having said this and to put some sensible clarification to this trend below some definitions of XAI will be cited.

According to DARPA [6], XAI aims to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, trust, and effectively manage the emerging generation of artificially intelligent partners."

The purpose of enabling applicability in machine learning, as established by FAT * [7], "is to ensure that algorithmic decisions, as well as any data that drives those decisions, can be explained to end-users and other interested parties in terms, not technicians".

FICO [8], the organizer of XML Challenge, sees XAI as "An innovation to open the ML black box" and as "a challenge to create models and techniques that are both precise and provide a good reliable explanation that satisfies the customer needs".

XAI focuses on removing this myth from the "black boxes", this implicitly suggests a

responsible AI, with this it would help build and produce transparent models that are not affected by the precision with which decisions are made, that is why that in both AI and machine learning it is very common for there to be a trade-off between the precision of how decisions are made and their interpretability. Accuracy is closely related to the quality and quantity of the training data.

To finish, XAI belongs to a group of new generation AI technologies called third-wave AI, this new wave of AI aims to generate precision in algorithms that can explain themselves. Thus seeking to achieve systems with a level of intelligence that resemble the human. A difficult, but not impossible, goal is that known as AGI.

1.2 Motivation

While the prospects for AI, in general, are uncertain, it is possible to conclude that computer scientists are quite optimistic about the progress of technology and consider themselves fortunate to have the opportunity to learn from it. In addition to the overall assessment of the importance of AI, even large organizations like Google and the intelligence industry are actively investing in A.I. as one of the key pathways for other AI approaches. With over ten years of development at Google, there is no shortage of examples of AI-related advancements that have been critical.

These include projects like Google's image recognition technology, which enables customers to take and evaluate hundreds of miles of image resolution. Now, a characteristic of the intelligence of the human being and that over-selling is the ability to try to explain the logic of another human being in making decisions.

This characteristic is not only important in the field of social interactions, seeking to be a "normal" individual since a person who does not express his intentions and thoughts is more likely to be considered a person with a "strange" personality, but It is also important in the educational field, where students at any level try to understand the reasoning of their teachers.

Explaining how decisions are sometimes made is a previous step to gain trust between people, a clear example is that of a doctor when he explains to his patient the reason for his treatment.

However, these social aspects are irrelevant to AI systems and their models, there are many arguments that cover the explainability behind AI. The most important are:

- System check: It is very common that most applications use black-box systems and it

is just as common that people do not trust these systems, the use of models that can be interpreted and verified by experts are of great need. In the work that shows [9] we appreciate that an AI system that was trained to predict the risk to which a person is subjected to become ill with pneumonia, and sadly leads to erroneous results.

This black box system increases the death rate in people who are related to pneumonia. In summary, this model expresses that patients who have asthma and heart problems are less likely to die from pneumonia than a healthy person, this error would be quickly detected by a qualified doctor and would deny the claim of the system. Problems (asthma and cardiological) are factors that negatively affect the prognosis for recovery.

However, the model, since it does not have a pure understanding of asthma, heart problems, or pneumonia, makes inferences it has about the data that is it gave. These data may be systematically biased because people with asthma and heart problems are often under strict medical supervision, and healthy people are rarely under this supervision.

Due to this information treatment, the system drew a conclusion in which asthmatic patients had a lower risk of dying from pneumonia. As expected, this correlation has no causality, and for this reason, it should not be taken as the basis for the prescription of therapy against pneumonia.

- Ameliorate the system: The purpose of improving an AI system is to understand its weaknesses, and in black box systems it is more difficult to perform such compression than in models that are somehow explainable and interpretable. In addition, biases must be detected in the data that is treated as well as in the model, with this it is more logical and easier to understand what the model is doing and how it comes to make its predictions.

  Interpreting a model is also useful to later be able to make a comparison between different models or architectures since it is not a secret that several models can have the same classification performance, but they also have differences in many terms of the characteristics they use for the shot. of their decisions, if at any time we want to know which is the most "appropriate" model, this model must be explainable, moreover, the more we know about the operation and what our model is doing and where it can fail, the easier it is to work with him, and so we can make improvements to make him a better model.

- Digging into the System: Today it is very common that an AI system is trained with countless examples, and this is why these systems can observe patterns in the data that human beings do not see due to the limitation of being able to learn a certain

number of examples, once the XAI is merged into AI systems, it will be possible to try to extract the knowledge that the system distills and in this way enrich the knowledge. In the branch of sciences (physicists, chemists, and biologists), they are very interested in systems that can enrich the hidden knowledge of their professions and it is there that a system that does not provide an interpretation or explanation is useless.

- Respecting the legislation: As we show AI systems are increasingly present in our lives, and this does not only mean that this would bring good things, it is for this argument that we must think that the use of AI in our daily lives would also affect the legal aspects, for example, if a system is wrong in its decision, who can be assigned responsibility for the acts of a system? it is a topic that is in vogue and has received greater attention today. since it is very difficult to find an answer that satisfies those legal problems using black-box models, and because of this it is that.

  AI systems that are implemented in the future for legal decision-making must be more explainable, another terrain in the fact that regulations are a force that drives systems to be more explainable are the individual rights of people who are affected by the decisions of an AI system, for example, a bank revenue, people might want to know why a system decided on them for better or for worse, and only XAI systems have the ability to offer this information. This concern caused the European Union to adapt the regulations where a "right of explanation" is implemented, where a user can request an explanation of an algorithmic decision that was made about him or her [10].

The above reasons are motivating enough to say that explainable AI systems are necessary. But this is not something new, history also gives us evidence that the concern that a system can be explained is not new.

1.3 Project goals

The main goals that are intended to be met in this document are:

- Analysis of the state of the art of XAI.
- Exploring algorithms and benchmarks for XAI techniques.
- Analyze eXplainable Artificial Intelligence (XAI) trends.

1.4 Structure of this document

In this section, we will briefly summarize the chapters contained in this document. This structure is the next one:

Chapter 1 **Introduction**

Chapter 2 **State of the Art**

Chapter 3 **Datasets**

Chapter 4 **Comparative Analysis of Current Trends**

Chapter 5 **Results**

Chapter 6 **Conclusions**

CHAPTER 2

State of the art

In this chapter, we will talk about the state of the art of explainable artificial intelligence, a couple of brief definitions of what explainable artificial intelligence is, in addition to the history of artificial intelligence and how it has evolved over time with previous and modern works. to understand what it was and is doing now to try to get artificial intelligence that is capable of explaining the reason for its resolutions.

2.1 Introduction

In order to do this work, it is necessary to know where is the current knowledge of what is intended to be applied, analyze And to compare, it is for this reason that in this chapter we will see a compilation as deep as possible of the result of other investigations, works, finds, or publications that tell us where the explanation of artificial intelligence is currently located.

For this, several steps have been followed in this way, we have ensured that the previous research is as accurate as possible, and it is for this reason that a selection of bibliography has been made, in which we will find very interesting doctoral theses or works in their studies. , Articles of a scientific nature, published in magazines and serious repositories, and the searches were made only through the academic google.

In addition to this, a classification has been made in order to better classify the structure of this chapter, starting with the story behind the explanation of artificial intelligence, but without delaying both the timeline or the work previously done.

Then the perception that human beings have about artificial intelligence is exposed, since not long ago and contemporaneously this type of technology is in impressive exponential growth in all aspects of human life without distinction, it is for this reason that the human being is in the middle of the knowledge loop, and it is good to know and say what is the role that humans currently play in this technology, to know if people are really interested in the answers or decisions that can take a machine learning model are of importance to people in general, the human being is so curious as to want to know how an algorithm has decided on it. It is an interesting approach that needs to be addressed in order to continue carrying out this work.

After this, it is also interesting to address the issue of whether or not a model can give an explanation that is the closest thing to human reasoning, this has also been a subject of much discussion in the past and that is why there are tests that they can model in such a way that one can have the conviction that the system can reason almost like a human being. Then that leads us to the fact that the model must give us an explanation that is as friendly as possible for our understanding. It is useless for a model to solve very large and critical problems, if the answer it gives us is too complex to understand, and already there in that point you can say the model actually made a prediction, learned something in the process or just guess or did mathematical probability calculations to get to the explanation

2.2 History of eXplainable Artificial Intelligence (XAI)

2.2.1 Bayesian Network and Expert Systems.

Rules-based systems are a good model from which to start when it comes to the story where you started asking for explanations to a machine learning system, these systems were treated as system design tasks, in other words they had the task to design a system that had the capacity to generate segregations in decision-making, but it was not until the 1970s, as we had previously said when there was a need to explain the decisions of expert systems. [11] carried out a work detailing a framework to create expert systems that have the capacity to give explanations and was one of the pioneers in exalting the importance of being able to justify the explanations that the systems provide. [12] is work done after prior analytics to justify explanations from expert systems, but both works were done for rule-based systems and are based on a domain-specific taxonomic knowledge pillar and strategic knowledge base separated. Subsequent to this work, in [13] a further step was taken as knowledge was further separated into three layers, adding a communication layer between the previously mentioned taxonomic domain and the strategic layers. With this separation, the communication layer was intended to create a communications expert to create solutions independent of the specific system and domain.

On some occasions, probabilistic decision-making systems, called Bayesian Networks (BN), are regarded as an expert and experienced systems, which are the successors of rule-based systems. The little work that is available for this type of system (BN) makes them describe themselves as explanations of expert systems, [14] shows in their work a survey of methods of explanation of these systems (BN), in addition to making an analysis of these methods in terms of some of their ways of explanation. There is a particular interest in the classification of the explanatory approach from the reasoning, the model and the evidence of the decision. Most of the explanation of Bayesian networks has been within the narrow context of a particular system, and is based on the production of a text that shows the actual posterior probabilities of each node, but does not provide an explanation of what they symbolize.

2.2.2 Recommendation Systems.

They are online services that are under the service of a large number of users, we offer recommendations on a personal and individual level for the needs of each of its users and be these means or products. Specifically, these recommendation systems want to create an

intuitive and very brief justification, so that it helps its users to decide whether to accept this recommendation or not. [15] In his work he presents us with an experiment that consists of measuring the satisfaction of users before various systems that offer a variety of justifications for a film recommendation system, where they discovered that the simplest methods were those that were most accepted by users and were more conclusive, these systems consist of showing the ratings of other users or showing some important feature of the film, such as an actor who participates in it. Justifications were obtained under the concepts of machine learning as a trust model in addition to other more complex justifications, 86% of users suggested that these justifications had to migrate to other systems, other studies that were done at the beginning of the year 2000 also demonstrated users specify that they are more satisfied with systems that somehow justify the reason for their recommendations [16]. Years later in 2009 [17] a way of justifying that focused on the most important feature merged into the User's record regarding that function. A study of users showed that styling the way something was justified was much more satisfying to users than previous methods. [18] further defined a classification for the explanations of the recommendation systems in three types:

- Explanations based on elements that were previously chosen by users.
- Explanations based on the choices of other users with similar tastes.
- Explanations based on characteristic.

In addition, a hybrid recommendation type was defined that combines two or more of the recommendations mentioned above, and by analyzing one user, the study was able to conclude that explanations based on characteristics were the best of the three main types, also that the hybrid explanations gave even better results. [19] stated that in previous studies the persuasion of justification and not justifiability had generally been assessed, their work showed that to justify, these feature-based justifications were superior to those based on neighbor and history of the user.

2.2.3 Experimentation In Other Similar Fields.

Generating explanations for users have also been seen to migrate in problems where users have a role where they interact in solving a specific problem and given this it is necessary that the resolution of said problem is understandable and understood because certain choices were made. In the work done by [20], and in [21] they explore a method that presents explanations for any allocation decision made by their program.

An assignment can be explained by identifying the above tasks so that they form a firm enough basis to justify the current one. Applying this at each point in the subsequent assignment creates what the authors call an "explanatory tree." Context-aware systems are systems that have the ability to detect changes in the environment and also respond to those changes. Some have been trained in order to provide their users with explanations that demonstrate the behavior of the systems. A 2007 study [22] showed mental models developed by the users of a System that predicted the interruptibility of their managers, although, they concluded that the contributions of low-level characteristics that were shown to their users were not very useful to be able to understand the system and it was suggested to use a higher level of concepts. In 2010 [23] they designed a toolkit to be able to use context-sensitive applications, which generate eight explanatory variants of four of the most widely used models (rules, decision trees, Neive Bayes and the hidden model of Markov (HMMs)).

In addition, the study on the explanation of Markov decision processes (MDP) has continued. Depending on the context and the particular state an MDP is in, in some circumstances it is better to explain to the user what is the best current course of action and why. Work has been done to explain the Markov decision processes (MDP). In the context of a particular state in an MDP, it is sometimes convenient to explain to a user what is the best current course of action and why. [24] describes an explanation system that helps plant operators carry out the necessary operations; [25] in their work they explore how to give the minimum sufficient explanation for certain activities, such as in the student environment, choosing the next course in a university study plan; [26] They also propose a dialogue system, instead of a single fixed explanation system, which allows the user to debate and question. The case-based reasoning community has also made very great efforts to find an explanation in probabilistic systems as an example we will take [27] who proposed a case-based explanation method for decision support systems, where specimens Alternatives are selected and the explanations provided by these specimens focus on how they differ (if the decision is different) or if they have similarities (otherwise). To speak of causal discovery is to refer to determining the direction of causality between the variables in a model, which would certainly help to explain the behavior of the models. [28] carried out an investigation of nonlinearity as non-Gaussianism of the real data in order to identify the causality between the variables, even adding noise, trying to demonstrate causal relationships is very useful to justify the predictions based on these models that are provided to users.

In forensic science, there is a work, [29] that explained legal cases by merging the Bayesian network model with a narrative language. They called it "scenario", considered statistical evidence, and also maintained a narrative framework that allows a judge or jury

to understand the arguments that are given to suppose certain things and thus be able to distinguish what the system does and the relationships between them. In this way you can get to know the structure of the statistical model, which is crucial for humans, to be able to make an informed decision in a legal case. In [30] we can see that a very similar approach was used to generate explanations for Bayesian networks in legal cases. Their work is based on defining a support graph directed towards the variable of interest, then using it to construct an argument.

Another characteristic that has been studied a lot is the interpretability in the contexts of the communicating agents. [31] experimented with neural agents trying to learn to communicate with each other about images. Then making use of human supervision whose task is to lay the foundation for learned communication in a way that would be understandable to humans. [32] also carried out studies on the messages passed between agents of a system with deep learning in communication policies. In this way they were able to develop a strategy to translate these messages into natural language based on the underlying beliefs implied by the messages. This seems to want to understand the belief that any model has.

One field that is very closely related to explanation is Natural Language Generation (NLG). This NLG is very adaptable and can be used in a different way and in different kinds of sophistication to be able to generate explanations. In addition to having the ability to explain machine learning and others, however, AI systems have worked on explanations of other types, for example [33] they generate interesting explanations of user interactions with a software system intended for administration, while that [34] in their work explain how the weather forecast was made and shows that it particularly helps readers to decide whether to believe the forecast or not.

Another work is that of [35] in which we can see a proposal to generate explanations as a tool to debug a description system based on logic that could be used by the developers and users of said system. First, the rules of inference are broken down into atomic descriptions; Corresponding atomic explanations are created using subsumption rules and chained together to form tests supporting the system's conclusions. In the constant investigation to be able to make an AI system explainable, several methods and strategies have been proposed to be able to carry out an explanation that adapts to the needs of the users and is also reliable, now it will give an overview of the methods of existing interpretation.

Most research discusses explainability in machine learning algorithms and will, therefore, be used to refer to this with the term "interpretability". Based on the literature search, [36] classifies these methods taking into account three criteria:

1. The complexity of interpretability.
2. The scoop on interpretability.
3. The level of dependency of the ML model used.

It can be said that, since being able to explain in AI is still in development, the classes of methods that belong to the proposed taxonomy are not mutually exclusive or exhaustive. However, this can be a good criterion for comparing and contrasting multiple methods. Next, we will describe the main characteristics of each class and give examples of current research.

2.2.4 Complexity Related Methods.

When it comes to the complexity that a machine learning model can have, we know that it has a close relationship with its interpretability. It is generally known that the more complex the model, the more difficult it will be to interpret and explain the model. Because of this, it was concluded that the easiest way to interpret a model would be to design an algorithm that is substantial and easily interpretable. Many scientific papers support this classic approach, to name a few: [37] showed the application of a learning method based on generalized additive models for the problem of pneumonia as previously mentioned. In this way they demonstrated the intelligibility of their model through case studies on real medical data. In the work they present in [38] they implemented a model based on attention that automatically learns by detailing the content of the images. This work also showed through visualization how the model can explain the results. [39] in their work, they presented dispersed linear models to implement a data-based scoring system called SLIM. The results of this work are dominated by the ability to interpret the proposed system to provide users with understanding due to its high level of scarcity and small integer coefficients A common obstacle that makes the use of this class of methods a difficult challenge to implement is the trade-off between interpretability and precision [40]. As noted in [41] "precision generally requires more complex prediction methods ... [and] simple and interpretable functions do not make the predictors more accurate." In a sense, you have intrinsic models that can be interpreted at a cost of precision.

A perspective from another angle on the machine learning interpretability approach is to build a high interpretable complex, this means creating a black box model with high precision and then using a separate set of techniques to perform a kind of reverse engineering to provide the necessary explanations without altering or even knowing the internal workings of the original model. Although this translates into a high degree of complexity as well as

an economic aspect, the most recent work carried out in the XAI field belonging to the post-hoc class includes explanations in natural language, visualizations of learned models, and explanations. With the information presented above, we can affirm that the general utility value of the interpretability depends on the nature of the prediction model that is being used. As long as the model is accurate for the task, and uses a suitably restricted number of internal components, the intrinsic interpretable models are sufficient.

However, it should also be borne in mind that in several works there is a group of intrinsic methods that are used for highly complex models and that are not interpretable. The purpose of this intrinsic method is to change the internal structure of a complex black-box model that, due to its own characteristics, is not interpretable, generally a DNN, to attenuate its opacity and, therefore, improve its interpretability [42]. The methods used can be components that add additional capabilities, components that belong to the model architecture [43], [44], is an example of this. as part of the loss function [45], or as part of the architecture structure, in terms of operations between layers [46], [47].

2.2.5 Scoop Related Methods

Interpretability involves understanding an automated model, this allows two new variants to be admitted, according to the scoop on the ability to interpret: understand all the behavior of the model or understand a single prediction. Consequently, we distinguish between two subclasses:

1. Global Interpretability:

   Global interpretability is useful to try to understand all the logic behind a model by following all the reasoning and leading us to all the different possible results. These types of are useful when machine learning models are critical to inform decisions at the population level, such as the percentage of drug use or trends or climate change. There are works and documents that propose generally interpretable models include additive models as well as sets of rules generated from a generative Bayesian model. However, these models are usually particularly structured, therefore conditioned on predictability to preserve interpretability. [48] in their work they proposed a way of interpreting the global model through recursive partitions called (GIRP), in this way a global tree of interpretation could be built for a wide range of machine learning models based on their local explanations. While conducting the experiments, the authors clarified that their method can reveal whether a particular machine learning model is behaving in a reasonable way or if it is over-trained for some irrational pattern. [49] raised a super vision, it was a new proposal, it was a new approach to information extraction, which

provides a global and deterministic interpretation. This work reaffirms the idea that representation learning can be successfully combined with traditional patterns-based startup models that produce patterns and these patterns are interpretable. [50] on the other hand, they proposed a vision based on maximizing activation, minimizing the preferred inputs for neurons in neural networks. The activation maximization technique was previously used by [51] Although multiple techniques are used in the literature to allow global interpretation. Arguably in the global model Interpretation is difficult to achieve in practice, especially for models that exceed a handful of parameters. Analogically like a human, who focuses effort on only one part of the model to understand everything, local interpretability may be more easily applicable.

2. Local Interpretability:

   Explaining why a specific decision or single prediction was made means that interpretability occurs locally. The first part of interpretability is generally used to generate separate explanations to test why the model makes instance-specific decisions. Some exploratory works have proposed local interpretation methods. The following is an overview of the interpretation methods explained in the review document. In [52] we are presented with LIME, which is a proposal for local interpretation. Explanation of the agnostic model. This model can approximate a black box model locally in the neighborhood of any prediction of interest. The newer, related and highly anticipated work by the creators of LIME, called anchoring, extends LIME using decision rules. (LOCO) [53] is another popular technique for generating local explanatory models that offer important local variables measured In another attempt to produce local explanations. We have [54]. In this work, the authors presented a method that can be used to explain the local decision of any nonlinear classification algorithm using the local gradients that characterize how a data point must move to change its predicted label. After this work, we found a series of documents with similar methods for image classification models [55] - [56]. In fact, it is a common approach to understand image decision classification systems by finding areas of an image that were particularly influential in the final classification. These approaches are also called sensitivity maps, salience maps, or pixelmap maps [57] and use occlusion techniques or gradient calculations to assign individual pixels an "importance value" that is intended to reflect their influence on classification final.

   Based on breaking down the predictions of a model into contributions of each function, Robnik-Sikonja and ˆKononenko [58] proposed to explain the prediction of the model, on the one hand, measuring the difference between the original prediction and the prediction made, omitting a set of characteristics. Although there are several

techniques to obtain local interpretations [59] - [60], the latest work by Lundberg and Lee [61] shows that there is equivalence between these techniques. They introduced a promising new technology with strong theoretical backing, called "Well-proportioned explanations that unify local approaches." An interesting and promising approach to work is to combine the advantages and benefits of global interpretability. The four possible combinations are:

- The interpretation of the standard global model responds as a model that makes predictions.
- The interpretability of the global model at the module level determines how various parts of the model affect the predictions.
- The local interpretability of a set of predictions explains why the model makes specific decisions for a set of instances.
- Finally, the usual local interpretation of a single prediction is used to demonstrate why the model makes a decision for a specific instance.

Another observation worth considering is that, in the reviewed literature, local interpretation is the most common method to generate interpretation in DNN. However, although these methods have been developed to explain neural networks, the authors often emphasize that their methods can potentially be used to explain any type of model, which means that they are agnostic models. Another way to classify explanations. It will be discussed below.

2.2.6 Model Related Methods

Another method of classifying interpretability is model technology, which is based on an agnostic model, meaning that it can be used in any of the types of machine learning algorithms or specific technology models mentioned above. This means that they only apply to a single class type or algorithm type.

2.2.6.1 Specific Interpretability Modely

Model-specific interpretation methods are limited to specific model classes. In essence, the inherent method is a specific model. The downside to this approach is that when we need a specific type of explanation, we are limited in providing model options for the model. As a result, interest in agnostic model interpretation methods has recently increased.

2.2.6.2 Agnostic Interpretability Model

In machine learning, the method of interpreting the agnostic model is not tied to a specific type of model. In other words, such methods separate prediction from interpretation. The agnostic interpretation of this model is usually a post-fact interpretation, generally used to explain ANN, and can be a local or global interpretable model. In order to improve the interpretability of AI models, statistical range techniques have recently been used, and machine learning and data science have developed a large number of models using these agnostic methods. Below are the usual technology groups, divided into 4 types of technologies:

Visualization:

The natural idea of understanding machine learning models (especially DNNs) is to visualize their representations to explore hidden patterns within neural units. As expected, some studies have tried to find content in black boxes with the help of various visualization techniques. Visualization techniques are essentially suitable for supervised learning models. In the reviewed literature, the most popular visualization techniques are:

- Substitute Models: The proxy model is a simple model to explain complex models. More specifically, it is an interpretable model (such as a linear model or a decision tree) that is trained in the predictions of the original black box model to explain the latter. However, there is little theoretical guarantee that simple alternative models can represent more complex models. The LIME method mentioned above is a prescribed method for building a local agent model around a single observation. [62] Using an alternative model approach, they extracted a decision tree representing the model's behavior. Another notable work by Thiagarajan [63] proposed a method for constructing TreeView visualizations using alternative models.
- Partial dependency graph (PDP): PDP is a graphical representation that can help visualize the average local relationship between one or more input variables and black box model predictions. Jobs that use PDP to understand supervised learning models include Green and Kern [64]. He used this technique to understand the relationship between the predictors of voter mobilization experiments and the mean value of conditional therapy and to make predictions through regression. Additives. In [65], they rely on increasing stochastic gradient and used PDP to understand how different the environment is that affects the distribution

of a particular freshwater. Berk and Bleich [66] demonstrated the benefit of using random forests and associated PDPs to accurately model the low-cost asymmetric classification prediction-response relationships that are common in criminal justice settings. Recently Welling, in collaboration with other authors[67], proposed a method called forest floor to visualize and interpret random forest models. The proposed techniques are based on contributions of method characteristics rather than PDP. As the authors argued, Forest Floor's advantage over PDP is that interactions are not masked by the average. Therefore, it is possible to locate interactions that are not visualized in a particular projection.

- Individual Conditional Expectations (ICE): The ICE framework extends the PDP, and the PD framework provides an excellent view of how the model works, and the ICE chart reveals individual interactions and differences by decomposing the PDP in the output. Recent work uses ICE instead of classic PDP. For example, Goldstein introduced ICE technology in [68] and demonstrated its advantages over PDP. Later, Casalicchio and others proposed the importance of local characteristics in [69]. Such mapping of local importance (PI) and individual conditional importance (ICI) are used as basic methods of visualization tools.

Knowledge Extraction:

It is difficult to explain how the machine learning model works, especially when the model is based on ANN. The multi-layered energy network is a general approximator. However, since the learning algorithm modifies the cells in the hidden layer, this can be an interesting internal representation. Therefore, the task of extracting explanations from the network is to extract the knowledge acquired by the ANN during the training process in an understandable way and encode it in an internal representation. In the research literature, several works have proposed methods to extract knowledge embedded in artificial neural networks, which can be based on two techniques:

- Extraction of rules: The job of obtaining information on highly complex models is to use rules [70] [71] [72] for extraction. This technique proposes a method of using rule extraction to provide a symbolic and understandable description of the knowledge learned by the network during its training. These rules use inputs and outputs of ANN to approximate the decision process in ANN. Incidentally, this is a type of knowledge used by experts in traditional artificial intelligence systems. The survey carried out by Ras [73] classified the previously proposed rule extraction strategies and proposed three methods to extract rules:

1. extraction of pedagogical rules
2. extraction of decomposition rules
3. Extraction of eclectic rules.

The decomposition method focuses on extracting rules at the level of each unit in the trained artificial neural network, that is, the vision of the basic artificial neural network is one of transparency. Although the teaching method treats the trained ANN as a black box, that is, the view of the underlying ANN is opaque, the orthogonal search-based rule extraction algorithm (OSRE) that he proposed[74] is a successful teaching method It is commonly used in biomedicine. The third category (commitment) is the hybrid rule extraction method, which combines the decomposition technique and the rule extraction teaching method [75]

- Distillation Model: It belongs to another technology of knowledge extraction. This is a distillation model. Distillation is a compression model used to transfer information (dark knowledge) from the depth of the network ("teacher") to a superficial network ("student") [76], [77]. Model compression was originally proposed to reduce the computational cost of the model at run time but has since been used for explanatory purposes. Therefore, together with other researchers, they studied how to use distillation models to distill complex models into transparent models. Other researchers together with Che [78] introduced in their article a knowledge extraction method called "interpretable imitation learning". This method learns interpretable phenotypic characteristics, thus mimicking the deep learning model. At the same time, a reliable prediction was made. The latest work by Xu and other collaborators introduce DarkSight [79], a visualization method used to explain the predictions of black-box classifiers in data sets inspired by the concept of dark knowledge. The proposed method combines knowledge extraction, rationalization, and DNN visualization.

Influence Methods:

This technique estimates the importance or relevance of the characteristics by changing the input or internal components and recording the degree to which the changes affect the performance of the model. Impact technology is generally visualized. In the reviewed literature, there are three alternative methods to obtain the correlation of the input variables:

1. Sensitivity Analysis:Sensitivity refers to how the ANN output is affected by its input interference and/or weights [80]. It is used to verify that the model's be-

havior and output remain stable when it intentionally interferes with the data or simulates other changes to the data. The results of the Visual Sensitivity Analysis (SA) are considered agnostic interpretation techniques since showing the stability of the model as the data changes over time increases confidence in the results of machine learning. SA has been used increasingly for the interpretation of general ANN, especially the classification of DNN images [81] and [82]. However, it is important to note that SA does not interpret the value of the function itself, but only its variants. The SA is generally used to test stability and reliability models, it can be used as a tool to find and remove unimportant input attributes, or as a starting point for more powerful interpretation techniques (such as decomposition).

2. Layer Relevance Propagation (LRP): In [83], another technique for calculating correlation is proposed, such as a hierarchical correlation propagation algorithm. LRP begins to propagate from the network's output layer to the input layer so that the prediction function is redistributed backward. The key feature of the redistribution process is called map retention. Compared to SA, this method explains the prediction about the state of maximum uncertainty, that is, it identifies the basic properties of the "rooster" prediction.

3. Characteristic Importance: Variable importance quantifies the contribution of each variable input (characteristic) to the prediction of complex ML models. After replacing the function to measure the importance of the function, the increase in the model prediction error is calculated. The arrangement of the important characteristic values increases the model error. By exchanging unimportant values, the model will ignore the characteristics, keeping the model error unchanged. Based on this technology, Fisher proposed an independent version of the feature importance model in [84], called model class dependency (MCR). The work mentioned in [76] above proposes a local version of important functions for the permutation-based Shapley feature, called SFIMP.

Example-Based Explanation: An example-based interpretation technique selects specific instances of the data set to explain the behavior of the machine learning model. Example-based explanations are largely model-independent because they make any ML model easier to interpret. The subtle difference from the agnostic method model is that the example-based interpretation method interprets the model by selecting instances from the dataset without acting on the model's features or transformations. Based on the review conducted, we identified two promising interpretation techniques based on examples:

1. Prototypes And Reviews: Prototypes are the choice of representative data instances, so project membership depends on their similarity to prototypes that lead to excessive generalization. To avoid this, top anomalies, also known as critical anomalies - those prototypes don't represent instances well, Kim developed an unsupervised algorithm to automatically find prototypes and comments from the dataset, called MMD-critical. When applied to unlabelled data, you will find prototypes and criticisms that characterize the dataset as integers.
2. Contractual Explanations: Wachter et al. [85] proposed the concept of "unconditional interpretation of facts" as a new type of automatic interpretation of decisions. Counter-factual explanations describe the minimum conditions that lead to alternative decisions (for example, bank loans) without having to describe the full logic of the algorithm. The point here is to explain the single prediction rather than the confrontation. As research contributions to such methods continue to grow, new technical models regularly propose agnosticism. Finally, it is worth noting that the main advantages of the agnostic model are the "flexibility" of the model, the level of interpretation, and representation. However, although agnostic model interpretation techniques are convenient, they are often based on alternative models or other methods, which may reduce the precision of the interpretation they provide. Although model-specific interpretation techniques tend to use direct interpretation models, this can lead to more accurate interpretations.

2.2.7 XAI Measurement: Evaluating Explanations.

Doshi-Velez and Kim [86] questioned the interoperability of measurement and evaluation. In fact, although more and more research produces interpretable machine learning methods, there is little work to evaluate these methods and quantify their relevance (only 5% of research studies focus on this topic). This may be due to the subjective nature of in-

terpretability. However, considering the number of existing interpretation methods, these methods should be compared, verified, quantified, and evaluated. Doshi-Velez and Kim established a baseline of evaluation methods and proposed three main types of evaluation interpretability:

1. application-based: put the explanation in the application and let the end-user (usually a domain expert) try it out. This guy assesses the quality of an explanation in the context of his final assignment.

2. Human-based: This involves conducting a simplified app-based assessment where experiments are conducted with lay humans rather than expert domain. This type is more appropriate when the goal is to test more general notions of the quality of an explanation.

3. functionally based: this type does not involve humans, it is more appropriate once we have a class of models or regulators that have already been validated.

Based on Doshi-Velez's evaluation classification, Mohseni and Ragan [87] proposed human-based evaluation as a benchmark for evaluating the interpretation of text and image data examples. They demonstrated this by comparing the interpretations generated by the classification model with the reference annotation metadata and can assess the quality and adequacy of local interpretations. Therefore, to show how to use human-based assessment as a method to assess local machine learning, Huysmans and others [88] studied decision trees, decision tables, propositional rules, and tilt rules to understand which is the most explainable. To this end, they conducted experiments with end-users to compare. They found that decision trees and decision tables are usually the easiest to explain, but different tasks make trees or tall trees more desirable. Backhaus and Seiffert [89] proposed a quantitative method to compare the explanatory power of the LD method. These tests take into account various machine learning methods learned from actual spectral data. Poursabzi-Sangdeh believes that quantifiable interpretability means that interpretability is defined consistently with a set of human interpretable concepts, and he proposes a general framework, ie "network anatomy to quantify the interpretability of the potential representation of ANN", Thus identifying the semantics of the hidden units of any unit. Given a neural network, then combine it with concepts understandable to humans.

Bau [90] has different interpretations of the interpretation capacity, which is considered as a potential attribute, which can be affected by different isolation factors (such as the number of inputs, the complexity of the model, and even the user interface), and can affect different results. Measured (such as the end user's ability to trust or debug the

model). They worked hard at work related to the manipulation and measurement of model interpretation. This is an interesting experiment that involves changes in the factors that are considered to make the model more or less interpretable and to measure how these changes affect the Decision People pay attention to two factors: the article number and if the model is transparent or black box. Paul's claim is based on the fact that various methods have been proposed to improve and evaluate the interpretation of the theme model, and he discusses how to apply the ideas of the theme model (such as human feedback and automated indicators) to the evaluation of the theme model. The interpretability of ML. Gilpin's latest work [91] proposes a methodological method to evaluate the interpretation of AI models in taxonomy, which can distinguish three types of interpretation: simulation processing, interpretation representation and interpretation of producer networks. Human factors are a common factor that directly affects the quality of interpretation and has been addressed from different perspectives in previous studies.

2.3 XAI Perception: Human In The Loop.

Interpretation and understanding are two different actions: interpretation depends mainly on the content to be explained (i.e. the original model) and the form of interpretation (i.e. the interpretive method), and understanding also depends on these elements. Who accepts the explanation (in other words, the person who explains, in other words, the person). To be interpretable, machine learning models must be understandable by humans. This is challenging. XAI is designed because it involves complex human computing processes that require HCI skills and ML experience. Furthermore, due to the explanation of human behavior, it has long been studied in the field of philosophy and psychology. Therefore, we must refer to this field to simulate the process of human interpretation and draw inspiration from the models developed in these fields, so keeping people informed is the deciding factor in determining the full value of interpretation. However, the literature review carried out has identified a few documents on the effects of human factors on the XAI. In this axis, we investigate works that discuss human characters from two perspectives:

1. the first focuses on how to produce explanations that simulate the human cognitive process.

2. the second focuses on how to produce human-centered explanations.

2.3.1 Human Explanations

Miller [92] is probably the most important attempt to clarify the connection between human science and XAI and provides in-depth research in the article. Research on philosophical, psychological, and cognitive sciences that deals with interpretive issues. The author notes that the latter can be a valuable resource for progress in the XAI field. He highlighted three main findings:

1. The explanations are contrasting: people do not ask why event E happened, but why event E occurred at the site of some event F.
2. Explanations are selective and focus on one or two possible causes and not all causes of the recommendation.
3. Explanations are social conversation and interaction for knowledge transfer, which implies that the explainer must be able to modify the mental model of the explained while participating in the explanation process. He believes these three points are imperative to consider if the goal is to build a useful XAI.

In [93] someone requested the use of social science models at XAI. The author of this article believes that most of the existing literature on XAI methods is developer-based. Intuition, not focus on the target users. Based on a brief bibliographic survey, they showed that scientific research is rarely carried out on current XAI research and made some important discoveries in the field of XAI-related human sciences. Returning to the taxonomy of interpretation methods, it is worth noting at this point that postmortem interpretation techniques are similar to how humans interpret decisions. As Lipton said: "In a way, we believe that humans are interpretable, which is applicable to interpretability (after the fact)." Furthermore, the example-based method of analysis is clearly inspired by science. Cognition of human reasoning Specifically, human reasoning generally has a prototype, using representative examples as the basis for classification and decision making. Similarly, Kim's method uses representative examples to interpret and group data.

2.3.2 Friendly Explanations

Returning to the taxonomy of interpretation methods, it is worth noting at this point that postmortem interpretation techniques are similar to how humans interpret decisions. As Lipton noted: "In a way, we believe that humans are interpretable, which is applicable to interpretability (after the fact)." Furthermore, the agnostic method of example-based interpretation is clearly inspired by science. Cognition of human reasoning Specifically,

human reasoning generally has a prototype, using representative examples as the basis for classification and decision making. Similarly, Kim's method uses representative examples to interpret and group data.

Recently, Zhu [94] noted that most of the existing work focuses on new methods of interpretation, rather than usability, real interpretation, and effectiveness in real users. They introduced a derivative research field called eXplainable AI for Designers (XAID) and proposed a human-centered approach to make it easier for game designers to co-create with AI / ML technology through XAID. Abdul investigates how HCI research can help end-users effectively develop practical and interpretable systems. The author has conducted a considerable analysis of the data-based literature to establish an interpretable research agenda for human-computer interaction. They also pointed to more related works that try to facilitate human understanding through interfaces, textual forms, or visual interpretation.

Among agnostic methods, visualization is the most human-centered technology. In fact, this method provides a better explanation for viewing the black box, but unfortunately, some technologies that belong to this method can produce visualizations, although visually Interestingly, your audience doesn't fully understand them. In Hohman's latest work, he recognized the importance of generating DNN visualizations and interpretations, these visualizations and interpretations are understandable and exposed to human work trying to produce such visualizations.

2.4 XAI Antithesis: Explain Or Predict

So far, we have submitted documents supporting XAI from different angles. What is different is that before presenting our comprehensive ideas, we suggest openly challenging typical methods along this axis, adjusting intuitive beliefs, and guessing previously discovered work on XAI. To seek holism, the objective here is to propose a structured proposal that respects the triad: essays, opposition, and synthesis. In his paper "Is Interpretation Always Important?" Bonter asks questions about the importance and subsequent use of interpretation techniques in the system as soon as possible to help users make low-cost decisions. Based on their research, they discovered that despite the lack of meaningful or easy-to-understand explanations, these opaque smart systems are still generally considered positive. They noted: "Although some users are interested in accessing more information, the main answer is that the application is transparent enough, or the cost of viewing the instructions will exceed the benefits." Therefore, XAI is not ready to accept entry into this type of smart systems market. "Too much, too little, or just correct?" It is a document proposed by Kulesza [95] in which they presented their results on how explanations affect the mental models of end-

users. Interestingly, they suggest that the integrity of the explanation is more important than the robustness: increasing integrity through certain types of information helped the user's mental models and, surprisingly, the perception of the profitability of participating in the explanations. They also found that, contrary to the philanthropic explanation, we believe that oversimplification could be a problem: "If robustness were too low, the user would experience greater mental demand and lose confidence in the explanations." This reduces the likelihood that users will heed such explanations. "

Although Holliday's work [96] entitled "User confidence in intelligent systems: a journey through time is work", it is confirmed on the basis of experimental studies that explain the effects of trust. This paper questions the typical approach that takes trust into account: smart systems are only recorded as a key quantitative figure at the end of a task. Most research papers on machine learning have proposed and contributed to a stricter interpretation of interpretability. In stark contrast to this wave of thought, Offer suggested in his work "I know when I see it" that the shortcomings of intuitive concepts of interpretability should also be better understood. This is that we must interpret accurately, not use intuitive considerations to determine where it is damaged. Wang [97] proposed in their work "Commercial Interpretability for Precision" a sparingly oblique additive Models that sacrifice a certain degree of interpretability for precision to achieve completely sufficient precision. From a statistical point of view, Shmueli [98] debated " A Explain or predict the dilemma? " Giving special emphasis to the field of machine learning.

Finally, using machine learning as a model, where prediction is more important than explanation, Yarkoni and Westfall [99] believe that, in psychology, "more attention to prediction than explanation can eventually lead to a better understanding of behavior". Call your work: "Pick Predictions for Psychological Interpretation: Lessons from Machine Learning."

CHAPTER 3

Datasets

For this thesis we will use some data sets that will help us in the future to be able to discern between the models that will be seen in the next section. For this, we have thought about choosing data sets that contain the amount of information necessary so that the treatment of the data is the best possible and thus avoid biases in the value of the information of the data. In general, the more data we collect, the better, the information we collect is free and can be found in the UCI machine learning repository.

For the thesis in which it is proposed to explain artificial intelligence and how to arrive at decision-making, it is not prudent to use data sets with little information, since these data sets represent those that have little theoretical understanding and less practical experience. Instead, we would like to exploit those data sets that have been used to generate knowledge about the decision-making process (or in the case of specific decision metrics), and this can be achieved from some concrete examples. This process can be modeled as a series of decision-making steps that have certain objectives, defined objectives and obtain useful information about the objective. It is possible to form knowledge about the decision-making steps from the respective objectives.

3.1 Introduction

Datasets refer to any file that contains records, so we can name any group of named records as datasets. In these datasets we can find a lot of very varied information. Being able to understand the different types of data for their treatment is a previous step that is of vital importance so that an exploratory data analysis (EDA) can be performed in this way and to be able to perform functional engineering for the AI/ ML models that these data will be implemented later. These data must be adequately treated and converted into variables so that the models can make the appropriate decisions and in this way we can have a visualization of the data more easily.

The data can be classified into 4 groups from the machine learning perspective

- Numerical Data: This type of data is any data that has exact numbers, statisticians also enter this classification, these data have a meaning of "measurement". The numerical data can be continuous and discrete. the continuous data have with characteristic that they can take any value within a predetermined range, on the other hand the discrete data have different values
- Categorical Data: This is data that represents characteristics, but it is also data that can have numerical values. This data is generally used to convert a string type data into a numeric. It still represents a category but with a numerical value without a mathematical value.
- Time series data: This type of data is a sequence of numbers collected at regular intervals over a period of time. This type of data is used especially in the field of finance, this type of data has a peculiarity and is that it has an attached time value such as a date or some time stamp that could look for trends over time.
- Text type data: These data are basically words, generally what is done to this type of data is to convert them into numbers, using functions that in most cases are interesting to use.

Depending on the data, it can be used in machine learning, and could have some repercussions, depending on the algorithm to be used for future modeling.

3.2 Cancer Wisconsin (Diagnostic) DataSet

Artificial intelligence is used specifically in bioinformatics, especially in the diagnosis of breast cancer. Will use learn how to detect mom's cancer. Cancer diagnosis is one of the most studied problems in the field of medicine. Some researchers are committed to improving performance and have achieved satisfactory results. Early detection of cancer is essential for a quick response and better chances of cure. Unfortunately, because there are no symptoms of the disease at first, it is often difficult to detect cancer early. Therefore, new knowledge must be discovered and explained to prevent and minimize the risk of adverse consequences.

Machine learning is used specifically in bioinformatics, especially in the diagnosis of breast cancer. Will use to understand the problem more precisely, we need tools to help oncologists choose the treatment necessary to cure or prevent a relapse by reducing the harmful effects of certain therapies and their costs. In artificial intelligence, machine learning is a discipline that allows machines to continuously develop throughout the process. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from the University of Wisconsin Hospital is used to classify tumors as benign or malignant.

This is one of three domains provided by the Institute of Oncology that has repeatedly appeared in the machine learning literature. This dataset includes 569 instances. Instances are described by 32 attributes, some of which are linear and some of which are nominal. The data set to be used has the following attributes. **Attribute Information:**

1. ID number.
2. Diagnosis (M = malignant, B = benign).
3. a la 32.

   Ten real-valued features are computed for each cell nucleus:

   (a) Radius (mean of distances from center to points on the perimeter).

   (b) texture (standard deviation of gray-scale values)

   (c) perimeter.

   (d) area.

   (e) smoothness (local variation in radius lengths).

   (f) compactness ($\frac{perimeter^2}{area-1.0}$)

   (g) concavity (severity of concave portions of the contour).

(h) concave points (number of concave portions of the contour).

(i) symmetry.

(j) fractal dimension ("coastline approximation" - 1).

The following table shows the general information that the data set has, in which we can see sections such as:

- The characteristic of the data set.
- The characteristics of the attributes of the data set.
- The tasks associated with the dataset.
- The number of instances the dataset has
- The number of attributes in the dataset
- Whether or not missing values exist
- The data set application area
- When the dataset was donated for open use
- The number of visits to the web

Feature type	**Value**
Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Task	Classification
Number of Instances	569
Number of Attributes	32
Missing Values?	No
Area	Life
Date Donate	1995-11-01
Number of web Hits	1239522

Table 3.1: Cancer Dataset

Below is a table that was obtained in the treatment of the data that shows us the 32 attributes that it has and that must be previously treated, in this way the different models can be used in the future, in addition to this we can mention that types of data that we have in this data are:

- For id: we have data that is of type int64
- For diagnosis: we have data that is of the object type
- For the rest of the data: we have data that are of type float64

Attributes of dataset	
id	diagnosis
radius_mean	texture_mean
perimeter_mean	area_mean
smoothnessv_mean	compactness_mean
concavity_mean	concave points_mean
symmetry_mean	radius_se
texture_se	perimeter_se
fractal_dimension_mean	area_se
smoothness_se	compactness_se
concavity_se	concave points_se
symmetry_se	fractal_dimension_se
radius_worst	texture_worst
perimeter_worst	area_worst
smoothness_worst	compactness_worst
concavity_worst	concave points_worst
symmetry_worst	fractal_dimension_worst

Table 3.2: Table of Attributes that there are in the dataset.

Now we will be able to appreciate several graphs that we have generated in the jupyter notebook to be able to carry out this work, in which we will appreciate several things that will help us understand how to clean the data, and later on they will also be useful for us to be able to do the treatment. of these data and therefore be able to apply models that help us explain how decision-making or predictions are reached.

In our first graph, what is shown is the number of malignant tumors and the number of benign tumors that we have with a simple bar graph. We can better appreciate this data, of course, we have also shown that it is interpretable with numbers and letters having as

data that benign tumors amount to 357 and malignant tumors to 212

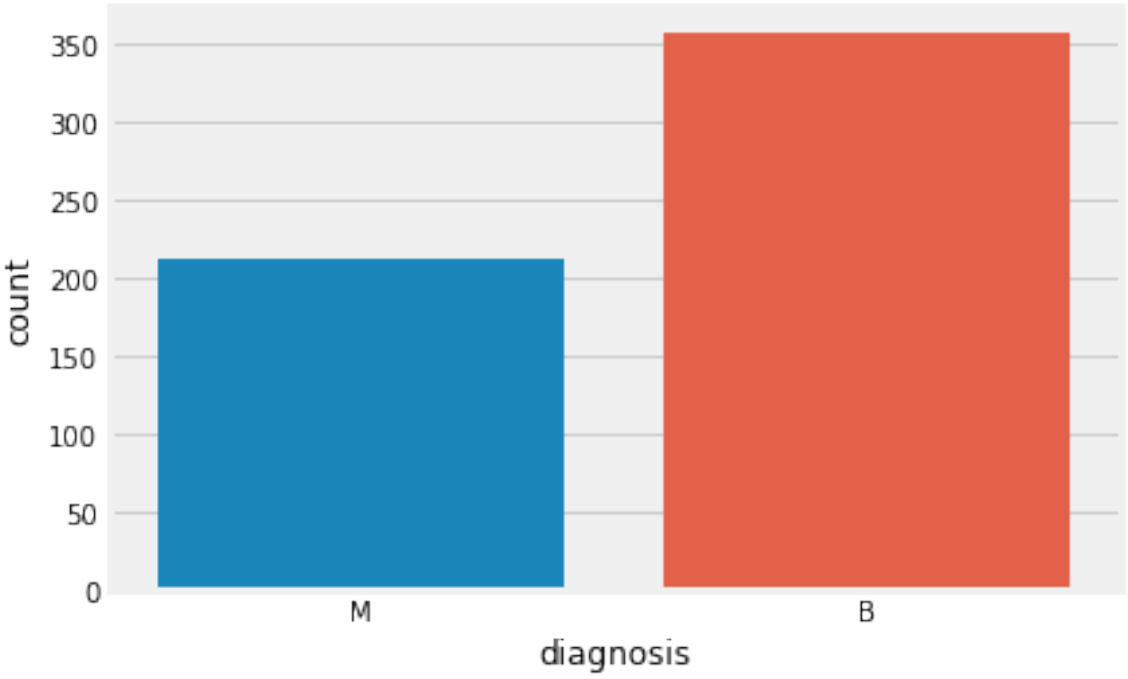

Figure 3.1: Classification of Benign and Malignant Tumors

The histogram that we will see next is a graphic representation of the variable "radius_mean" that alludes to the average of the radius that have been obtained from both malignant and benign tumors, in addition it is also observed that the surface of each bar is proportional to the frequency of the represented values.

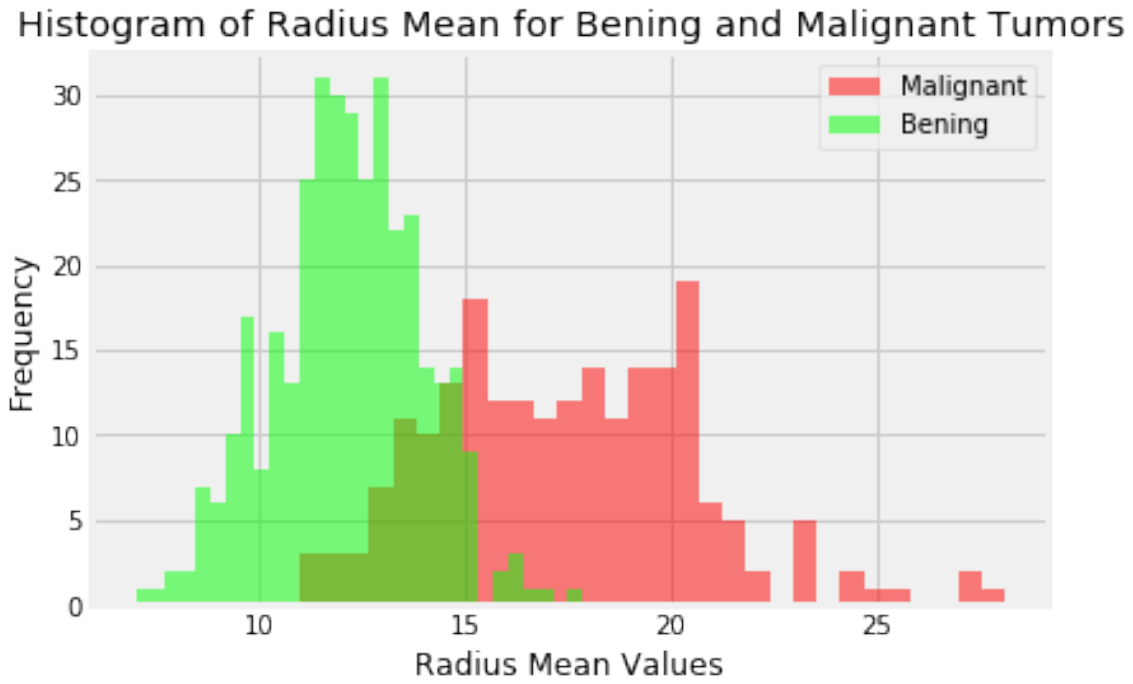

Figure 3.2: Mean Values of Radius Tumors

Since the differences between the values of the characteristics are too high to be able to better observe the plot. It was done by tracing characteristics in 3 groups and each group includes 10 characteristics so that it can be better observed.

To interpret the following figure, let's take the texture_mean characteristic as a reference. In addition, we can see that there is a separation of the median between malignant tumors and benign tumors, so we can assume that it would be a good characteristic for our future plans, not with the fractal_dimension_mean characteristic, We can see that the median of malignant and benign does not seem separate, and we cannot say that it provided us with good information in the future.

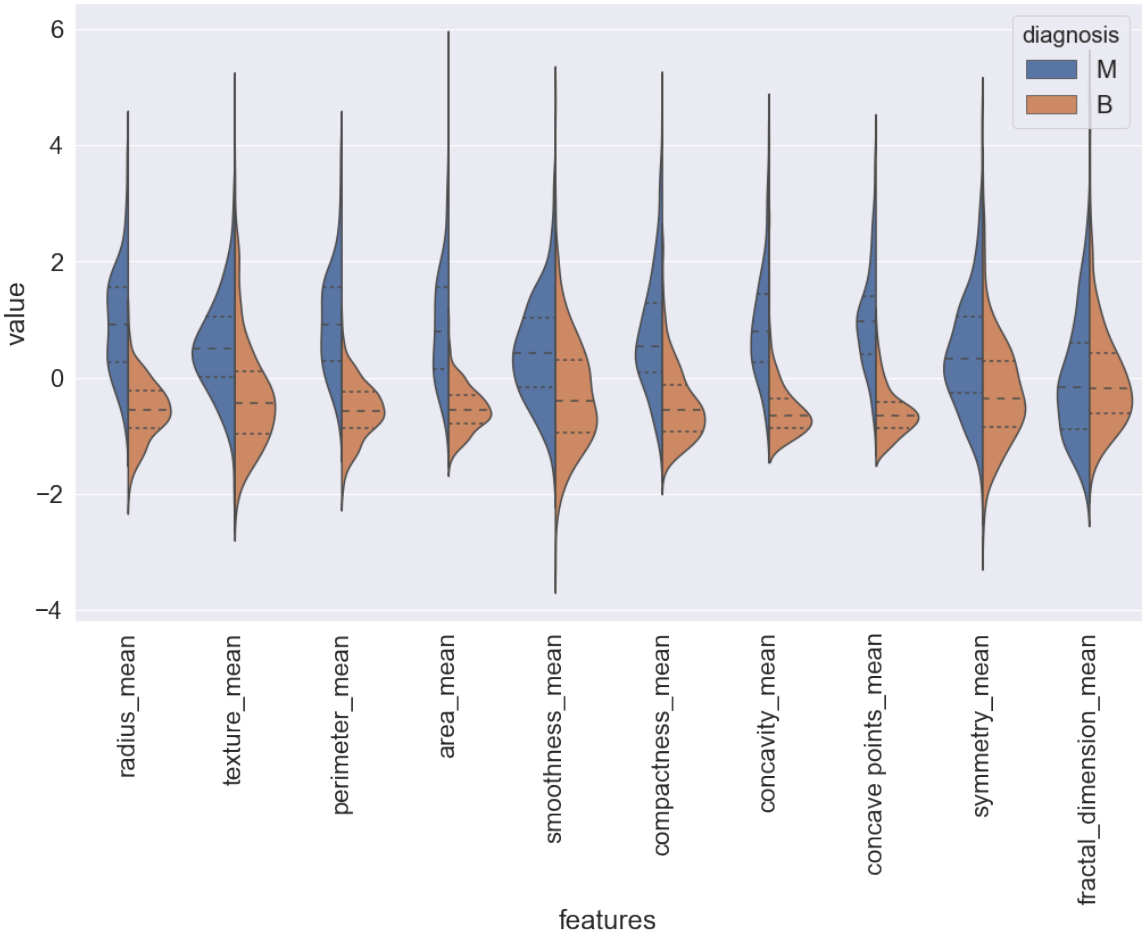

Figure 3.3: Violin diagram of the first 10 features

Continuing with the work, two more violin graphics will be shown that will serve as a reference to know which features are more or less important, to be treated and thus help us in the future.

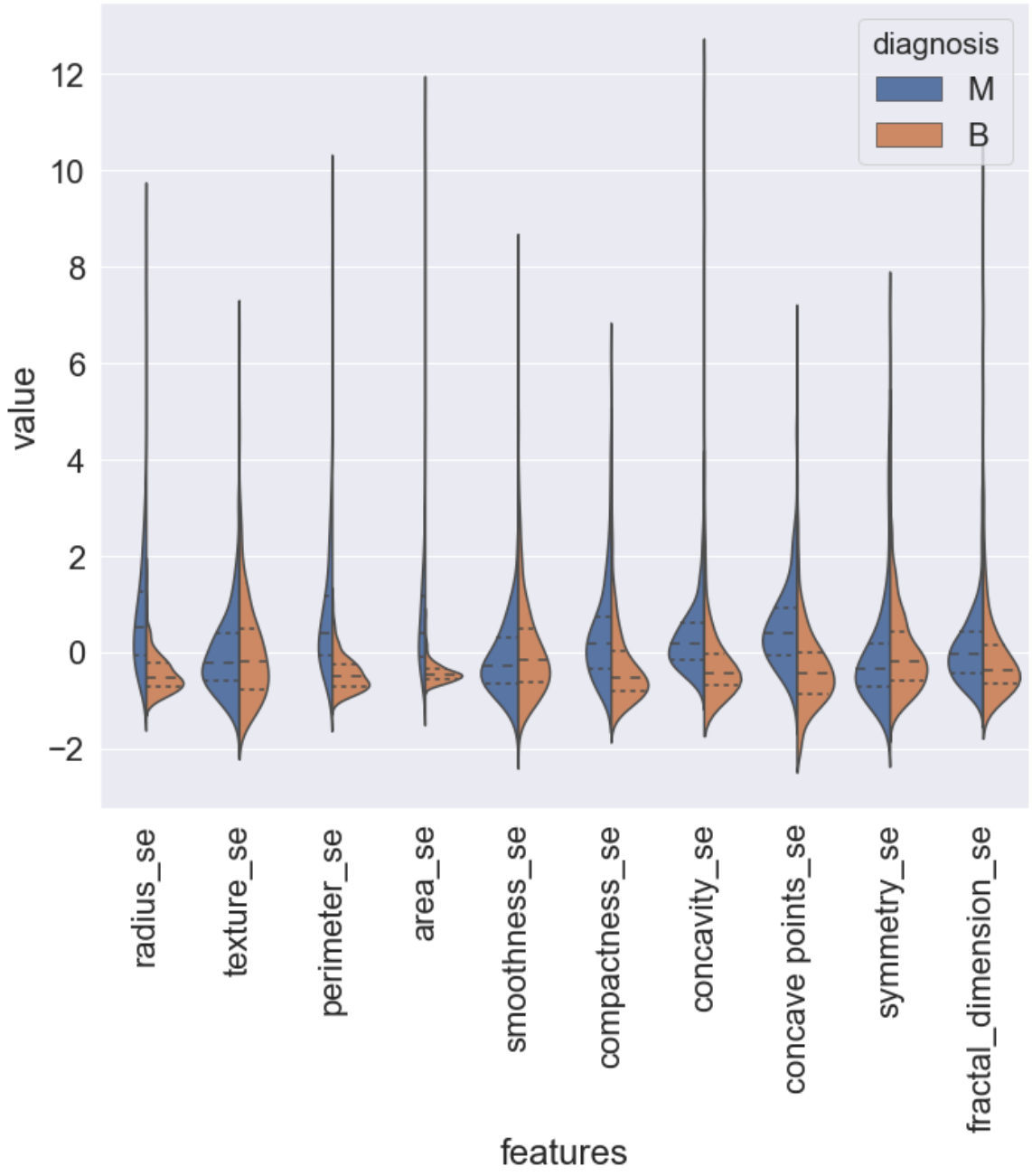

Figure 3.4: Violin diagram of the second 10 features

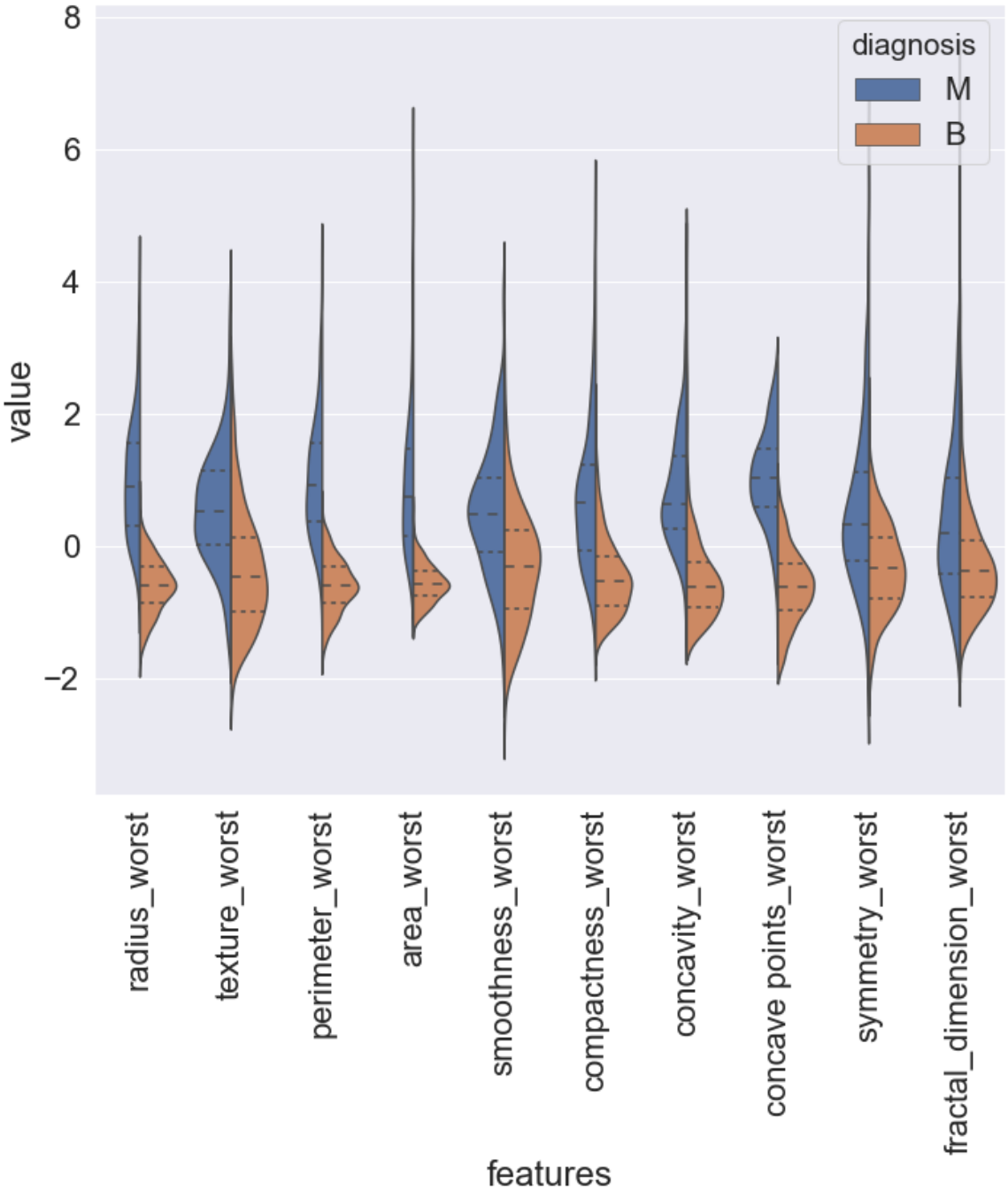

Figure 3.5: Violin diagram of the latest features

As we can see in the last graph specifically, there are two features that are very similar (concavity_worst and concave points_worst), what we will do next is interpret using another graph if these two features not only look alike but we will also decide if they are correlated among themselves, the latter is not always the case, but if the features are correlated with each other, we can discard one of these features.

To make the comparison of two characteristics more in depth, we will use an assembly diagram.

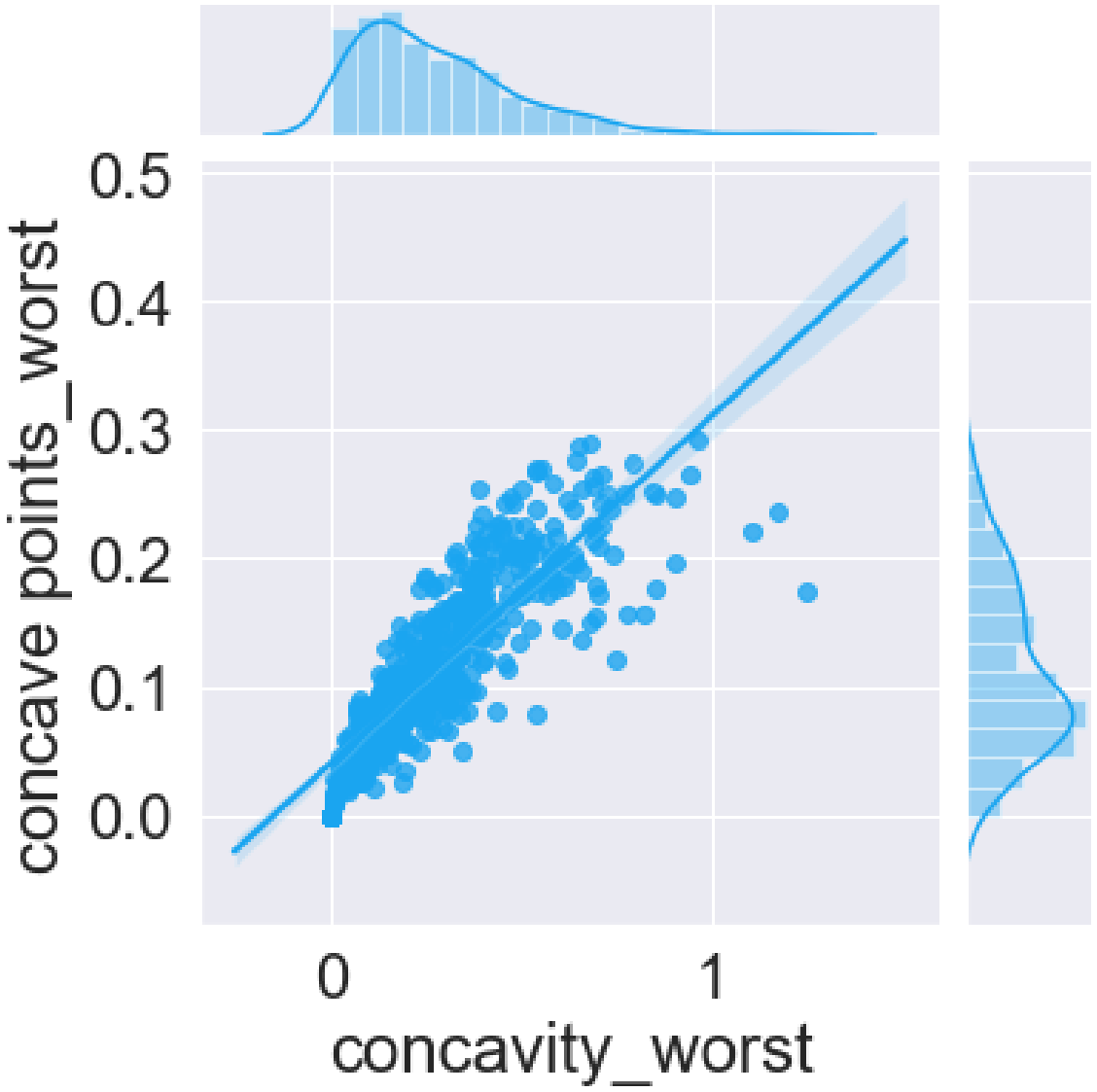

Figure 3.6: Assembly Diagram to compare features

As we can see, these two features are directly correlated. The Pearsonr value is the correlation value and 1 is the highest. and as we can see the value we obtain is 0.86 and with that data it is more than enough to say that they are correlated. We must not forget that we are not yet making a choice of the functions, we are just looking to have an idea about them.

We can use other techniques to know what others features are correlated but not only using two features only, if we want to do this but with more features we can use pair grid plot, The following graph shows us how this was done and we were able to determine that we have described that radio_worst, perimeter_worst, and area_worst are mapped to. We definitely use these discoveries for feature selection.

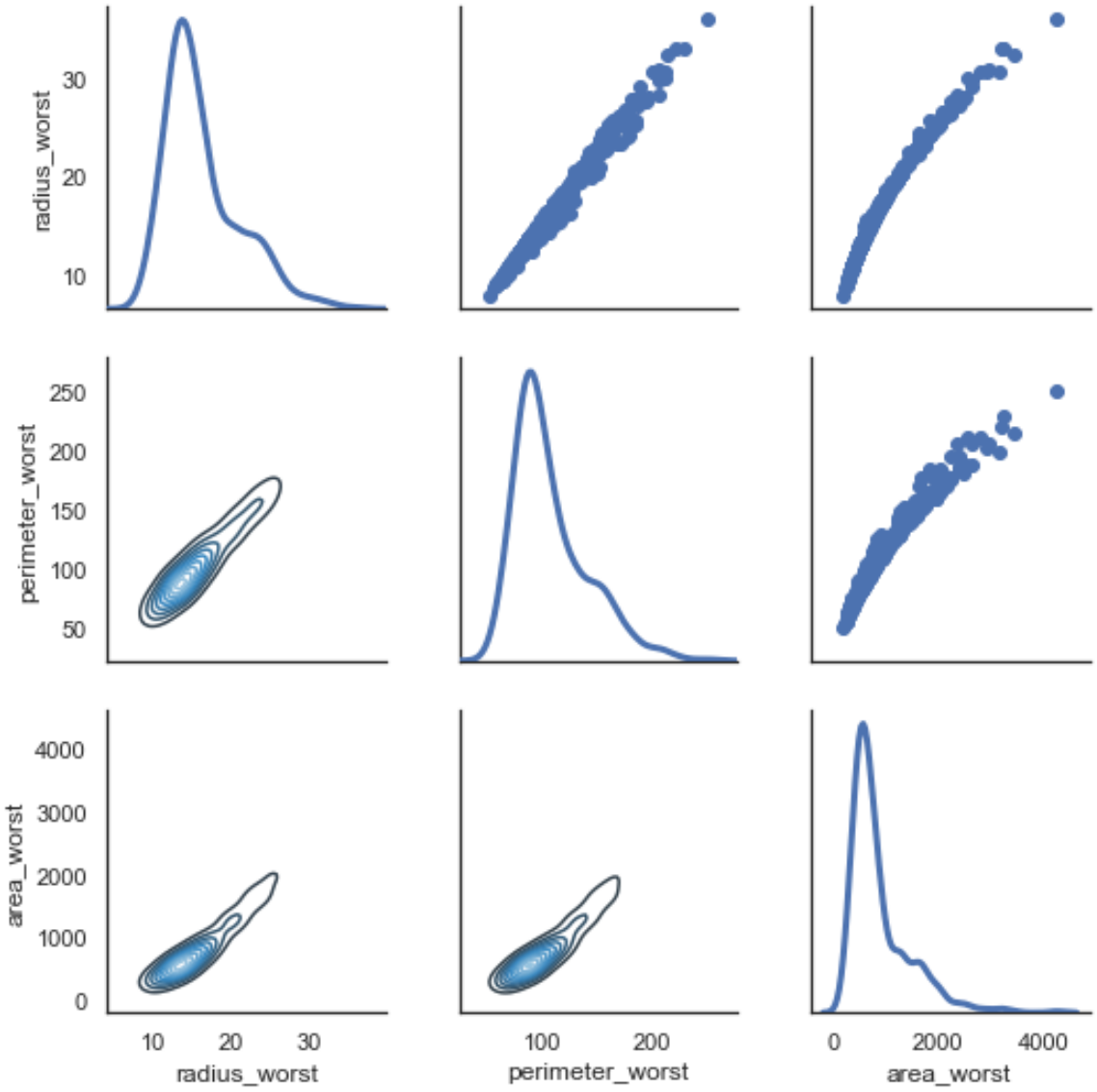

Figure 3.7: Pair grid Diagram of features

3.3 Red Wine Quality DataSet

Another field in which AI is used very frequently is in the field of oenology, which is the science or art of producing good quality wine, the dataset contains information that, when treated by a model or several models of AI You can determine if a wine is of good or bad quality. This assumes that the model becomes a pseudo expert who is responsible for analyzing these characteristics. This would greatly help vineyard owners and wine consumers, in order to guarantee in some cases that the wine is of good quality and thus be able to certify it, and in other cases as a consumer it can be said that these wines they have acquired are from the best quality for consumption. since in the field of gastronomy a good wine makes the difference between consuming a quality meal or not, but not only that. Several studies have shown that the consumption of red wine is capable of protecting the heart and it is also possible to control cholesterol levels among other benefits. These benefits have been attributed to the alcohol levels and antioxidants in the grapes. Unfortunately, not all wines are created equal. Wines like Malbec have a lot of resveratrol. These types of wines come from grapes with thick skin and other famous wines like Cabernet Sauvignon have many antioxidants in their compound. Once both the commercial benefits and the health benefits have been considered, we have taken into consideration this dataset that we have taken from the UCI machine learning repository, from here we can obtain physicochemical data, this data set contains 1600 instances. instances are discovered using 12 attributes. the attributes that the dataset has are the following.

Attribute Information:

1. **Fixed Acidity:** Acid is the main ingredient in wine and greatly improves the taste of wine. In fact, acid imparts acidity, which is the basic characteristic of wine flavor. Acid-free wine is "flat". Chemically.

2. **Volatie Acidity:** The main volatile acid in wine is acetic acid, which is also the main acid related to the smell and taste of vinegar.

3. **Citric Acid:** Citric acid is generally added to wine to increase acidity, complement specific flavors, or prevent iron damage. It can be added to finished wine to increase acidity and impart a "fresh" flavor.

4. **Residual Sugar:** It is given from the natural sugars of the grape that remain in a wine after the alcoholic fermentation ends.

5. **Chlorides:** chlorides as a major contributor to wine salinity

6. **Free Sulfur Dioxide:** Sulfur dioxide is used in the wine creation process to control its oxidation and also prevent microbial growth.
7. **Total sulfur Dioxide:**Total sulfur dioxide (TSO2) is the portion of SO2 that is found freely in wine in addition to the portion that has been fused to other chemicals in wine.
8. **Density:** Wine density values are mainly used to have wine quality parameters.
9. **pH:** The pH mainly affects the taste of the wine, in addition to the microbial stability of the wine
10. **Sulfates:** It is the product of the fermentation process and it works as a preservative for certain yeasts and bacteria (which will quickly destroy a wine if they start to multiply)
11. **Alcohol:** The ethyl alcohol (CH3-CH2OH) that I find in wine is given by the fermentation process of the grape
12. **Quality:**score between 0 and 10.

The following table shows the general information that the data set has, in which we can see sections such as:

- The characteristic of the data set.
- The characteristics of the attributes of the data set.
- The tasks associated with the dataset.
- The number of instances the dataset has
- The number of attributes in the dataset
- Whether or not missing values exist
- The data set application area
- When the dataset was donated for open use
- The number of visits to the web

Feature type	**Value**
Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Task	Classification, Regression
Number of Instances	1600
Number of Attributes	12
Missing Values?	No
Area	Business
Date Donate	2009-10-07
Number of web Hits	1238000

Table 3.3: Wine Dataset

Below is a table that was obtained in the treatment of the data that shows us the 12 attributes that it has and that must be previously treated, in this way the different models can be used in the future, in addition to this we can mention that types of data that we have in this data are:

- Most of the data: we have data that are of type float64
- For Quality: we have data that is an int64 type

Attributes of dataset	
Fixed Acidity	Volatile Acidity
Citric Acid	Residual Sugar
Chlorides	Free sulfure dioxide
Total sulfur dioxide	Density
pH	Sulphates
Alcohol	Quality

Table 3.4: Table of Attributes that there are in the dataset.

Next, and in the same way as with the previous dataset, several graphics that we have generated in the jupyter notebook will be shown to be able to carry out this work, in which we will appreciate several things that will help us understand how to clean the data, and later they will also be useful. so we can do the treatment. of these data and, therefore, be able to apply models that help us explain how decision-making or predictions are reached.

In our first graph, what is shown is the quantity of good quality and poor quality wines that we have with a simple bar graph. We can better appreciate these data, of course, we have also shown that it is interpretable with numbers and letters with data that good wines are few and amount to 217 and poor quality wines to 1382

It should be noted that in this graph you can see values of 1 and 0 this has been done thinking that the data shown are the values equal to 0 equal to poor quality wine and the values equal to 1 are wines of good quality

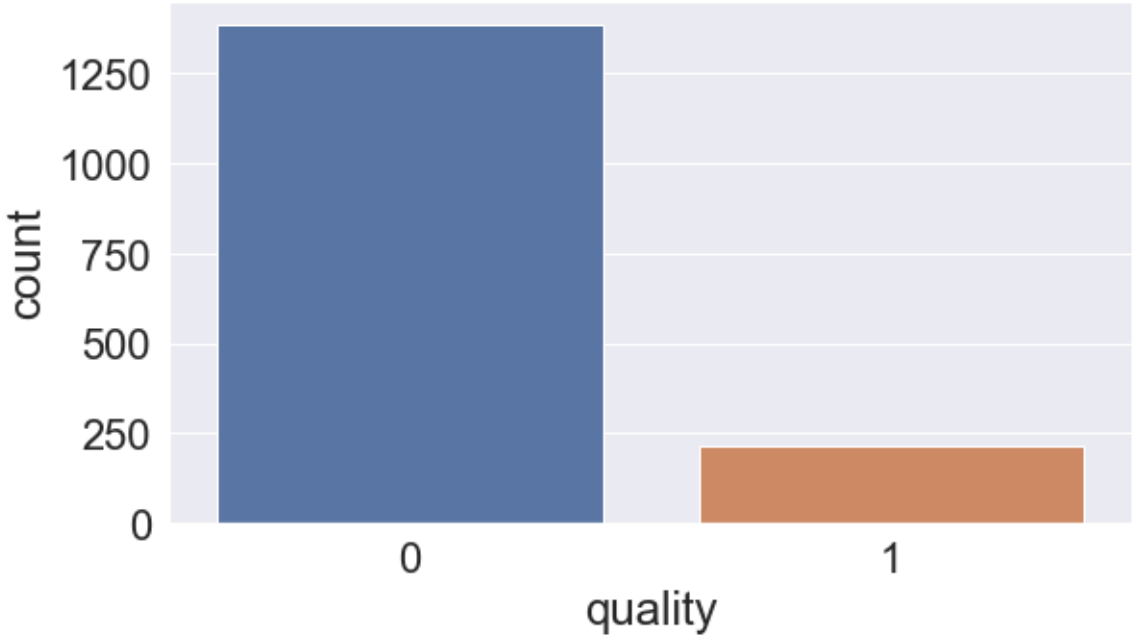

Figure 3.8: Classification of Good Wines and Bad Wines

The next figure shows the Distributions of the features that we have in our dataset. The distribution of features helped us to understand what type of characteristics we were dealing with, in addition to what values could be expected from these characteristics. This helped us to understand what type of characteristic they were treating and what values can be expected to have from these characteristics. After that it was seen if the values were centered or scattered. As we can see, the characteristics, such as 'sulfur dioxide' or 'sulfates', have a correct and skewed distribution. With this graph we have a general idea of all the features that our dataset has and with the data that we work with.

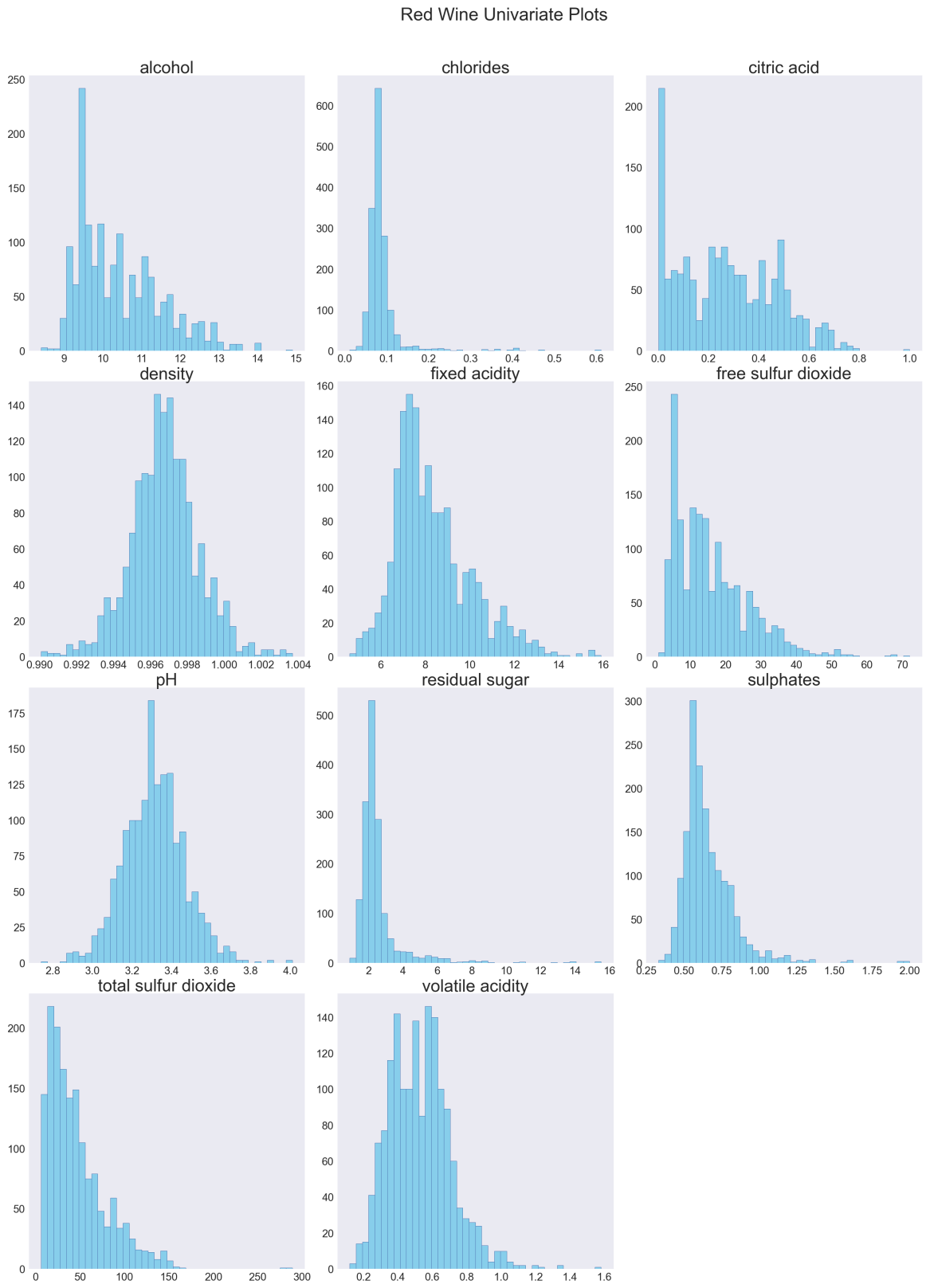

Figure 3.9: Diagram of Distribution of the Features

In this data set we do not have many features as in the previous one, and that is why we can have all the data represented in a single plot, in this way we can see the features so that it can be better observed.

To interpret the following figure, let's take the characteristic of alcohol as a reference. In addition, we can see that there is a separation of the median between the qualities that the wines have, so we can assume that it would be a good characteristic for our future plans, not with the residual sugar characteristic, we can see that the median of good wines and Bad wines do not seem to be as separate as the other characteristics, and we cannot say that it provided us with good information in the future. although it will not be a feature that we are not interested in said feature

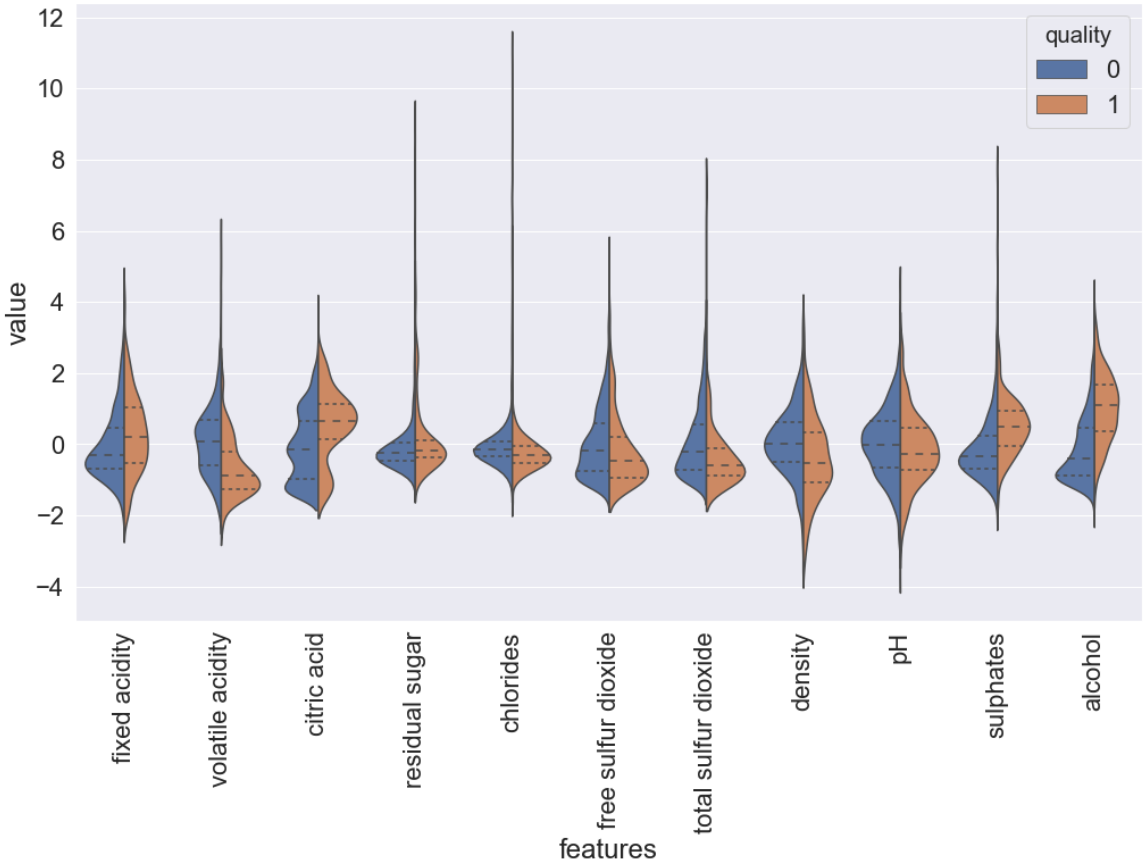

Figure 3.10: Violin diagram of the Features

As we can in the graph, there are two characteristics that are very similar (density and pH), what we will do next is interpret using another graph if these two characteristics not only look alike, but we will also decide if they are to correlate between them, the latter This is not always the case, but if the characteristics are correlated with each other, we can rule out one of these characteristics.

To make the comparison of two characteristics more in depth, we will use an assembly diagram.

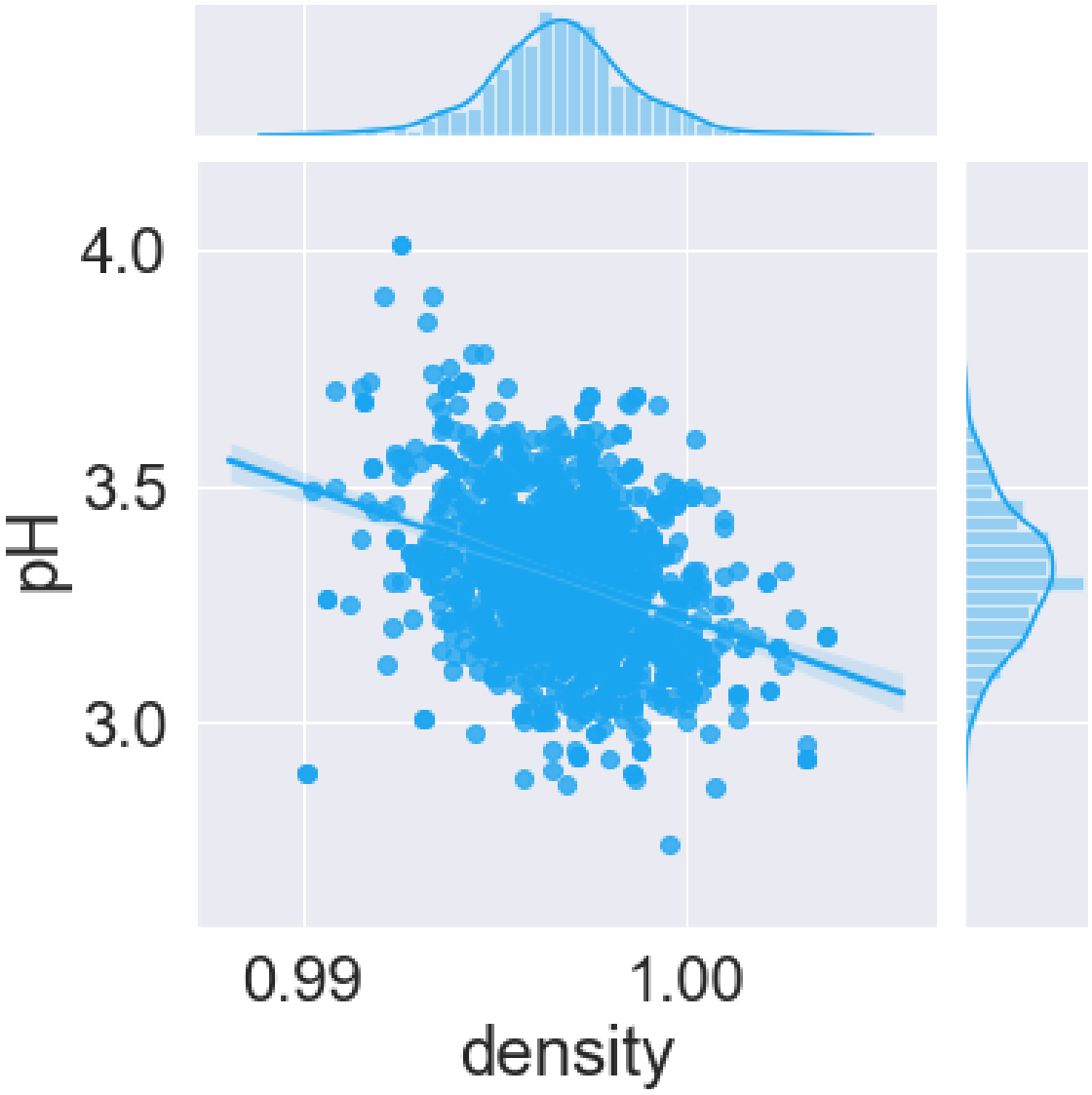

Figure 3.11: Assembly Diagram to compare features

As we can see, these two characteristics are directly correlated. the Values that mark the correlation are very close to 1 and that would be enough to be able to say that the characteristics shown are correlated. We must not forget that we are not yet choosing the functions, we just want to have an idea about them.

We can use other techniques to know what others features are correlated but not only using two features only if we want to do this but with more features, we can use pair grid plot, The following graph shows us how this was done and we were able to determine that we have described that residual sugar, sulphates, chlorides are mapped to. We definitely use these discoveries for feature selection.

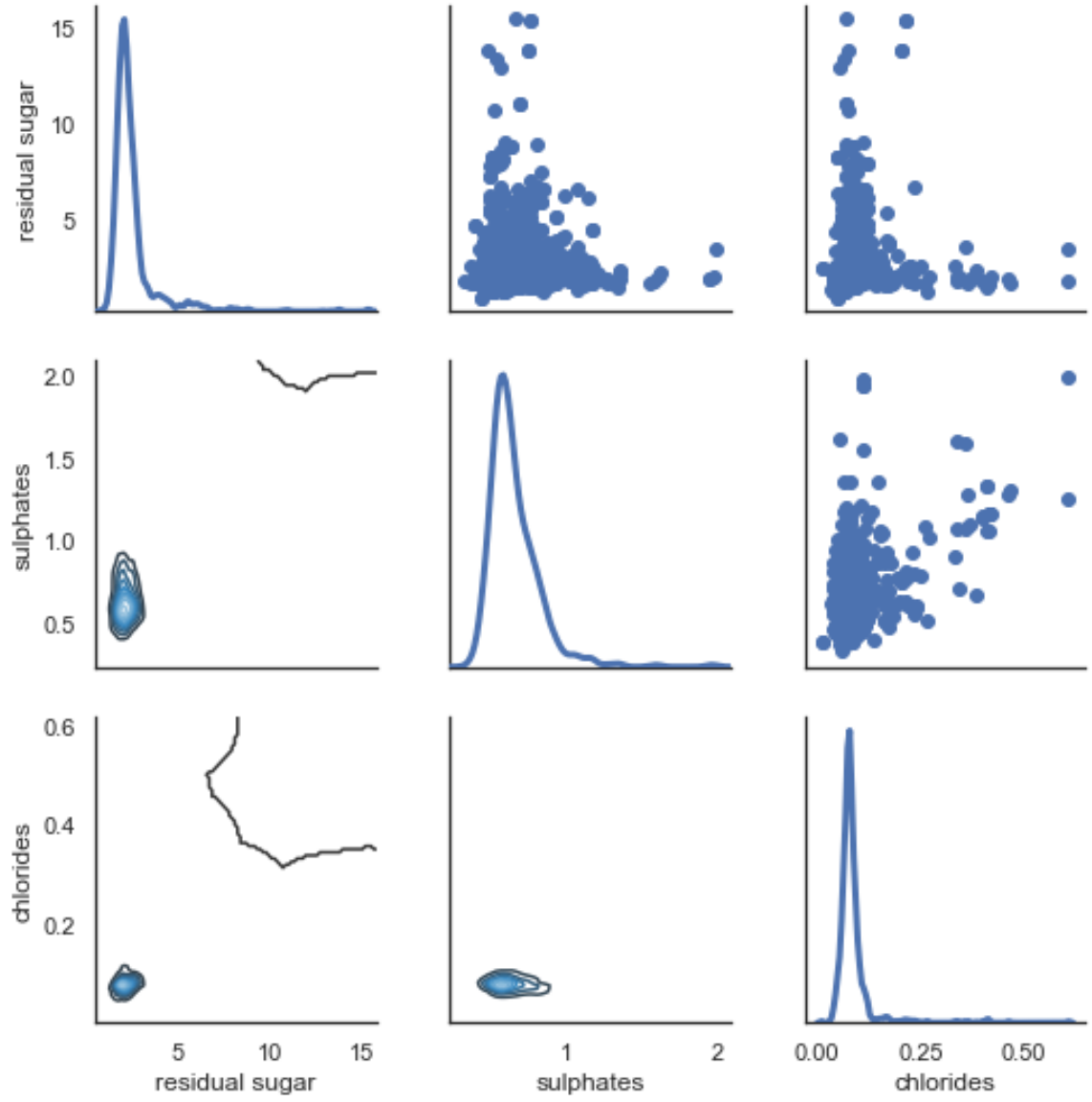

Figure 3.12: Assembly Diagram of Distribution of the Features

As we have shown in this chapter on datasets, a proper understanding of the characteristics we have in each of them will be of great importance so that in later chapters we can better understand how each of the models is shared. When working with these data, and even more, explain with other techniques how each model drew its respective conclusions.

CHAPTER 4

Comparative Analysis of Current Trends

In this Master's Thesis, we have applied AI algorithms so that each model that has been applied to each dataset can make a decision or make a classification based on the characteristics of each of the datasets. For the model to have been able to function properly, what had to be done was to make a correct treatment of the data that we showed in the previous chapter very briefly.

In this chapter, you will see the AI models that have been used once the data has been processed and then apply explanatory techniques to the models used, in this way we can have an idea of what we have at our disposal to explain how are those AI models make decisions or classify data.

4.1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML), as previously emphasized, can help not only in improving business operations but can go further in science or medicine. But this is not so easy to achieve. For this, the techniques are divided into two large groups, the first supervised learning and the second large group the unsupervised non-learning. now well within the first group, we also have classification models and regression models. And the second large group unsupervised learning is based on clustering or grouping algorithms, that is, the data is not labeled so we do not know if it has a certain category. Now with our data previously mentioned in the previous chapter, what we did was look for a function that helped us to represent these data in this way, we managed to generalize them in the best way for new data.

This chapter will consist of two fundamental sections: in the first Section, A description will be made of the models that have been used, the same for each dataset, to know how each model behaves according to its methodology with the data we have from our datasets. In the second Section, a description will also be made of the techniques that were used to explain how each model made one or the other decision or made one or the other classification.

4.2 Artificial Intelligence Models to Use

As it is already known, all the AI/ML models are oriented to learn a certain function (f), which gives us a more precise correlation between the input values (x) and the output values (y), being able to say that y = f (x). When we want to implement artificial intelligence models to do some mapping between the aforementioned values (x) and (y), the result should not for any reason be equal to 100% exactly, because this would suggest that you are doing a mathematical calculation without more and without having used any aforementioned AI/ML model. This is contrasted with the fact that the function (f) that would already be trained can predict a new constant (y) also using a new value of (x), this would allow predictive analysis. There are AI/ML models that fulfill this idea, using diverse and varied approaches.

Before starting with Artificial Intelligence models, we must make construction trains and test data sets, for this and to do it in a better way, in both data sets used, the objective data were transformed into binary classification data (tumor malignant and benign tumor) in one case, and (wine of good quality and wine of bad quality) in another case, from there, the construction of our data set was made to train and test in a classic ratio of 70/30.

Once we have this previous step ready, we can start training our classification model, which in our case will be:

1. **Logistics Regression :** Logistic regression is a type of algorithm widely used when trying to implement an artificial intelligence model that can provide binary results. This means that the model can predict the result and specify one of the two categories of Y value. This function is also based on changing the weights of the algorithm, but it is different due to the use of nonlinear logical functions to transform the Y value. Result . This function can be expressed as an S line that separates true and false values. For a logistic regression to be successful, the requirements are the same as when applying a linear regression we need to eliminate the repetitive input/output samples. and, reduce as much as possible the noise that would be the data that has no value. This is a model that is generally applied due to the fact that it is fast and in addition to being very effective for binary classification.

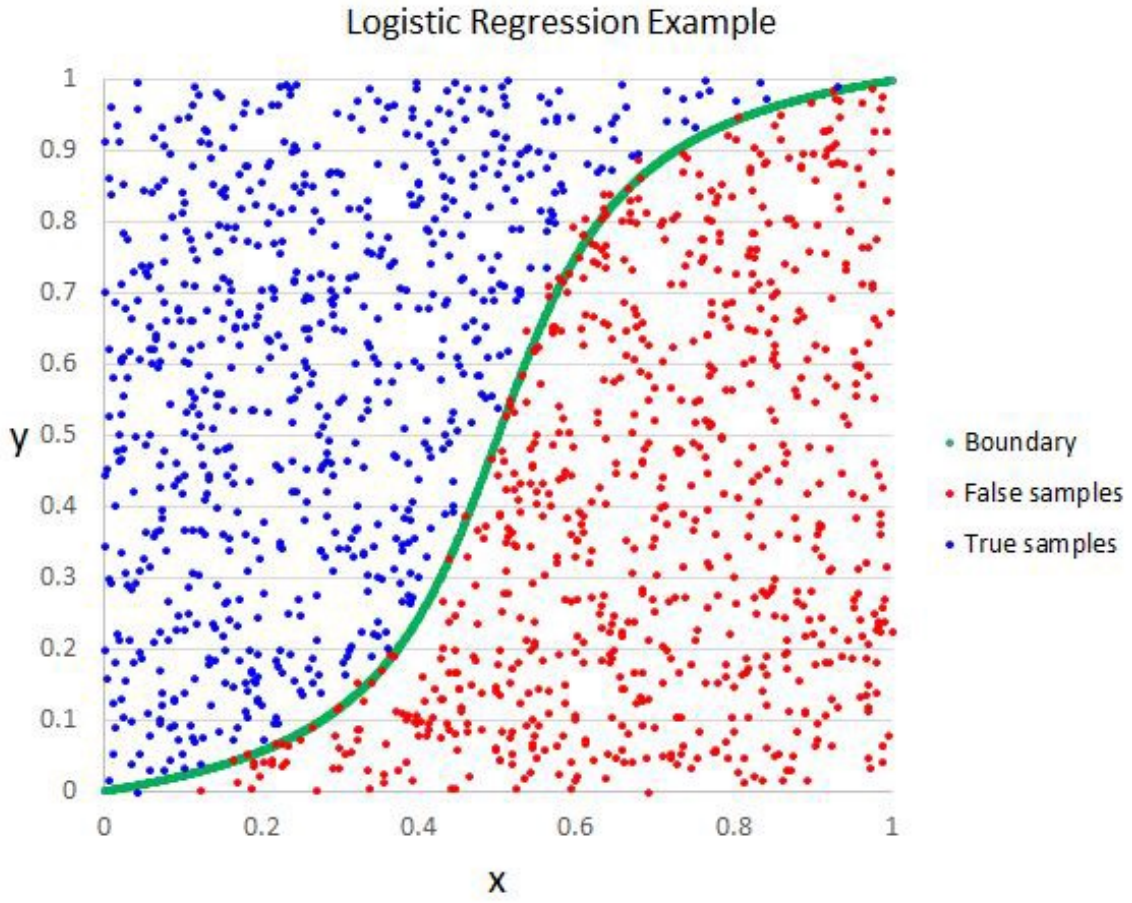

Figure 4.1: Logistic Regression Model [1]

2. **Naive Bayes:** The Naive Bayes algorithm is an AI/ML model that could be simple, but it is a model that has very strong characteristics in terms of solving problems with varied and complex qualities. and this is achieved thanks to the fact that it can calculate 2 types of probabilities

   - A chance for each class to appear
   - The conditional probability of the independent class, because there is an additional X modifier.

   This model is named "Naives" because it works by making assumptions, and those assumptions are that the input data values have no relation to each other. This is certainly not true in the real world under almost any circumstances, this model is applied to a multitude of standardized data, in order to predict or make predictions with a high degree of precision in their results.

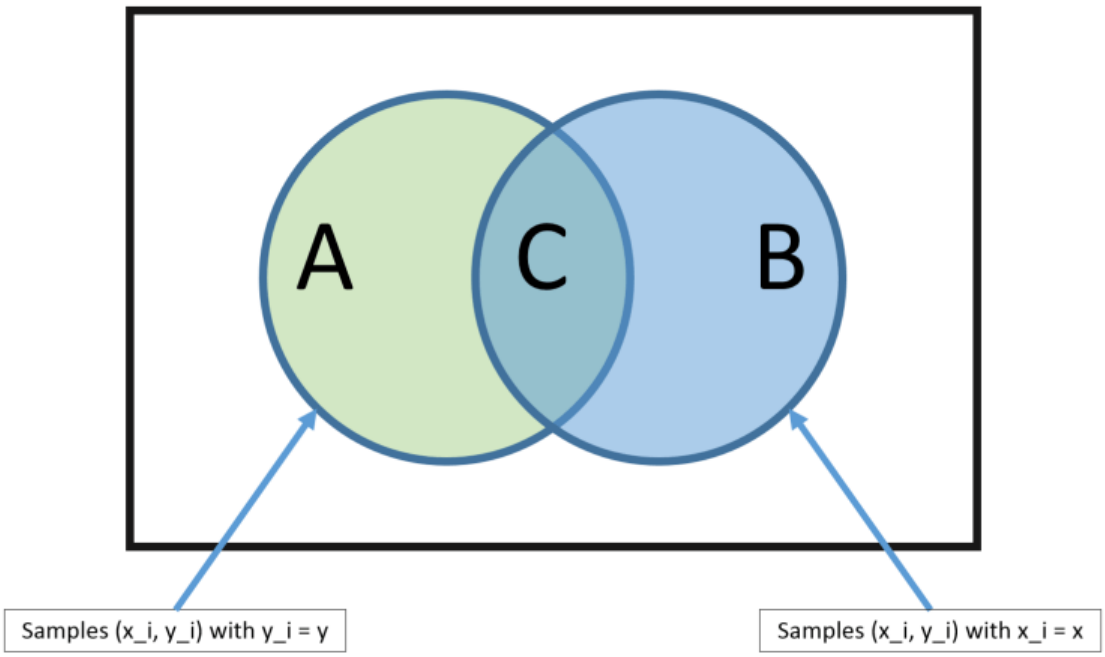

Figure 4.2: Naive Bayes Model [2]

3. **k-closest Neighbors:** This model is also a very powerful tool that is available in Artificial Intelligence, it uses the entire training data set. The way it works is that the decisions or predictions it makes are calculated by analyzing and reviewing the entire data set for the data nodes K with simple values. These nodes are called neighbors and to know the resulting value, use the Euclidean number that can be easily calculated according to the differences in value. These data sets require a very large amount of

computing resources, in this way it stores and processes the data, if the attributes are varied and must be constantly reviewed, the model has a precision run. Otherwise, it is a model that works very well, it is fast and precise, being efficient in finding necessary values from large data sets.

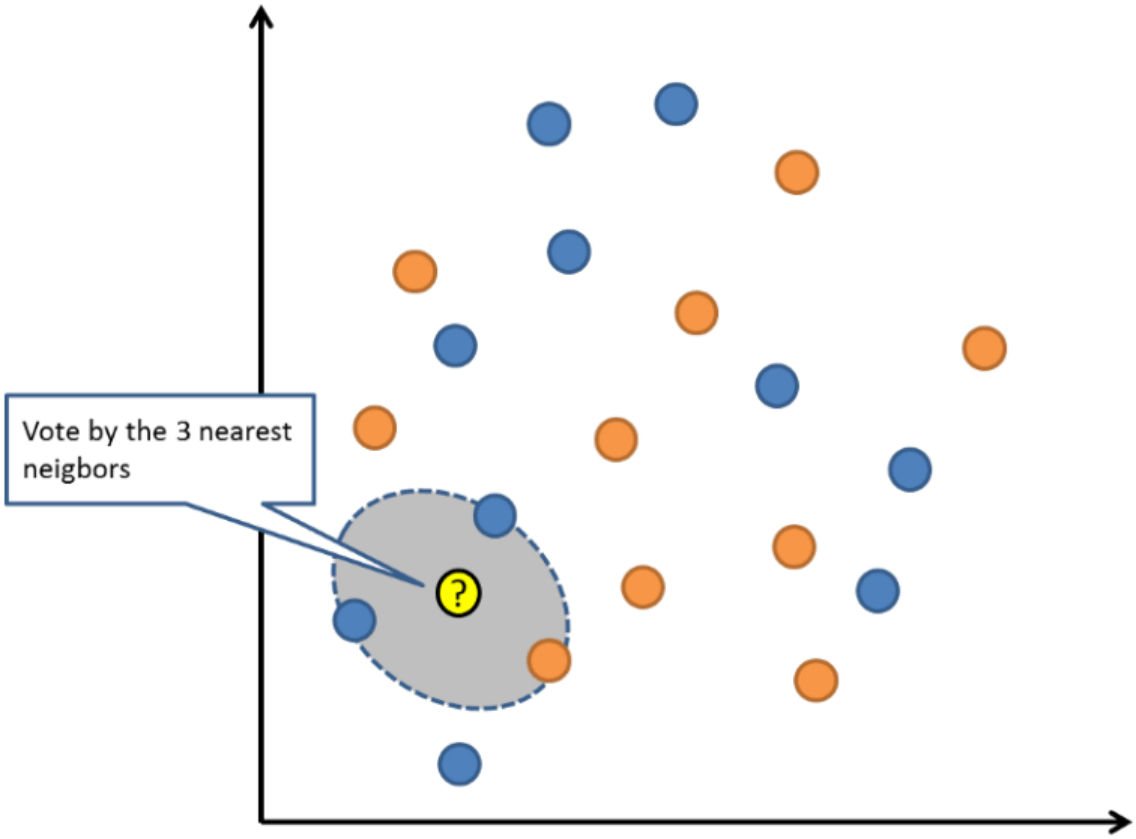

Figure 4.3: K-closest Neighbors Model [3]

4. **Random Forest Classifier :** The Artificial intelligence model random decision forests have as a component several decision trees, and for this, each decision tree has the task of processing multiple samples of the data and its results are saved as a collection of many samples in a bag in this way a more accurate output value can be found.

   In this model, we have several optimal sub-routes instead of a single optimal one, in this way we guarantee a much more precise general result if we work with decision trees and this model already solves the problem we have, random forests are a better approach that is taken into account to provide an even better result.

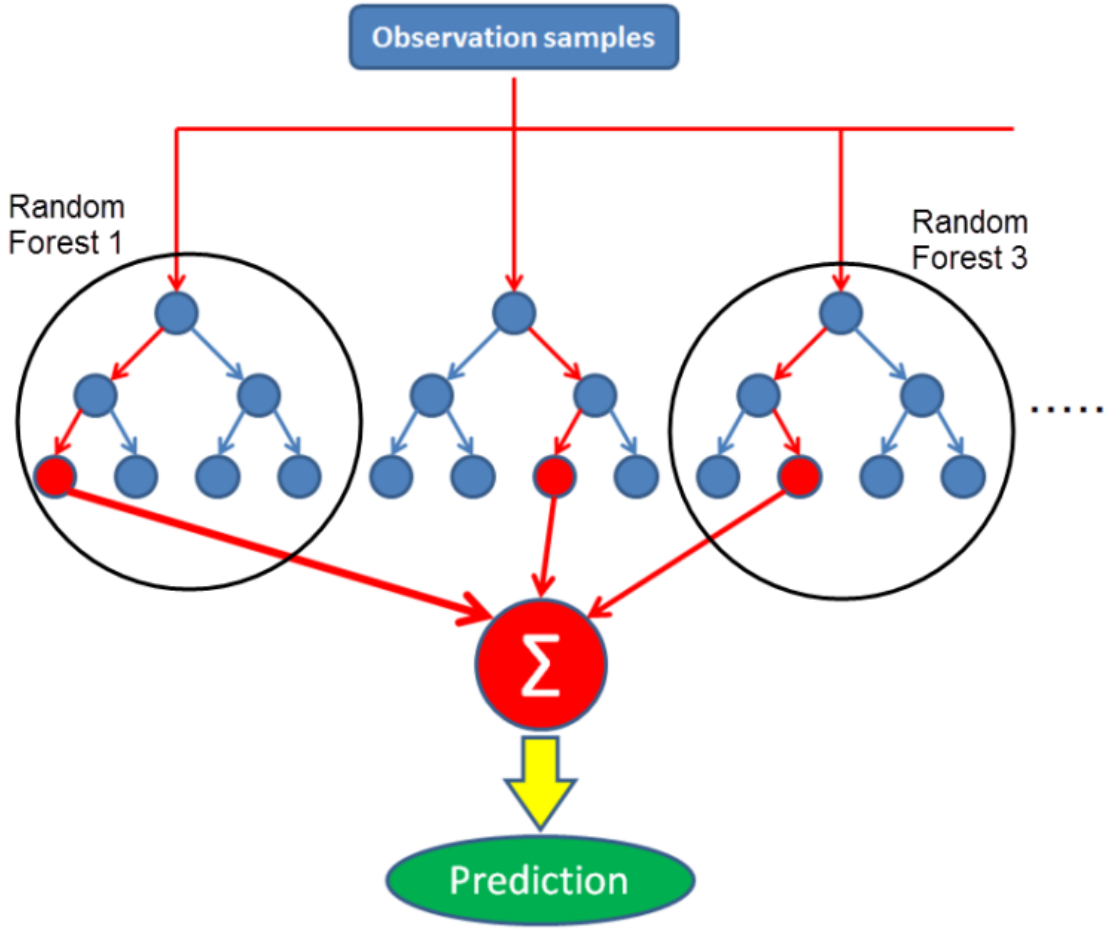

Figure 4.4: Random Forest Classifier Model [4]

4.3 Techniques for Explaining Artificial Intelligence Models

Trying to explain artificial intelligence is a job that has come from the moment machine learning is implemented for the first time. Over time, techniques have been developed to explain how a model makes its decisions. In this work, we have focused on the following explanation techniques, so that in this way we can make a comparison of how each model has behaved. The techniques we have used are the following:

4.3.1 ELI5

ELI5 is a Python software package that helps us debug machine learning classifiers and thus be able to interpret their predictions. Supports the following machine learning frameworks and software packages [104]

- scikit-learn

ELI5 is able to work with the tools that scikit-learn word processing has and in this way it could highlight your data. It is also possible to debug scikit-learn pipes that contain Hashing Vectorizer, undoing the hashing [104]

- XGBoost.
- LightGBM.
- CatBoost.
- CatBoost.
- Sklearn-crfsuite.
- Keras.

Inspection and debugging are easy for some classifiers and difficult for other classifiers. It's not a big deal getting coefficients from a linear classifier, associating them with feature names, and then displaying them in an HTML table. The goal of ELI5 is not just to deal with simple cases but even for simple cases, it is also valuable to have a unified API for inspection, so we can:

- make a call to a function that already exists in ELI5 and in this way get an instantaneous and well formatted result

- Format codes can be reused between artificial intelligence frameworks.
- You can reuse "drill down" code, like the function filter or text highlighting.
- ELI5 will handle many traps and minor differences.
- Algorithms like LIME try to interpret black-box classifiers through simple, interpretable classifiers that are locally adapted. This means that each additional "simple" classifier/regressor algorithm (such as LIME) automatically supports more options.

4.3.2 Partial Dependence Plot (PDP)

PDP shows that the relationship between the target and the input variable is linear, monotonic, or more complex. Its structure is very simple. In this technology, it is important to highlight two things:

- It does not matter which algorithm is used: random forest, neural network, or any other model since it is an agnostic method of the model.
- This is a global approach: it is built using all cases or a representative sample of them. In this way, technology can explain how variables affect the overall goal, which is different from other methods that focus on understanding the relationship between the model and the individual

The importance of variables depends on the algorithm and its parameter configuration. PDP graph is calculated only after the model has been adjusted. The model fits real data. This actual data can differ in innumerable ways, but after the model is adequate, we could start by taking all the characteristics of a single variable, then we use the model to make the prediction and we can modify the aforementioned variable.

It is also possible to visualize the partial dependence of two characteristics at the same time. This is known as **Bivariate Partial Dependence Plot (BPDP)**

4.3.3 Individual Conditional Expectation (ICE) Plot

The Individual Conditional Expectations (ICE) graph shows a line in each instance, showing how the instance predictions change when the function changes. The ICE graph shows the predicted dependency of each instance of the function separately. Compared to the regular line in the partial dependency graph, each instance produces only one line

There is a problem with ICE charts: Sometimes it can be difficult to know if the ICE curve differs from one individual to another because they start with different predictions. A simple solution is to concentrate the curve at a certain point in the function and only show the predicted difference up to this point. The resulting graph is called the centered ICE graph (c-ICE). Setting the curve at the bottom end of the function is a good option.

Another way to make heterogeneity visually easier to detect is to look at the various derivatives of the prediction function for the characteristics. The resulting graph is called an ICE-derived graph (d-ICE). The derivative of the function (or curve) tells you if the change has occurred and the direction of the change. Using the exported ICE graph, it is easy to discover the range of feature values, in which (at least some of) the instance's black box predictions will change.

4.3.4 SHAP (SHapley Additive Explanations)

SHAP is a game-based theoretical method that you can use to explain the results of various machine learning models. Its function is to use the classical value of Shapley's game theory and its related extensions to link optimal credit allocation with local interpretation. [100]

What shap does is, question each features how much was the contribution for the model to make a decision, then an average of all the permutations of each feature is taken and in this way an individual contribution is obtained, then to give a Shapley definition is the calculation of the average of the marginal contributions in all the permutations. Using shap has benefits that will be explained below:

- First, we have global interpretability. With the group SHAP values, it can be shown what value each predictor contributed, without taking into account that the collaboration was positive or negative.
- As a second benefit, we have local interpretability. It is logical to suppose that each observation will have its own SHAP table of values, what this does is that the technique increases its transparency.
- And as a third benefit, the SHAP explaining technique can obtain calculated values based on any tree model.

CHAPTER 5

Results

In this master's thesis work, activities have been completed from the choice of datasets, which allow us with the help of your data, and by applying artificial intelligence models, which we have also chosen, and after that apply interpretation techniques of those models. In order to get to this point of work, the moment when we will present the results we have obtained.

After having applied AI algorithms for each dataset, visible results could be obtained that we will present in this chapter, as well as for the techniques that explain how the models have made their decisions.

5.1 Introduction

In this chapter the results of everything we have applied and obtained in this working time are presented, the readers' ease is preselected several graphs explaining and comparing the results that were obtained after applying the models and techniques that were registered in the previous chapter.

As a first point, we will present the results obtained by applying the machine learning models that were applied to the datasets in Chapter 3, here we can see how each model has made a decision or made a classification as such. This is important since not all models work in the same way and not all will give us too optimal or poor results either, but it is necessary to know and see what each model did.

The next step was to apply explanatory techniques to these previously mentioned models, for this we have decided to use the most visual models possible but which in turn help us with the objective of comparing some models with others. The models when behaving in a different way with each other will not always give us similar or similar graphs, they all work in a different way but to notice just that, a variable will be taken, which stands out the most, and it will be used for all the techniques, this with in order to make the comparison as impartial as possible.

In order to compare the techniques applied to the models, we have divided it into two parts. In the first part, we will make a comparison between all the models using the techniques that can be used in each and every one of them, in this way we will be able to make a global discernment of the models and techniques. And the second part is to make a comparison of techniques that can be applied only to one or the other model, in this way we can explain how a model behaves with a technique that cannot be applied to another model due to its characteristics.

5.2 Evaluation of generated models

In order to analyze the models with which we work, we will present it in the following way, we will go model by model and for both datasets.To apply machine learning models, several steps will be followed to the two datasets.

- **Train our model:** When we are in the process of training an artificial intelligence model, what we must do is provide training data that will help the model to learn, these data were obtained in the data processing mentioned in Chapter 3 of this work. For this, the data that we provide to you will be the training data must have an answer that will be the "correct" one or in our case you will have made the "correct" choice.

- **Make predictions about the data delivered to the model:** Using the term "Prediction" is nothing more than referencing the final output of an algorithm once it has been trained on a "historical" data set and then applied to a new set of new data. When forecasting the probability of an obtained result, our algorithm creates values with probability for an unknown variable. Keep in mind that the term "prediction" in certain cases can refer to the fact that the model is going to predict an outcome in the future. But it should be borne in mind that on other occasions the term "prediction" has to do with cases of choice of some event that has already occurred, but what is done is an assumption with more category than what occurred.

- **Performance evaluation of the model used:** When talking about the term "performance" in machine learning we must bear in mind that it is a term that does not have an official definition, and this is because this term does not have limitations such as: The scope of the model to be implemented, the metrics that are intended to be used to be able to make an evaluation at the output of the model and the objective to be achieved when finished using the model. The use of any metric that you want to implement will be subject to the type of variable that you are trying to predict, some of them are: the root mean square error, the recovery, the precision, the F measurement, among others.

Once these previous steps have been explained, now it is time to present one by one the models that have been applied to the datasets, showing their results both in tables and graphs so that you have a more concrete idea of how the behavior of each of the models.

5.2.1 Random Forest Model

Once we have the data ready so that the model can be trained, what we do is instantiate the random forest classifier. To do this, we must take into account several characteristics of the model that we will not delve into, but it is important to mention to have an idea of what is in the classifier. therefore the instances that have been used have been the following:

- bootstrap=True
- class_weight=None
- criterion='gini'
- max_depth=None
- max_features='auto'
- max_leaf_nodes=None
- min_impurity_decrease=0.0
- min_impurity_split=None
- min_samples_leaf=1
- min_samples_split=2
- min_weight_fraction_leaf=0.0
- n_estimators=10
- n_jobs=None
- oob_score=False
- random_state=None
- verbose=0
- warm_start=False

These intakes have been kept the same for the two datasets in such a way that the comparison is as objective as possible and safer to have reliable data, once we have implemented the model, the next step was made predictions for test data.

Following the methodology, what we have to do is evaluate the performance of the model. We will help ourselves with tables. It will be easier to understand the values and

results obtained from each model for each dataset. For this, we design a simple classification evaluative function and in this way, we will obtain Accuracy, Precision / Recall and F1 Score Metrics.

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.99	0.99	0.99	105
Bening	0.98	0.98	0.98	66
Accuracy			0.99	171
Macro avg	0.99	0.99	0.99	171
Weighted avg	0.99	0.99	0.99	171

Table 5.1: Random Forest Performance table for the cancer dataset.

classification Report				
	Precision	Recall	F1-Score	Support
Low Quality	0.70	0.78	0.74	221
Hight Quality	0.79	0.71	0.75	259
Accuracy			0.74	480
Macro avg	0.74	0.75	0.74	480
Weighted avg	0.75	0.74	0.74	480

Table 5.2: Random Forest Performance table for the wine dataset.

As we can see, the model behaves differently for each of the datasets, and at the time of a classification it has values for one data set and other values for the other dataset. We will briefly analyze what we obtained and we will make a quick comparison of the model with the datasets.

- To calculate the precision it is necessary to understand that it is a function that

calculates what proportion of positive identifications was correct, now with this idea we can deduce if our model has an precision of a value "x", it can be assumed that the model was correct "X%" of times to a statement that we say in our cases: malignant tumors, benign tumors, wines of good quality and poor wines.

- To calculate the recall it is necessary to understand that it is a function that calculates that Proportion of real positives was correctly identified, certainly with this we can now say that in our model we have a recovery of an "x" value. In other words, it correctly identifies the "X%" of our wine qualities or our types of cancer.

- It's easy to come up with a new measure of model performance, and it's the F1 Score measure, which is a function of Precision and Recall. and it would be better understood if it is expressed as a formula

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{5.1}$$

We mostly use the F1 score when we are looking for a balance between precision and recall, in the previous items we were able to emphasize that the precision depends a lot on the amount of true negatives, and this in most of the times for us is irrelevant, however F1 score helps us to have an unequal class distribution which means that we have a lot of real negatives.

Once these brief concepts have been understood that have occurred on very complex topics, it is easier for us to say that the model behaves better in a dataset that contains cancer cases than with the dataset that contains wine data and how get to know its quality

Next, we will show the AUC/ROC curve that the model obtains with the two datasets to be able to compare what is shown in that graph.

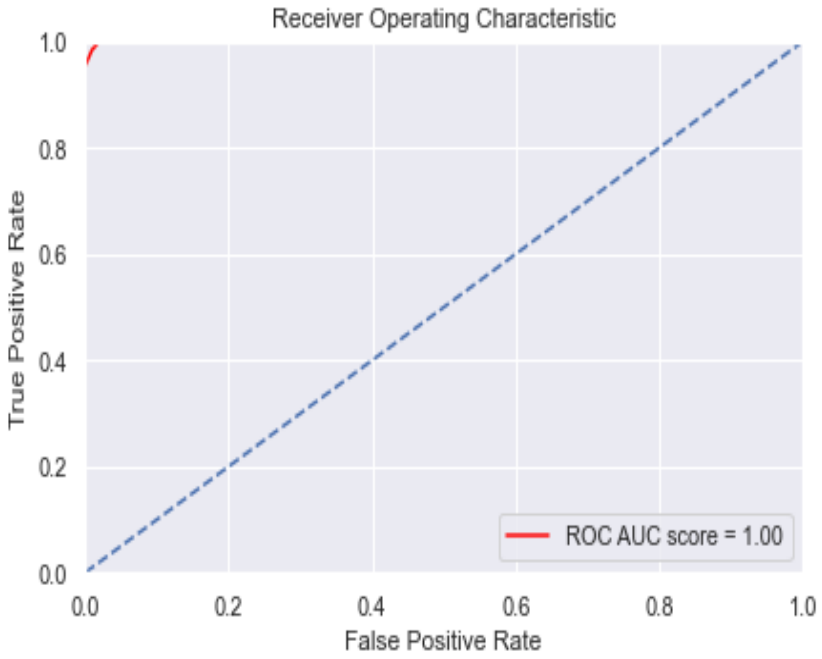

Figure 5.1: Random Forest's ROC/AUC curve in Cancer Dataset

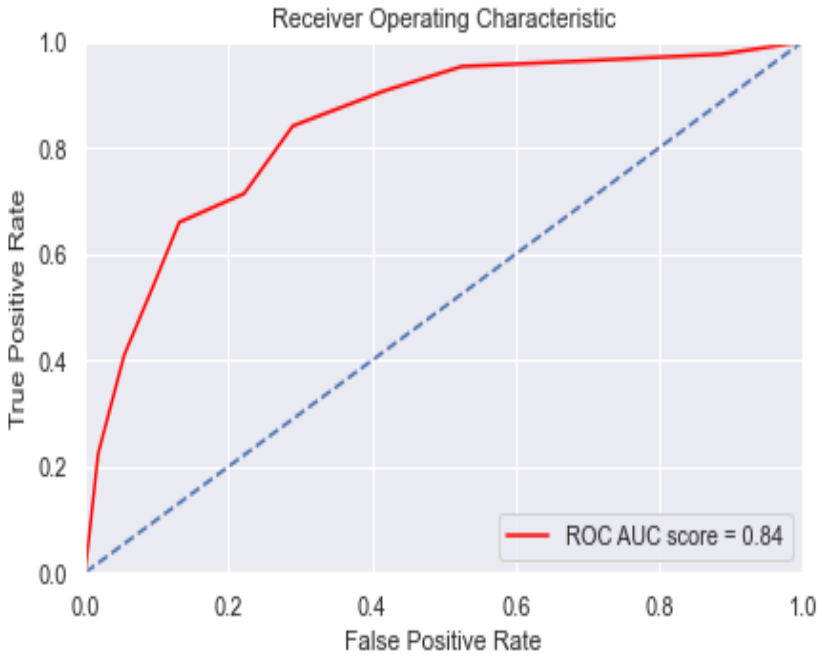

Figure 5.2: Random Forest's ROC/AUC curve in Wine Dataset

The receiver operating characteristics (ROC): curve is defined as the graph that helps us show the performance that our classification model has had, this under all the classification principles, it is necessary to consider that the curve draws two parameters

that are:

- **True Positive Rate (TPR):** In machine learning, the true positive rate, often referring to sensitivity or recall, is generally used to measure the percentage of real positives that are correctly identified, and it is calculated with the following expression.

$$TPR = \frac{TP}{TP + FN} \tag{5.2}$$

- **False Positive Rate (FPR):** is the number of negative cases wrongly identified as positive cases within the data, that is, the probability that there are missing alarms, and it can be expressed by the formula:

$$FPR = \frac{FP}{FP + TN} \tag{5.3}$$

The ROC curve plots the relationship between TPR and FPR under different in different classification measures. Lowering these thresholds can cause more elements to be classified as positive, which results in an increase in false positives and true positives.

Area Under the ROC Curve (AUC): it helps us give an added measure of performance across all possible rating thresholds. Interpreting AUC can have several approaches, but one of them is, as a model, it has a probability of classifying an element as positive rather than a negative element.

This means that the ROC curves indicate the sensitivity of a related test, which produces continuous results, taking into account the data known as false positives at different cut points.

With this information we can say that the model behaves in a better way with the dataset that contains cancer data, being able to make a better discernment between whether it is a type of benign or malignant cancer, and on the other hand the model has 84% probability of distinguishing between whether a wine is of good or bad quality.

5.2.2 k-closest Neighbors

Next, we will analyze the results obtained with the k-closest Neighbors model, after having done the data treatment and having adjusted and trained the model. in both datasets both to make the discernment between malignant cancer and benign cancer, as well as to be able to classify good and bad wines.

As before, this model must also be instantiated so that the classification characteristics are as close as possible to each other. therefore the instances that have been used have been the following:

- algorithm='auto'
- leaf_size=30
- metric='minkowski'
- metric_params=None
- n_jobs=None
- n_neighbors=9
- p=2
- weights='uniform'

These intakes have been kept the same for the two datasets in such a way that the comparison is as objective as possible and safer to have reliable data, once we have implemented the model, the next step was made predictions for test data. methodology, what we have to do is evaluate the performance of the model. We will help ourselves with tables. It will be easier to understand the values and results obtained from each model for each dataset. For this, we design a simple classification evaluation function and in this way, we will obtain Accuracy, Precision / Recall and F1 Score Metrics. As we can consider, with this model

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.68	0.98	0.80	105
Bening	0.89	0.26	0.40	66
Accuracy			0.70	171
Macro avg	0.79	0.62	0.60	171
Weighted avg	0.76	0.70	0.65	171

Table 5.3: KNN Performance table for the cancer dataset.

we have a performance of lower quality than the previous one in all aspects, however, in this work, as it is not oriented to obtain malignant results with the data but rather a saber because a model shows the results it shows is of It is very interesting for us to know how it differs from the other models once the comparison is made, which we will see later

classification Report				
	Precision	Recall	F1-Score	Support
Low Quality	0.62	0.62	0.62	221
Hight Quality	0.68	0.67	0.67	259
Accuracy			0.65	480
Macro avg	0.65	0.65	0.65	480
Weighted avg	0.65	0.65	0.65	480

Table 5.4: KNN Performance table for the wine dataset.

Now we will show the respective ROC/AUC curves for each model and thus we can give another opinion regarding the behavior of the model.

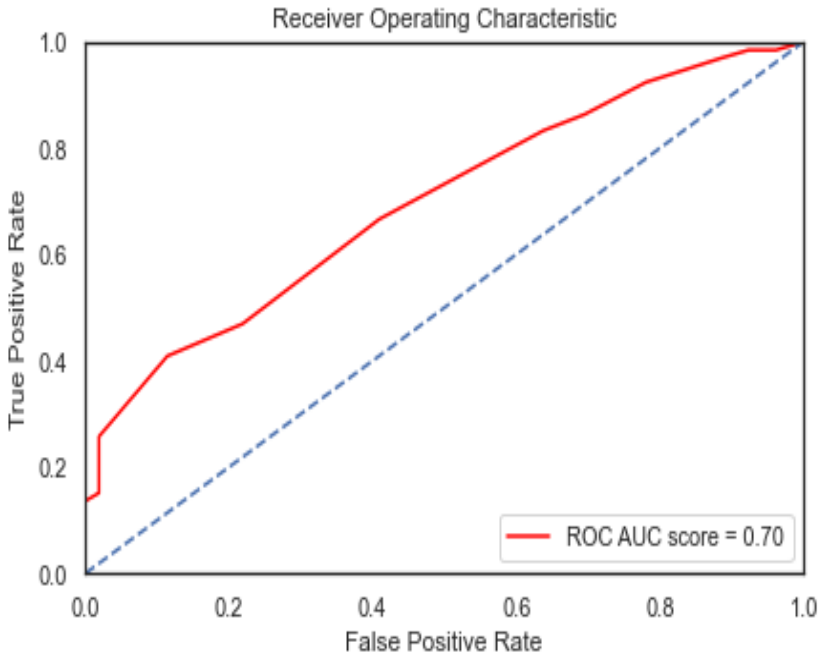

Figure 5.3: KNN's ROC/AUC curve in Cancer Dataset

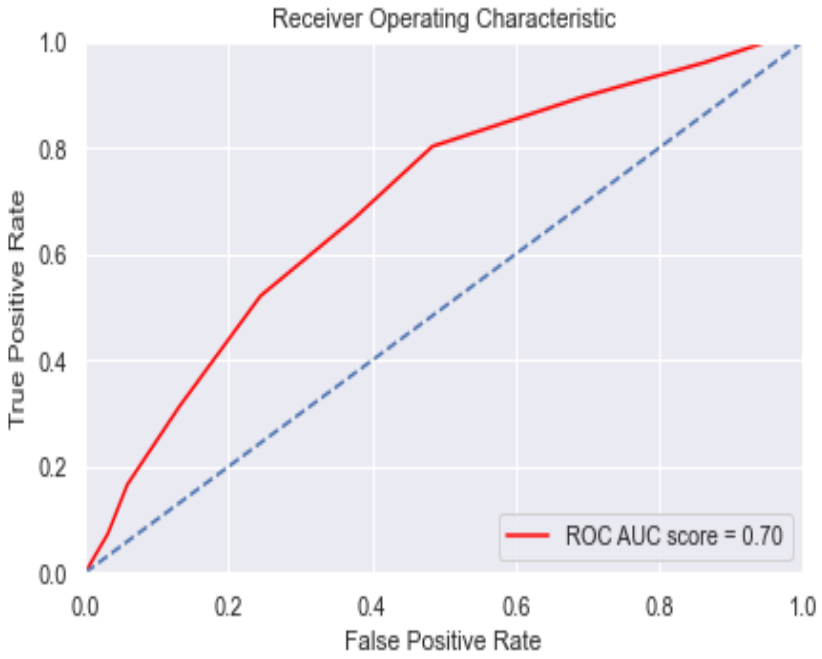

Figure 5.4: KNN's ROC/AUC curve in Wine Dataset

As we can see in the graph of and taking into account the brief explanation that it is an ROC curve and the AUC we can notice that the data from the dataset that contains information on the types of cancer in the previous model had a very good curve that could practically distinguish between which types of tumors were benign and which others were malignant, now is not the case the graph shows us that there is an overlap of identity, that is to say that the model could not so easily distinguish between one or the other this time, something similar but to a lesser extent it happens with the data set of the wines, the model with the features that it has acquires problems of distinction to know which wine is good and which wine is bad, However the value of AUC for both graphs is 0.70, that is to say that the model has a 70% probability of distinguishing between a benign and malignant tumor and whether a wine is of good quality or of poor quality.

Next we will see something a little more interesting since we will use a model called Naive Bayes, as previously explained it is a probability model and this is precisely what makes it interesting since we can approach this model from a perspective called Naive Bayes with a Bernoulli approach and Naive Bayes with a Gauss approach, we will show the results below.

5.2.3 Naive Bayes Bernoulli Model

Naive Bayes with Bernoulli distribution, the Bernoulli distribution is a concept that is applied to discrete random variables that can have only two possible events "success" and "not success". or as in our case for a dataset "malignant tumor" or "benign tumor" and for another dataset "wine of good quality" and "wine of bad quality".

Therefore, to be able to use a bernoulli distribution in Naive Bayes it is necessary that the samples are represented with binary values, if the model is fed with other data types, it is not a problem either since the model has an instance that can binarize the data, of course this depends on the binarization parameter that is used. Bernoulli can make a decision on a Naevi Bayes model based on the rule:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{5.4}$$

Once this model was implemented in the two datasets, we could obtain the following results: For this model with this Bernoulli characteristic and also treating the datasets in the same

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	1.00	1.00	1.00	105
Bening	1.00	1.00	1.00	66
Accuracy			1.00	171
Macro avg	1.00	1.00	1.00	171
Weighted avg	1.00	1.00	1.00	171

Table 5.5: Naive Bayes Bernoulli Performance table for the cancer dataset.

way to avoid that there are better or worse results, and thus a deeper discernment can be made on how the models behave, we can clearly see how with some data It behaves better than with others such is the case of the information that is in the dataset of the types of cancer when doing the classification does it almost perfectly not for the dataset of the classification of the wines. with the graphics that will be shown later we will explain something more about it

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.50	0.09	0.12	221
Bening	0.54	0.93	0.68	259
Accuracy			0.54	221
Macro avg	0.52	0.51	0.42	480
Weighted avg	0.52	0.54	0.44	480

Table 5.6: Naive Bayes Bernoulli Performance table for the wines dataset.

As we have seen in the model performance tables, for one dataset it has behaved marvelously, while for the other dataset it has not, then we show the ROC / AUC curves that have been obtained for this model, and thus give a clearer idea of what we will see

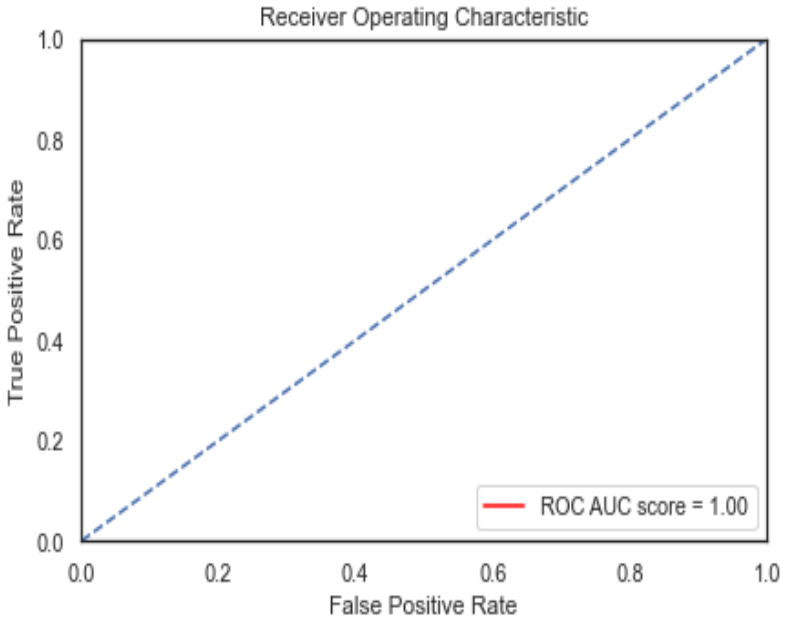

Figure 5.5: Naive Bayes Bernoulli's ROC/AUC curve in Cancer Dataset

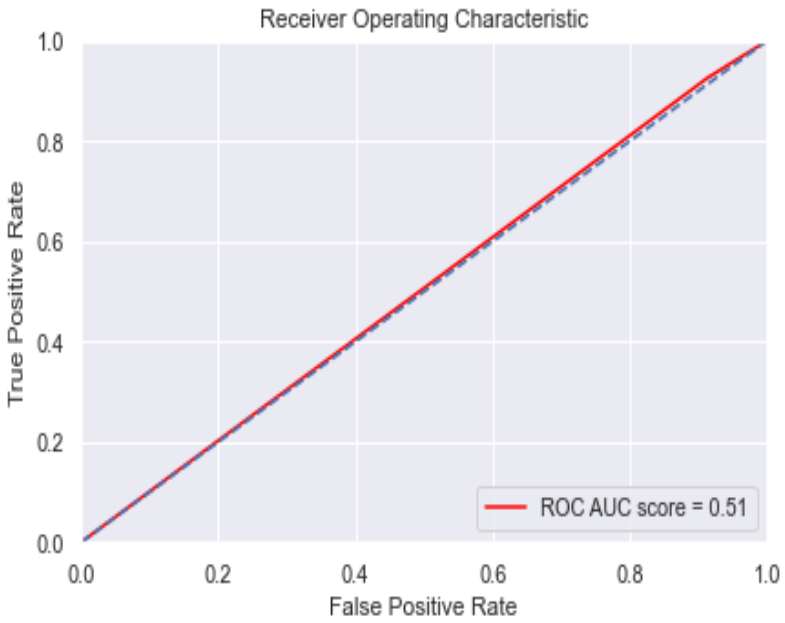

Figure 5.6: Naive Bayes Bernoulli's ROC/AUC curve in Wine Dataset

Now as we can see in the figures for the dataset that contains the information of the types of tumors, the model is doing a great job since it can easily and accurately distinguish one type of tumor or another type of tumor. This means that malignant tumors do not exceed the marked threshold and benign tumors do not exceed the marked threshold, which implies that we have neither false positives nor false negatives.

and therefore the AUC value for this model is 1, which means that it has a 100 % probability of discerning whether a tumor is malignant or benign. On the other hand we have the wine quality dataset, and here on the contrary we say that this model for this data set is useless since it has many values in the range of false positives and many values in the range of false negatives, it is say the classification charts are crossed with each other and exceed the threshold for both good quality and poor quality wines. and the AUC value gives us 0.51 and this means that the model does not have the ability to discriminate between two classes, it cannot know what type of wine is good and what type of wine is bad, possibly this dataset should be treated otherwise or simply do not apply this model.

5.2.4 Naive Bayes Gaussian Model

Naive Bayes with Gauss distribution, Now we know that there is another distribution to itself and it is the Gaussian distribution, this type of distribution is a continuous function that closely approximates the binomial distribution of events. This means that the Gaus-

sian Naive Bayes classifier adapts better to continuous data, and thus for the better its performance assumes that those at the input come with a normal distribution.

In this way and to carry out the classification. The probability of the characteristics is assumed to be Gaussian:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{5.5}$$

Below we show the results that have been obtained with this model for both datasets, and with this we can make a distinction of how it has behaved for both a set of data and the other.

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.62	0.98	0.76	105
Bening	0.50	0.03	0.06	66
Accuracy			0.61	171
Macro avg	0.56	0.51	0.41	171
Weighted avg	0.57	0.61	0.49	171

Table 5.7: Naive Bayes Gaussian's Performance table for the cancer dataset.

As we can see in this model we no longer have perfection, but it behaves in a way that is neither very good nor very bad, and this is due to the way in which the data is taken from input to the model, and since they are not necessarily binary data or have been binarized, because the model acts in a way for this type of data, and gives us a different result than what could have been assumed with the other model.

classification Report				
	Precision	Recall	F1-Score	Support
Low Quality	0.70	0.68	0.69	221
Hight Quality	0.73	0.75	0.74	259
Accuracy			0.72	480
Macro avg	0.71	0.71	0.71	480
Weighted avg	0.72	0.72	0.72	480

Table 5.8: Naive Bayes Gaussian's Performance table for the wine dataset.

As with the other dataset, this model behaves in a not very optimal way but neither does it do so as with the previous model in which it was not able to distinguish from a good wine or a bad wine. This assumes that this model is a little more versatile than that of Bernoulli, but we will see it when we enter the explanatory part of the models. has had the model

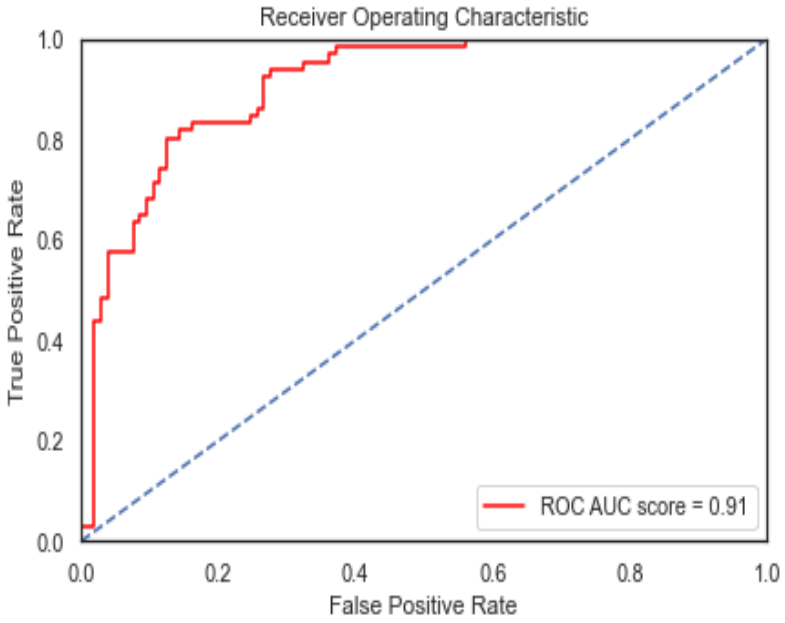

Figure 5.7: Naive Bayes Gaussian's ROC/AUC curve in Cancer Dataset

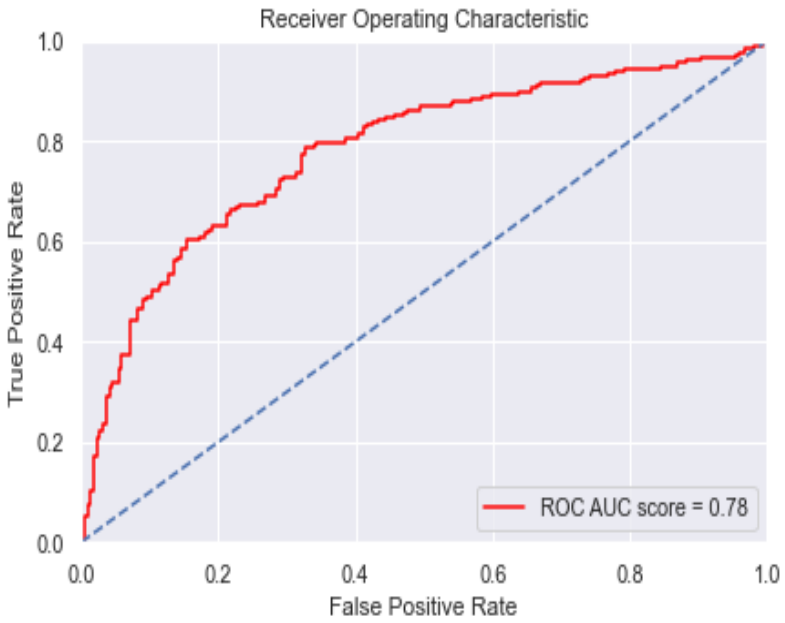

Figure 5.8: Naive Bayes Gaussian's ROC/AUC curve in Wine Dataset

On the other hand we have the data set of the wines and now we can see that the performance of the model is much better than with the previous model, since although it is true it still has a high value of false positives and false negatives, now the model it is able to identify if a wine is of good or bad quality and that at least 70 % of the time.

5.2.5 Logistics Regression

When we talk about logistic regression, we have to go back to the beginning of the 20th century, since it is at that time that logistic regression was used for the first time and used in the biology sciences, then it came to be used in many branches of the social sciences, and then already for most of the problems they have in everyday life. This model is used when the dependent or objective variable has binary values or in other words it is categorical, as in our cases there is a malignant tumor or poor quality wine, since we assign a value of 0 and if the tumor is benign or the wine is of good quality because we assign the value of 1]

As before, this model must also be instantiated so that the classification characteristics are as close to each other as possible. therefore, the instances that have been used are the following:

- C=1.0
- class_weight=None
- dual=False
- fit_intercept=True
- intercept_scaling=1
- l1_ratio=None
- max_iter=100
- multi_class='warn'
- n_jobs=None
- penalty='l2'
- random_state=None
- solver='warn'
- tol=0.0001
- verbose=0
- warm_start=False

These intakes have been kept the same for the two datasets in such a way that the comparison is as objective as possible and safer to have reliable data, once we have implemented the model, the next step was made predictions for test data. methodology, what we have to do is evaluated the performance of the model. We will help ourselves with tables. It will be easier to understand the values and results obtained from each model for each dataset. For this, we design a simple classification evaluation function and in this way, we will obtain accuracy, Precision / Recall and F1 Score Metrics. As we can consider, with this model

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.60	0.03	0.05	105
Bening	0.39	0.97	0.55	66
Accuracy			0.39	171
Macro avg	0.49	0.50	0.30	171
Weighted avg	0.52	0.39	0.25	171

Table 5.9: Logistics Regression's Performance table for the cancer dataset.

classification Report				
	Precision	Recall	F1-Score	Support
Malignant	0.71	0.70	0.70	221
Bening	0.75	0.76	0.75	259
Accuracy			0.73	480
Macro avg	0.73	0.73	0.73	480
Weighted avg	0.73	0.73	0.73	480

Table 5.10: Logistics Regression's Performance table for the wine dataset.

As we can see in the two tables, the model is shared differently for each dataset. The performance that is obtained when working with the dataset that has the data with respect to the classification of malignant cancer and benign cancer is less than the performance it has had. with the dataset of the qualities of the wine. This, for example, can be seen that the precision with malignant tumors is higher than with benign tumors, and with the recall it is the opposite, it is better with benign tumors than with malignant tumors, and the F-score scores thus demonstrate this.

With that information also and now showing the ROC graphs, more accurate comments can be obtained, and a better appreciation of how the model has worked with the data can be given.

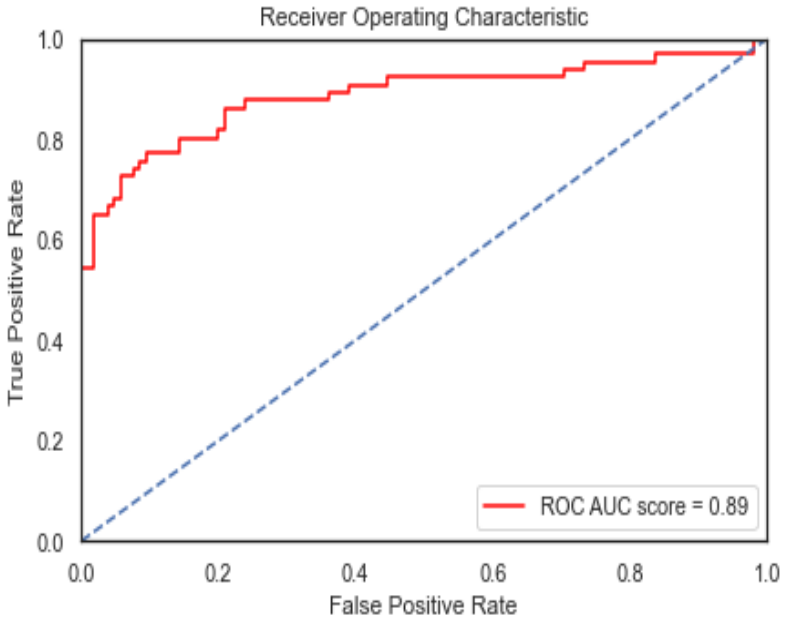

Figure 5.9: Logistics Regression's ROC/AUC curve in Cancer Dataset

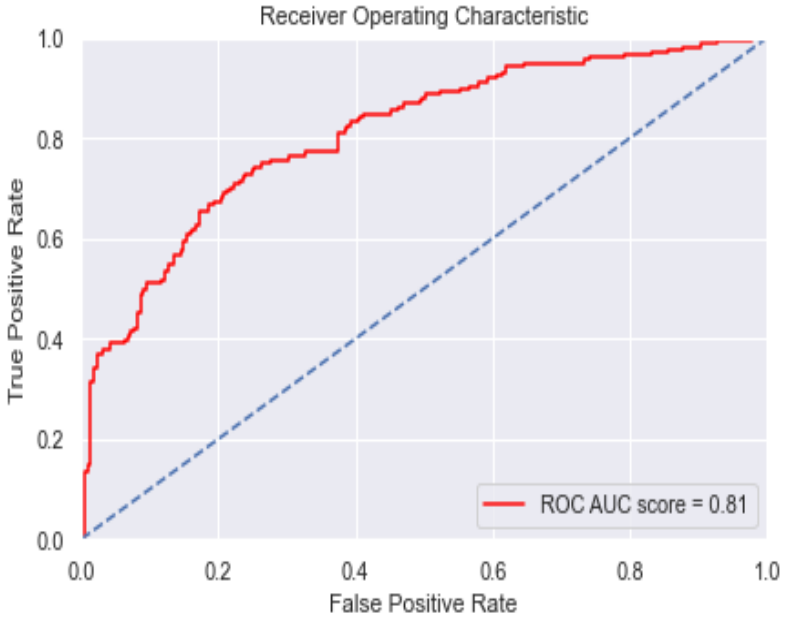

Figure 5.10: Logistics Regression's ROC/AUC curve in Wine Dataset

In the graphs where we can see the ROC curves, we can see how we have the decrease of the false-negative and false green values for the two models, in the model that analyzes the data on the quality of the wine it is evident that the graph is a little better than for data that has cancer information. Now as for the AUC values we can see that for the data corresponding to the classification of cancer types we have 89 % that the model can do the classification in the correct way and for the model of the classification of The probability that the model can make good judgment is 81 %. It is good to be able to do an analysis of explanatory tools in the future.

5.3 Comparative Analysis of XAI approaches: Global Analysis

Now we have reached the fundamental part of our work and that is to explain how it is that each technique applied to each model, tells us how it is that each algorithm is capable of making a decision or in our case it will make a classification, for this once again We will work to analyze the techniques that can be applied to all the models, taking as reference a particular feature that for our cases in all our datasets and applying the first model analyzed Random Forest, we obtained a graph of "Features Importance" with which we could Observe which feature had more relevance in the model and from that compare these features with the other models and the other explanatory techniques so that everything has a consistency.

The following table shows in the numerical form how each feature had to do so that the model can make the classification decision. On the other hand, a figure is shown that contains the chart of Importance of characteristics for the wine data set. We can see how it is that for this data set the model tells us that what was important to him was the characteristic of alcohol, but how can this characteristic influence a decision?

Later we will see how alcohol can influence a model to decide whether it is a good wine or a bad wine.

Feature Importance	
Feature	Importance
alcohol	0.211460
volatile acidity	0.111827
sulphates	0.106692
total sulfur dioxide	0.097851
density	0.092291
pH	0.068644
chlorides	0.068026
residual sugar	0.062705
fixed acidity	0.061695
free sulfur dioxide	0.061401
citric acid	0.057407

Table 5.11: Feature Importance's table for the wine dataset.

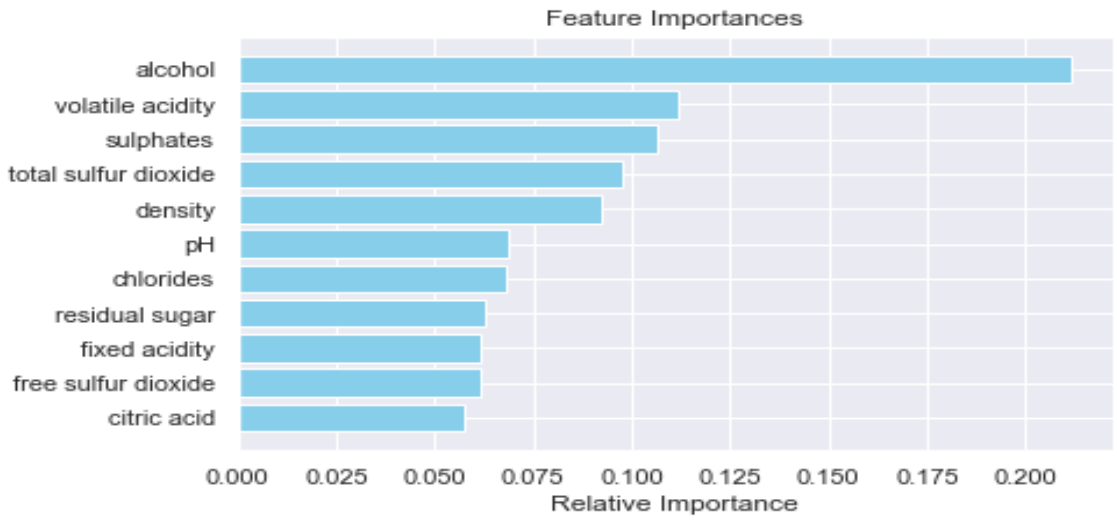

Figure 5.11: Feature importance graph in the wine dataset.

The following table shows in numerical form how each characteristic had to be done in order for the model to make the classification decision. On the other hand, a figure is shown that contains the Importance of Characteristics graph for the cancer dataset. We can see how the model tells us for this data set that what was important to it was the characteristic of the perimeter mean, but how can this characteristic influence a decision?

Later we will see how the perimeter mean can influence a model to decide if a tumor is malignant or benign.

Feature Importance				
Feature	**Importance**		**Feature**	**Importance**
perimeter_mean	0.242062		texture_worst	0.006500
area_worst	0.156980		radius_mean	0.005974
concave points_mean	0.153956		symmetry_mean	0.005055
concavity_worst	0.059181		id	0.004896
perimeter_worst	0.049458		smoothness_se	0.003196
perimeter_se	0.036465		compactness_worst	0.002495
radius_worst	0.030609		compactness_se	0.001706
concave points_worst	0.014953		fractal_dimension_se	0.001661
rea_mean	0.008800		symmetry_se	0.001011
concavity_mean	0.008761		fractal_dimension_mean	0.000512
smoothness_mean	0.008128		concave points_se	0.000506
texture_mean	0.008084		texture_se	0.000130
fractal_dimension_worst	0.008053		concavity_se	0.000000
smoothness_worst	0.008031		radius_se	0.000000
area_se	0.007259		compactness_mean	0.000000

Table 5.12: Feature Importance's table for the cancer dataset.

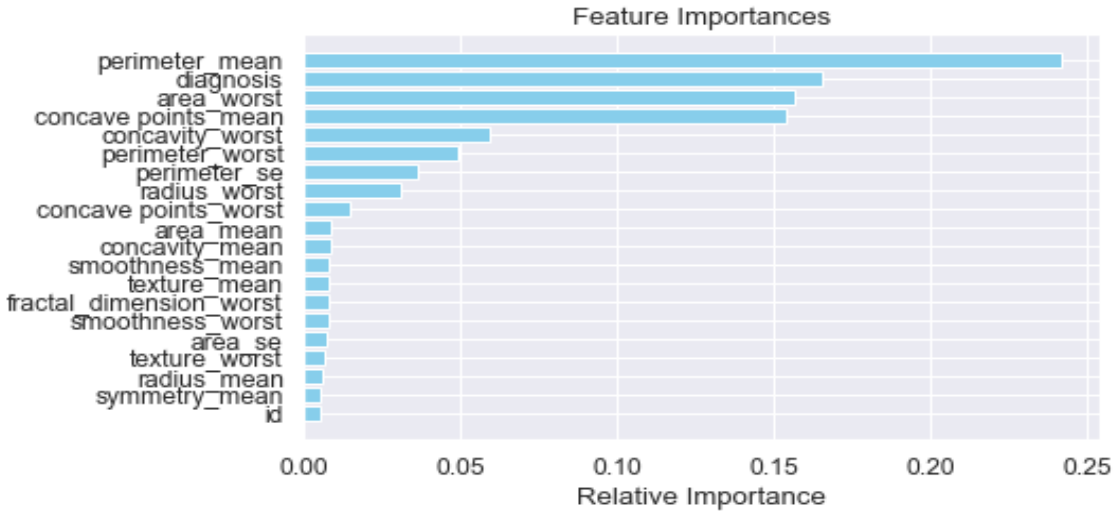

Figure 5.12: Feature importance graph in the cancer dataset.

Now that we have obtained the fundamental characteristics to be able to implement the explanation techniques, it will be easier for us to implement the tools in this way when it is done by characteristics that is the most general at the time of making an explanation, we will do it with a single fundamental characteristic that will give us a clearer idea of how a model behaves.

after this brief introduction of how it has been taken into consideration what characteristic. We will start with the techniques that have been used to explain each model in each work dataset. So we will start with:

5.3.1 Partial Dependence Plots (PDP), ICE Plots, and Bivariate PD Plots for all Models :

Now we will make an analysis of the functioning of the models with respect to a variable as we had already said in the previous section to make a better analysis of the behavior and not to delay the explanation much, many of the people who see these graphs, will not know what This is, but it is for this reason that what is happening will be explained and analyzed model by model, before starting just as a way to refresh knowledge, we will say that. Partial dependency charts have characteristics that show how each variable or predictor affects the model's prediction.

Below we will show the result of our work, then we will explain them. for practical purposes and to avoid confusion in this section, they are separated as follows: first the PDP charts obtained with all the models in the wine dataset are shown and explained, and then the PDP charts will be shown and explained that were obtained from the cancer dataset.

As already explained, the PDPs are graphs that will show us the average of a selected characteristic in our case, we already have it very clear what they will be (alcohol and perimeter _mean, respectively), the Individual Conditional Expectation (ICE) graphs are a Great tool to be able to make a disaggregation of the average obtained, these graphs will give us a vision of the functional relationship between what is predicted and the function, all this separately, and in each instance. they are a method of disaggregating these averages. The ICE graphs visualize the functional relationship between the predicted response and the function separately for each instance. In simpler terms, a PDP shows the average of the individual lines that are in the ICE charts

As we will see in the graphs below, each line represents an instance and gives us a vision of what happens when the characteristic of a specific instance varies, which is obtained by making the prediction in the model, taking into account that the other characteristics remain constants ICE diagrams highlight variations in fitted values over the range of a feature. In such a way that we could know where and how much the heterogeneity is.

PDP charts are a great way to try to explain something that a black box model does to a user who has no prior AI/ML knowledge, or to an audience that has no technical or engineering background. However, it is a somewhat limited visualization technique, to make it a little more agile you also have the option of explaining visually and with two characteristics at the same time this is done by combining one characteristic with another, in theory, it should be the combination of one numerical characteristic with a categorical characteristic, but if you only have numerical characteristics it works in the same way, although there is a degradation of the ability to interpret, this is done with color intensity

scale analysis as will be seen in the following graphs

5.3.2 Partial Dependence Plots (PDP),ICE Plots and Bivariate PD Plots for wine dataset:

With the data that we have from the dataset, it can be said that the quality of the wines differs greatly depending on the characteristics of the data set, initially it seems somewhat complicated to unravel the effects of these characteristics, so we will base ourselves only on the characteristic of alcohol that had more importance in the Random Forest model.

For this, the already adjusted model was used in which it made the classification of a good quality and poor quality wine, what is done is to repeatedly modify the value of a variable (alcohol), in order to make a series of classifications. the wine could be classified if the alcohol was only present in a percentage, which varies from 0% to X% of presence in the wine. let's see.

It is necessary to highlight some elements of the graph, for a better understanding.

The vertical axis is interpreted as a change in the classification of what would be classified in the baseline or in the leftmost value. On the other hand, the blue shaded area tells us the level of confidence that one has.

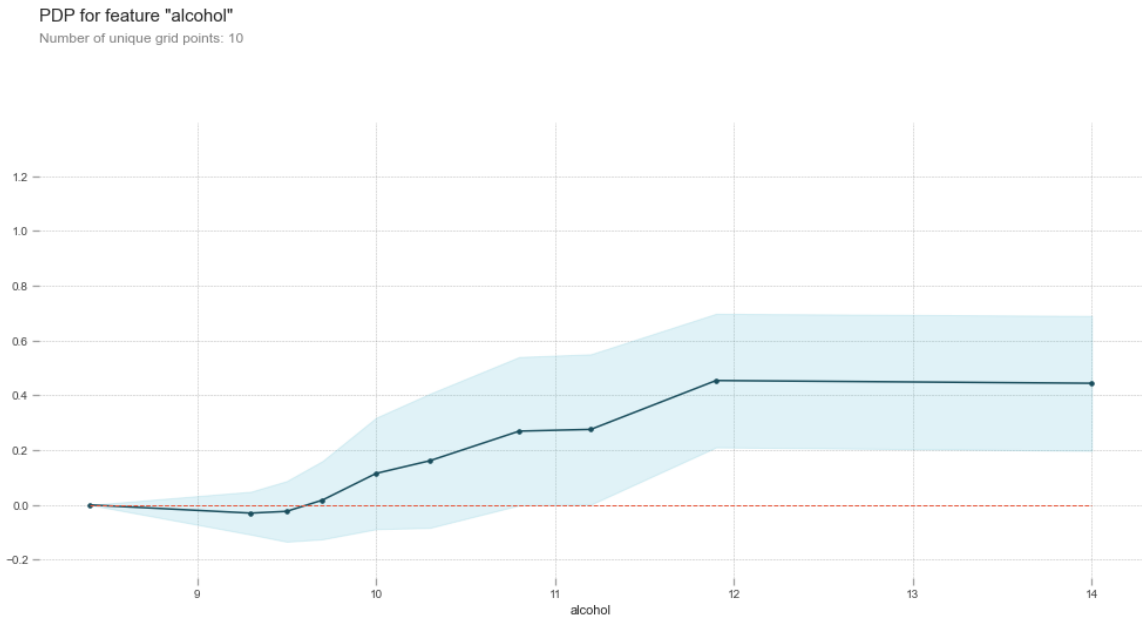

Figure 5.13: PDP's Random Forest Model for Wine dataset.

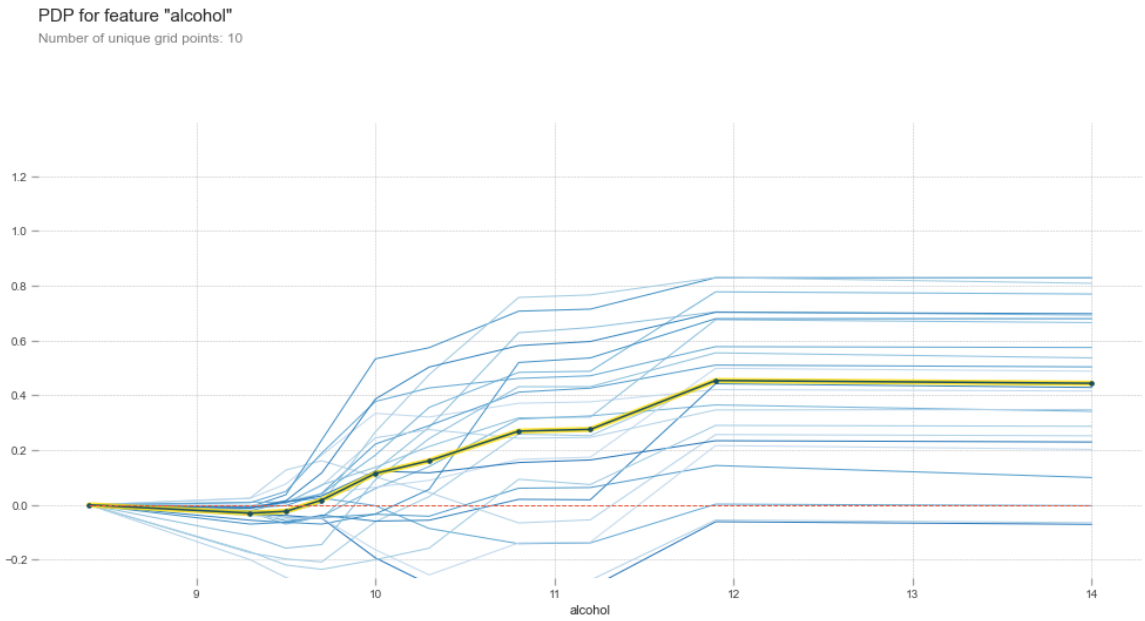

Figure 5.14: ICE Plot Random Forest Model for Wine dataset.

In this graph, we can see that the alcohol level has an influence that increases the prediction towards High-quality wines for values between 9 and 12, after that range, the importance continues to be positive but gradually decreases, before that the influence was neutral and there was a time when there was a slight negative influence

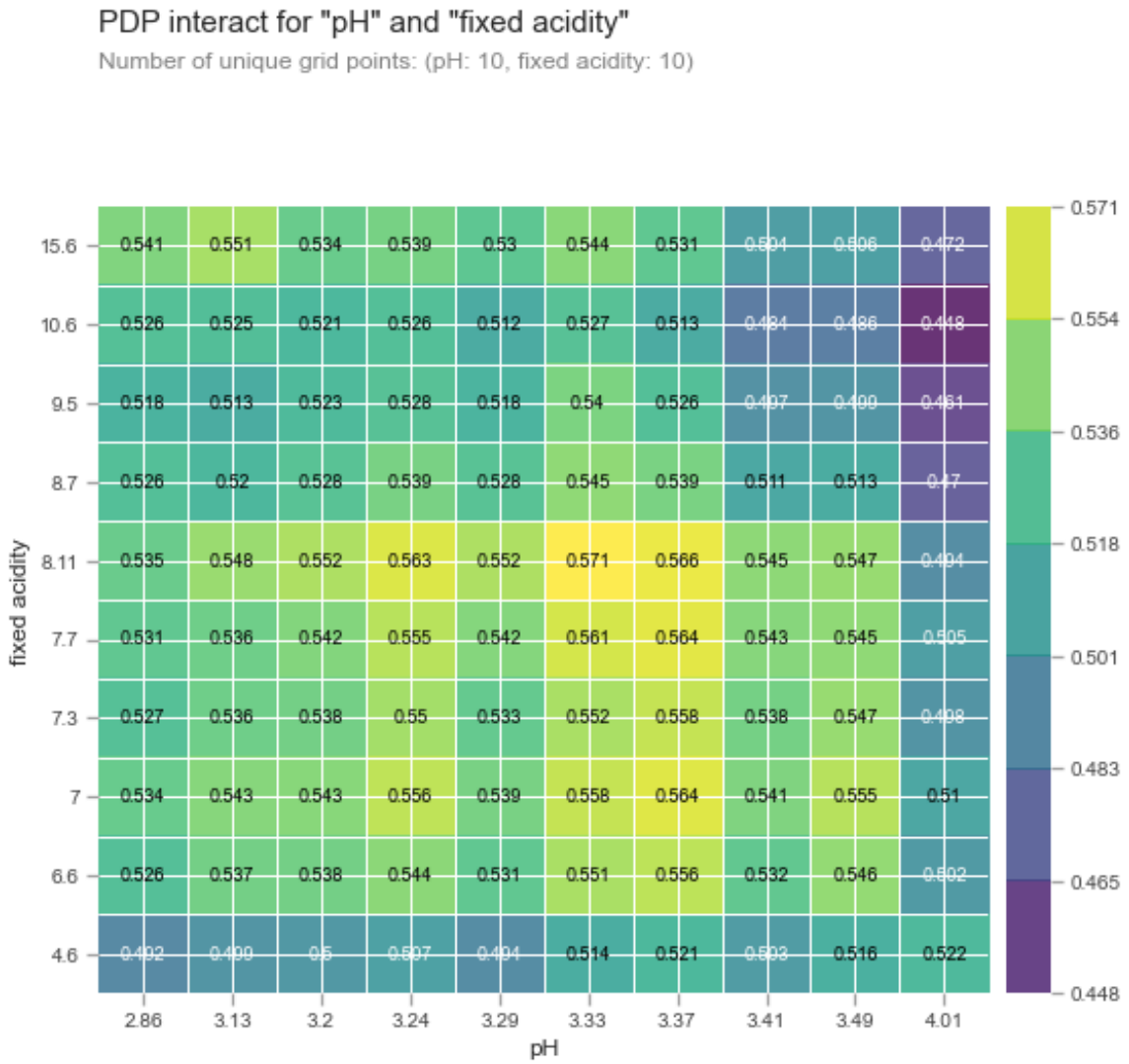

Figure 5.15: Bivariate PD Plot Random Forest Model for Wine dataset.

As we can see to the right of the graph we have a bar that indicates the probability of having a good quality wine, the best probability is found in dark violet color and the best probability of having a good quality wine is in a yellow color Now, when we have a graph showing the combination of characteristics (on the horizontal axis we have the pH and on the vertical axis we have fixed acidity), it is important to make an analysis of what is happening, so let's start with the worst case, when the pH has a value of 4.01 and the fixed acidity has a value of 10.6 the probability of having good wine is 44%, contrasting with the best case we would tend that if the pH has a value of 3.33 and the fixed acidity has a value of 8.11 so we would have the probability that the wine is good of 57.1%. As we can observe and realize when using two variables that, by the way, are variables that are

in the mean of importance and that do not differ much between them, the probability of having a good wine remains almost 50/50

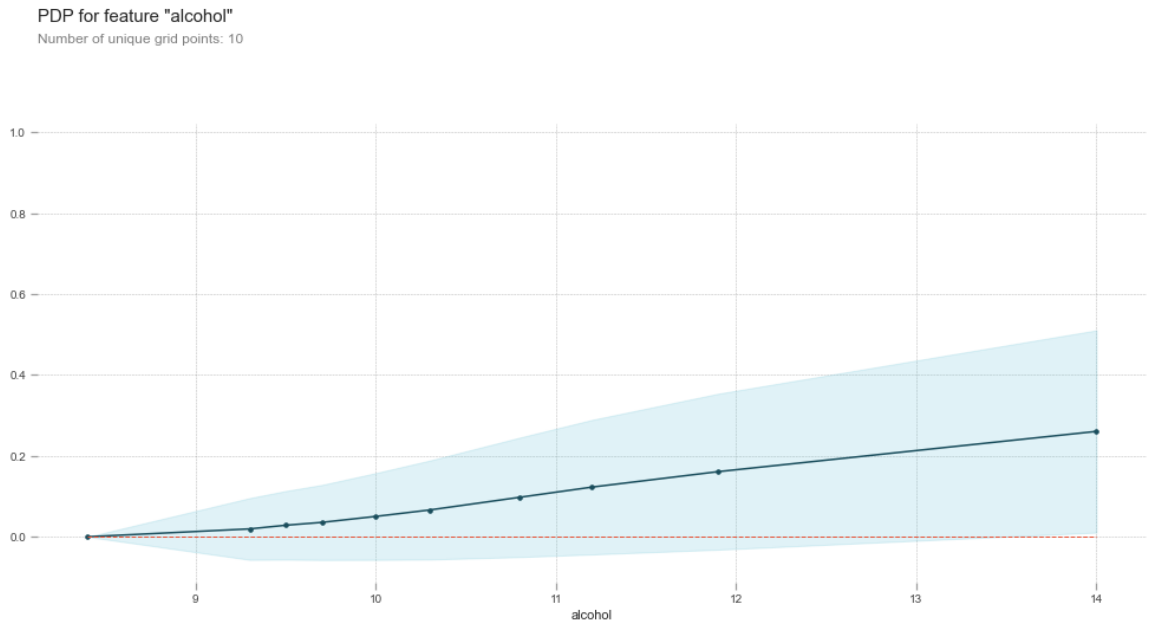

Figure 5.16: PDP's KNN Model for Wine dataset.

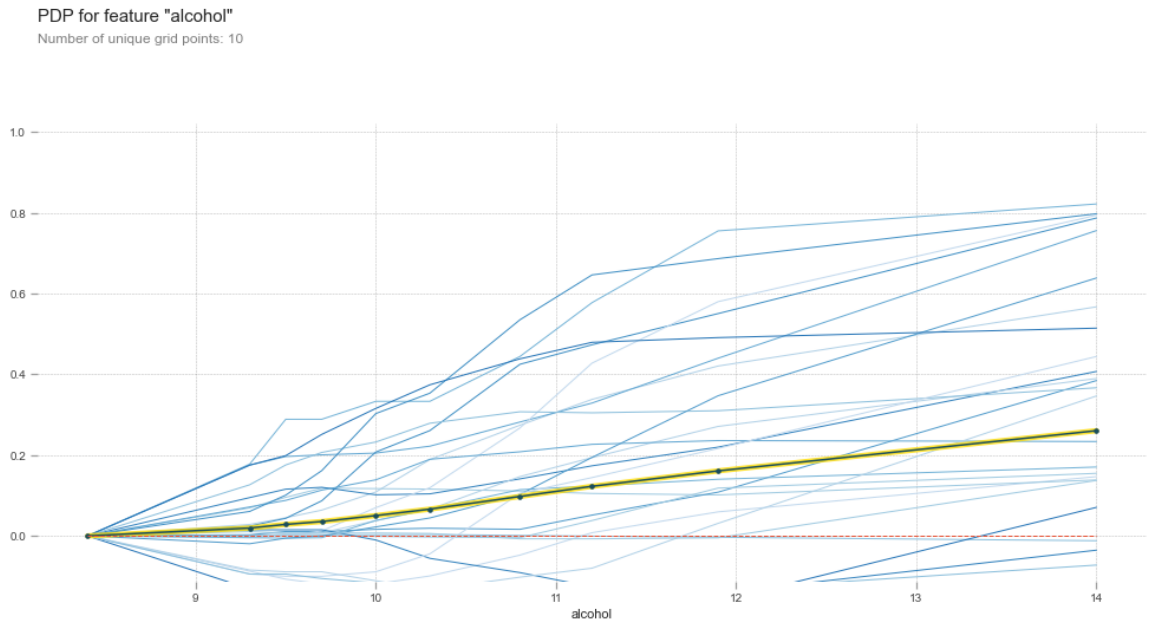

Figure 5.17: ICE Plot KNN Model for Wine dataset.

In this graph, we can see that the alcohol level has an influence that is only positive

the more the alcohol level increases, the wine will be of better quality for the model. With

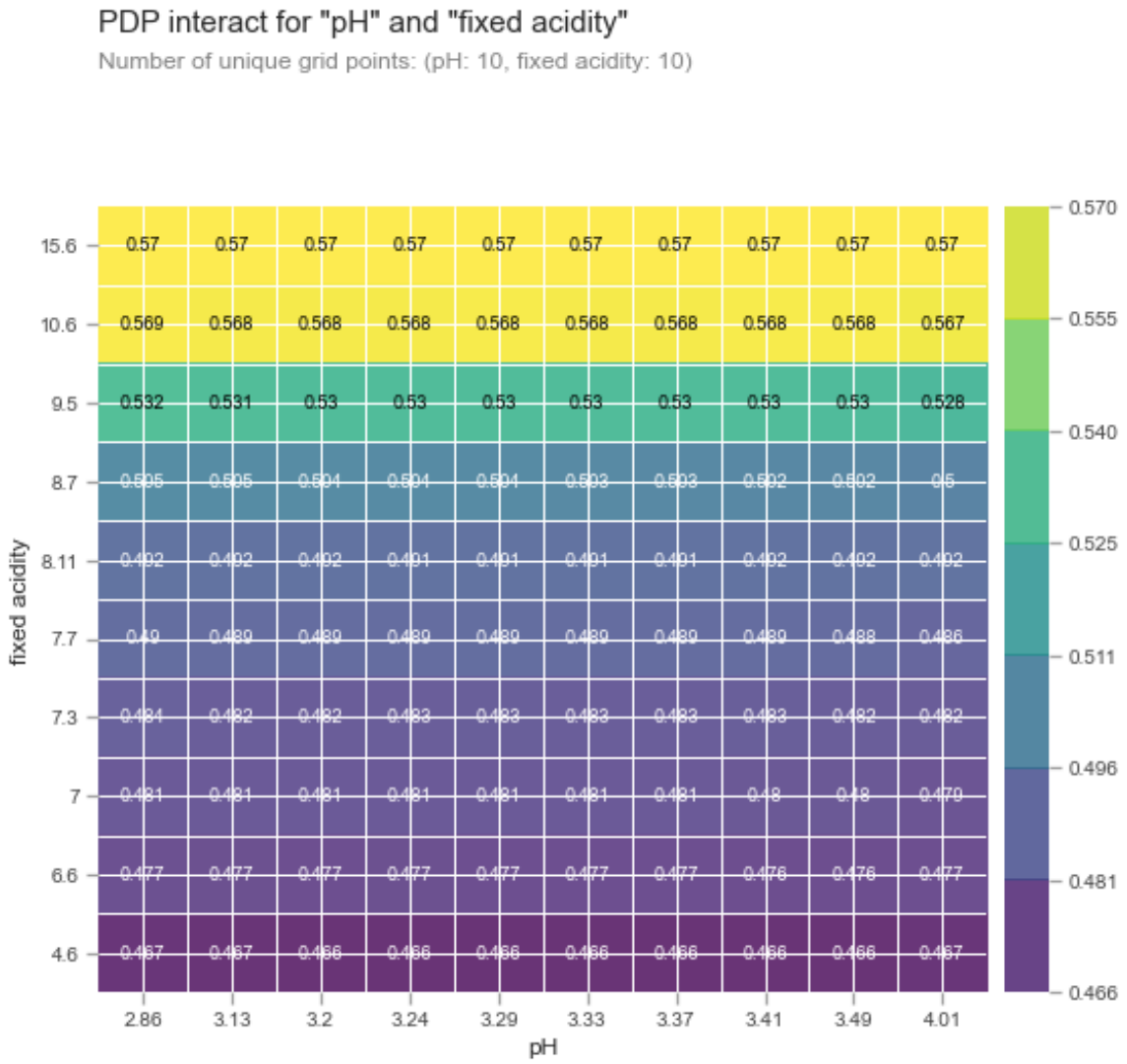

Figure 5.18: Bivariate PD Plot KNN Model for Wine dataset.

the KNN model it is different as we can see, they are very well determined when these two characteristics together can be a good wine and when they produce a bad wine, we can also choose that this model is well parceled, so we will analyze when it has less probability of being a good wine with a value of 46.66 % probability of being a good wine when the pH has a value of 3.2 and the value of fixed acidity is 4.6. On the other hand, the best probability of having a good wine is 57 % having pH values equal to 2.88 and fixed acidity equal to 15.6, in the graph we can see that it does not matter how high or low the pH value is. really important for the model is the value of the fixed acidity the higher its value the

better the probability of having a good wine. This model is interesting to analyze since

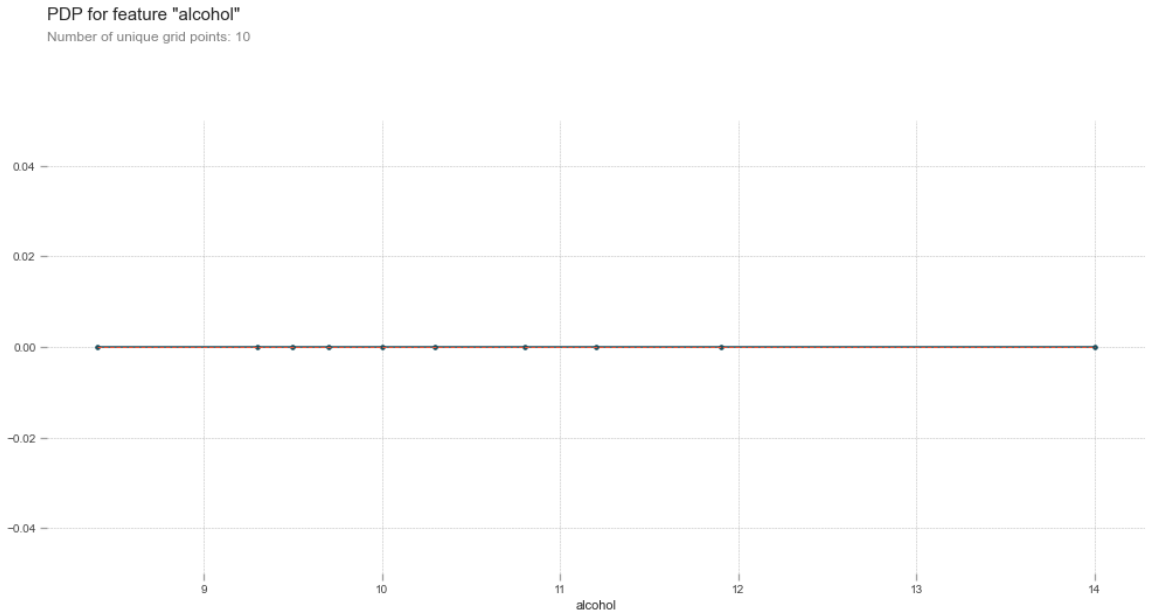

Figure 5.19: PDP's Naevi Bayes Bernoulli Model for Wine dataset.

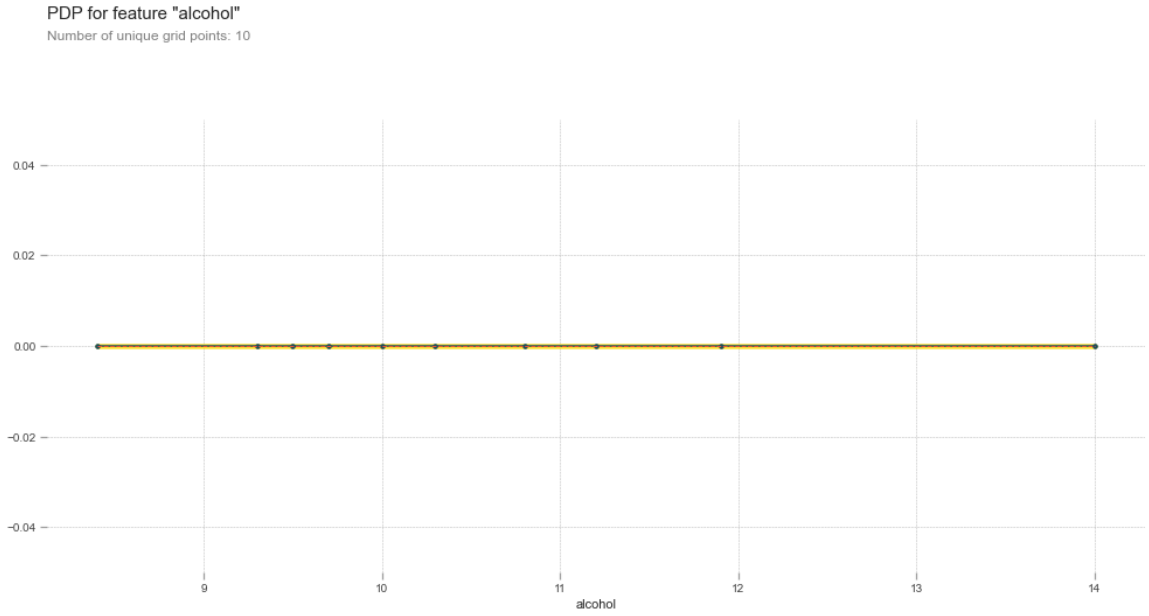

Figure 5.20: ICE Plot Naevi Bayes Bernoulli Model for Wine dataset.

the characteristic of alcohol has no significance for it, this may be because this model does not make a distinction between variant values, but rather does so in a "binary" (1 and 0)

way. way for the model the characteristic is trivial.

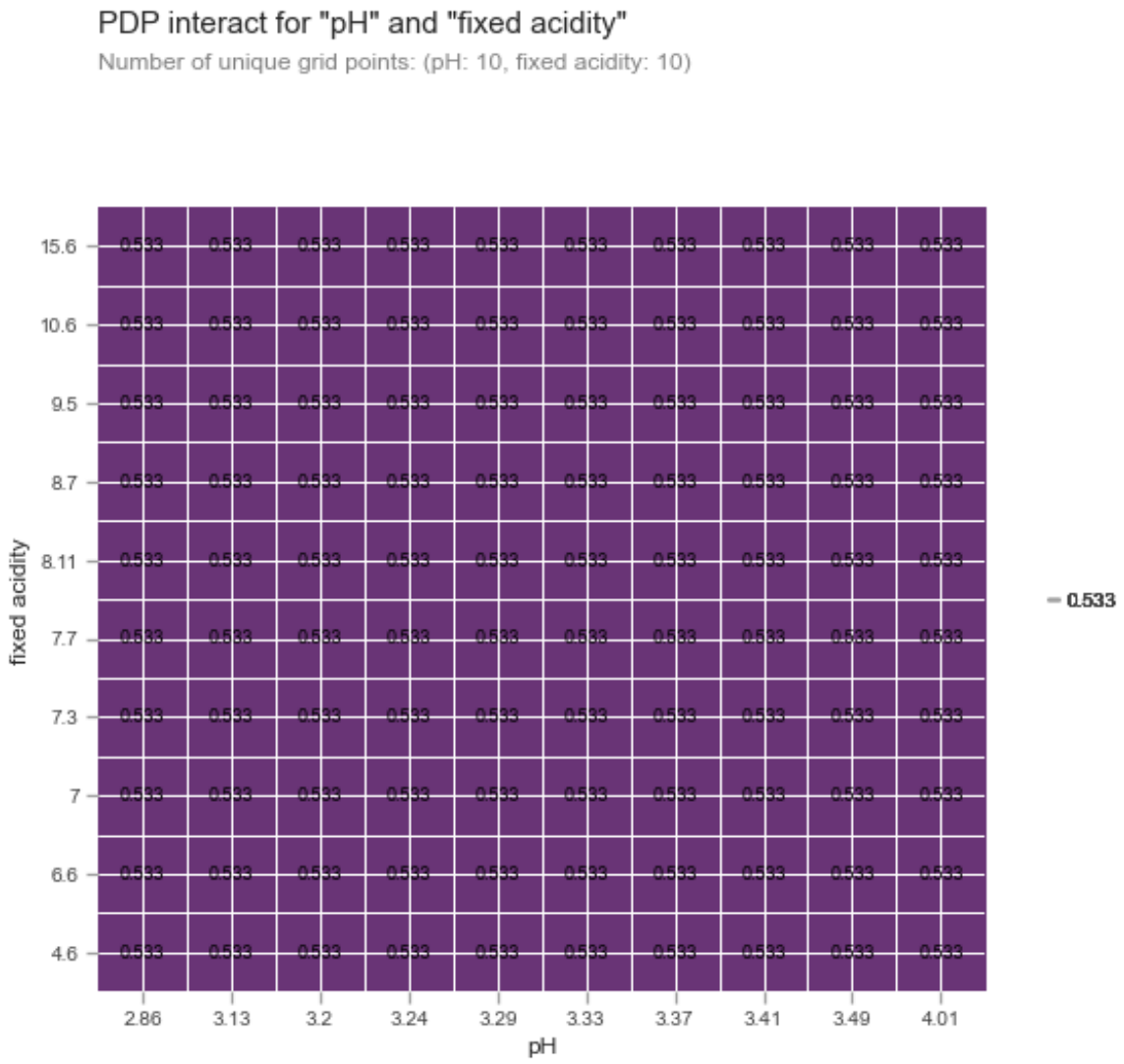

Figure 5.21: Bivariatee PD Plot Naevi Bayes Bernoulli Model for Wine dataset.

This model would give an erroneous impression before eyes that are not trained to understand what is happening, for the model the pH and the fixed acidity are inconsequential, that is to say, that it does not matter when one or the other varies the percentage of having a good wine quality will always be 53.33 %

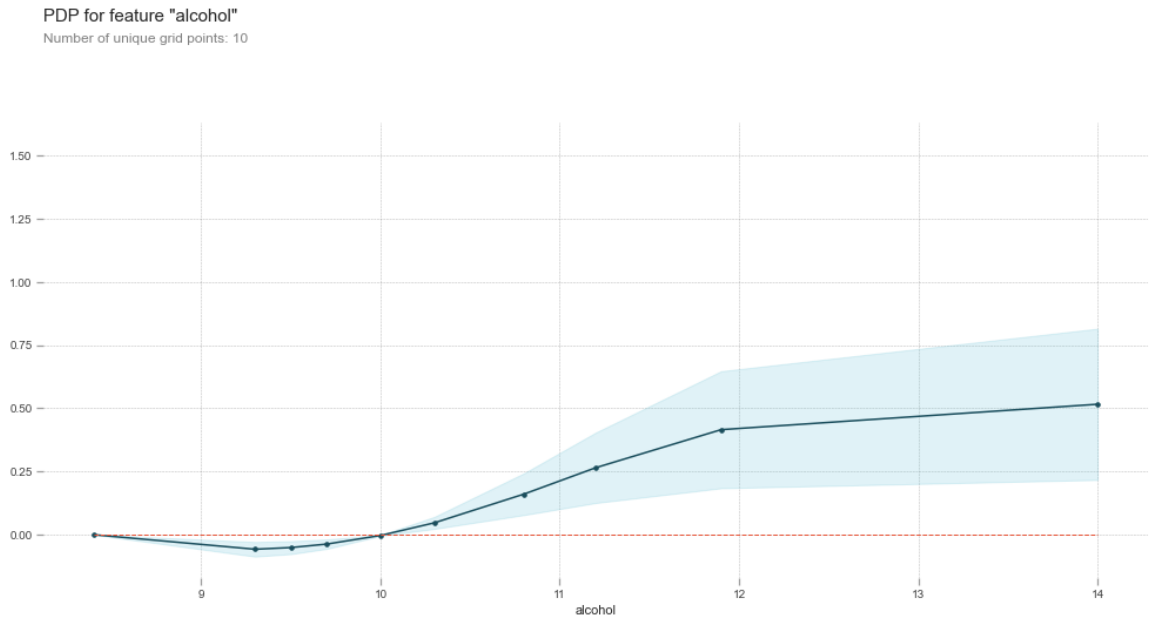

Figure 5.22: PDP's Naevi Bayes Gaussian Model for Wine dataset.

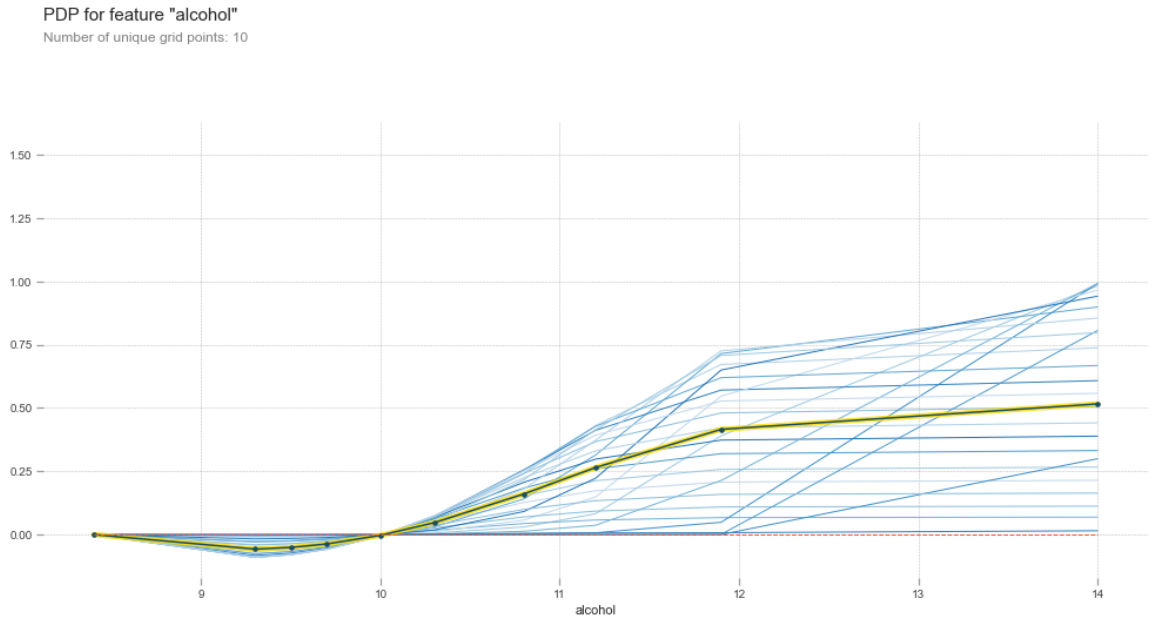

Figure 5.23: ICE Plot Naevi Bayes Gaussian Model for Wine dataset.

In this graph, again we can see how it is that the level of alcohol has an influence that increases the prediction towards high quality wines, the difference is that here up to the value of 10 alcohol was a characteristic that negatively influenced the model from from that value the characteristic begins to take positive values and to influence positively in a

gradual way up to 11.8 more or less and from there it remains constant in growth.

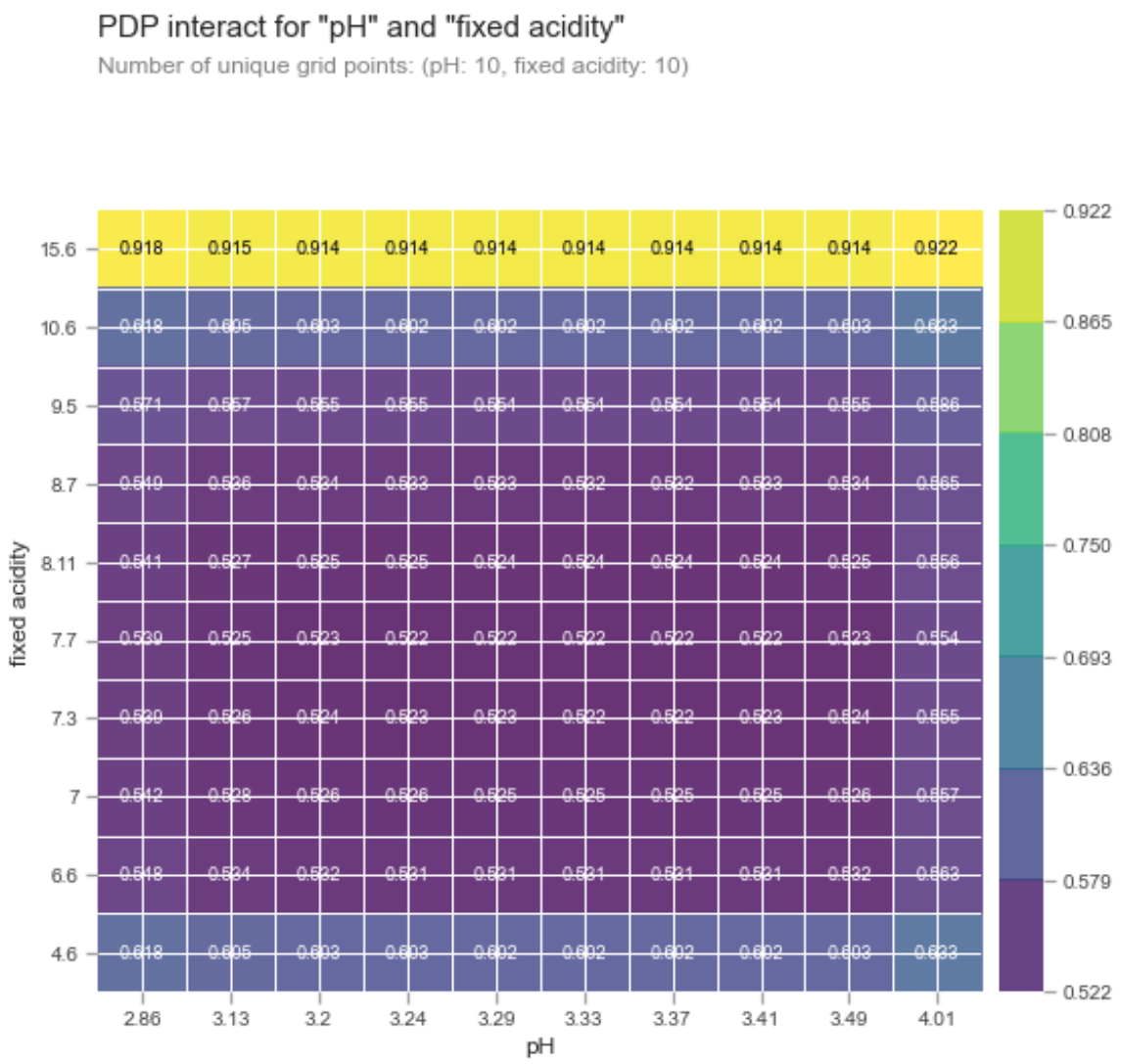

Figure 5.24: Bivariate PD Plot Naevi Bayes Gaussian Model for Wine dataset.

When we see the bivariate graph of the Naevi Bayes Gaussian model we can see something peculiar, the worst probabilities are concentrated in the center of the table, that is, if the pH s is between 3.13 and 3.49 and the value of fixed acidity is between 6.6 and 8.7 we will have the worst probability of having good wine is 52.5 %, but when the fixed acidity level reaches its maximum value, which is 15.6, the best probability of having good wine is when the pH starts at 2.86 with a probability of 91.8 % and when the pH reaches its maximum value which is 4.01, in that case, the probability of having a good quality wine is 92.22 % in the rest of the threshold, the probability remains at 91.4 %

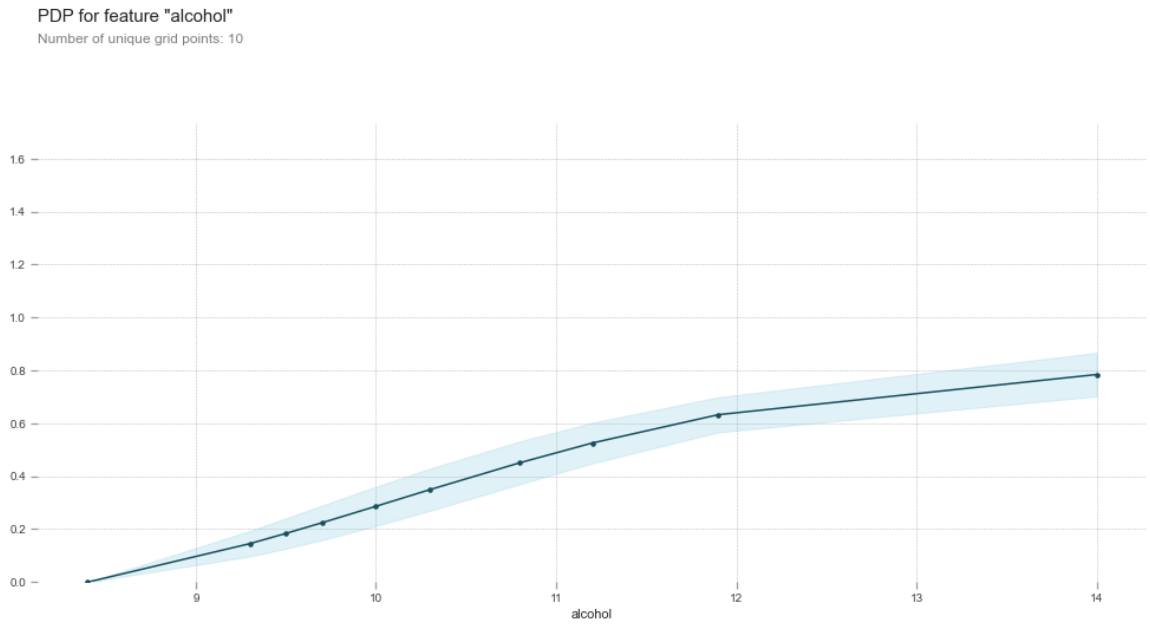

Figure 5.25: PDP's Logistics Regression Model for Wine dataset.

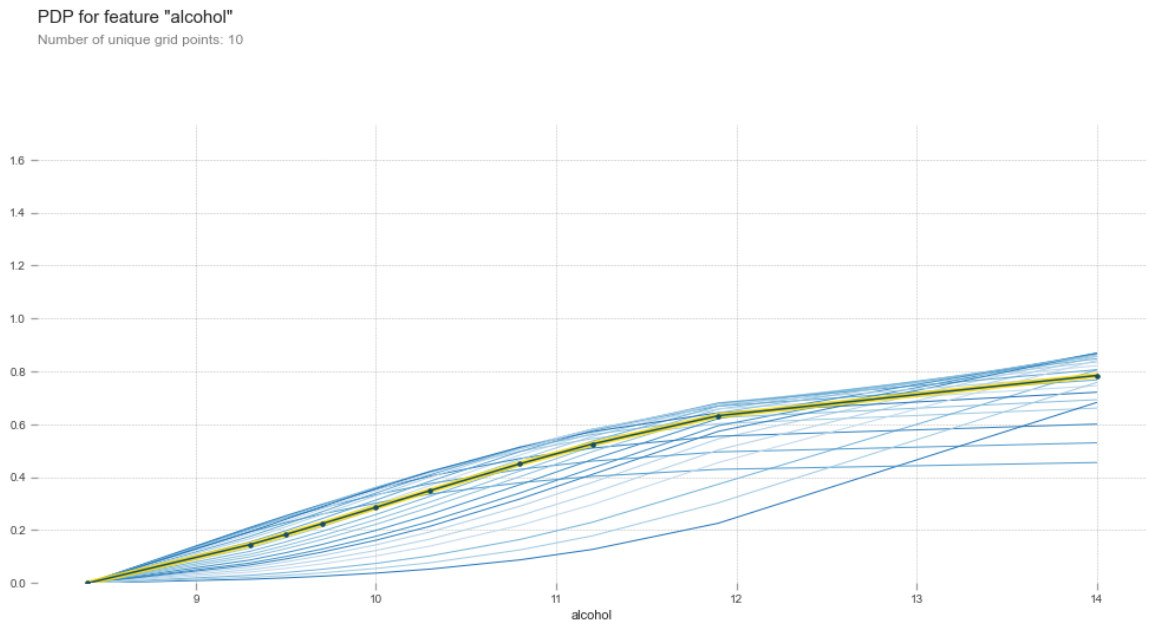

Figure 5.26: ICE Plot Logistics Regression Model for Wine dataset.

As we can see in the logistic regression graph, the growth of importance increases as the amount of alcohol in the wine. If you have the concept of AI / ML for linear or logistic regression models, you can interpret the partial dependency graphs to be similar to the coefficients in these models.

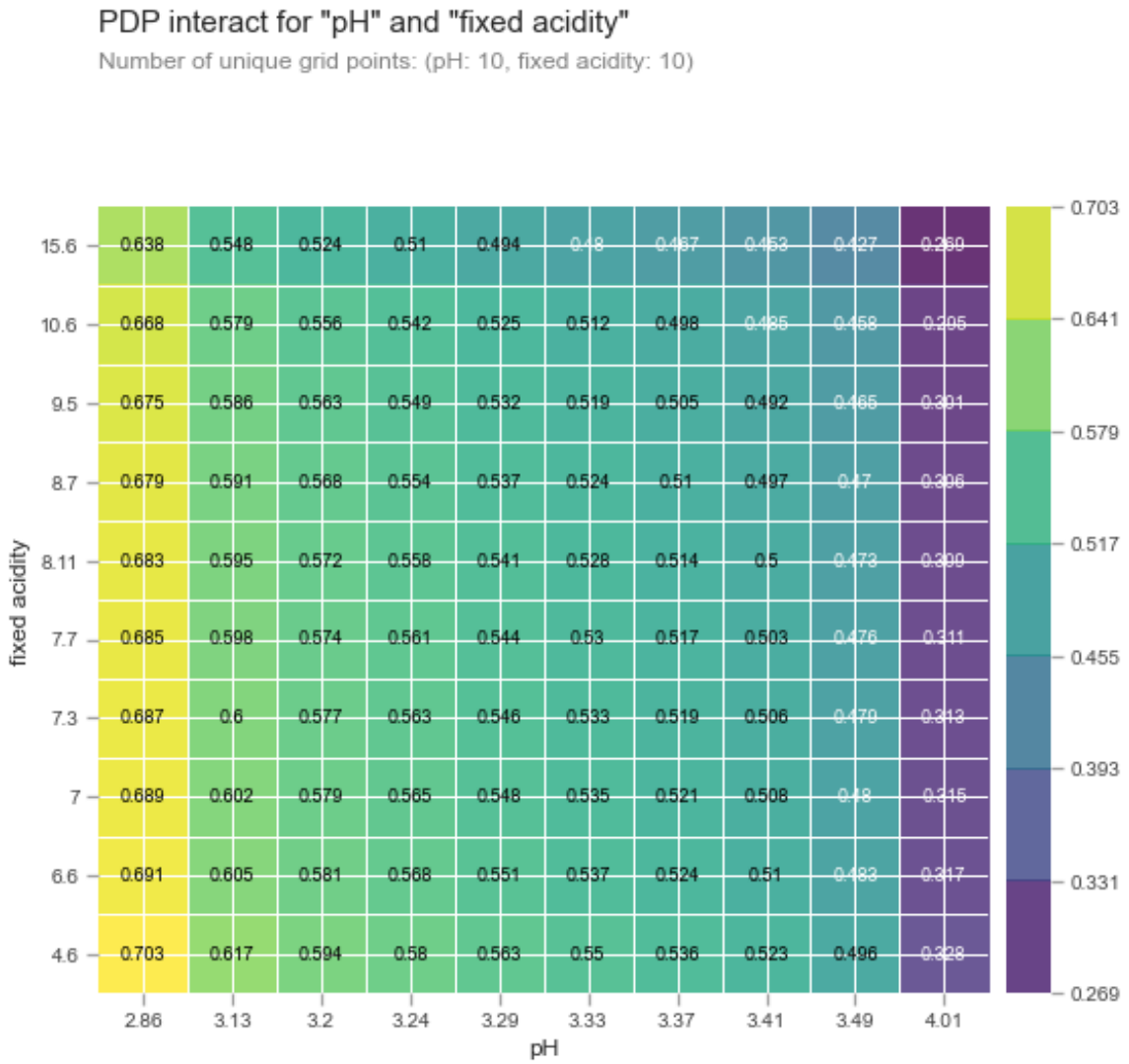

Figure 5.27: Bivariate PD Plot Logistics Regression Model for Wine dataset.

This bivariate graph also has its particularity, as we can see here the best probabilities do not depend on the fixed acidity values but rather on the pH as we see when the pH is more valuable no matter the fixed acidity the probability of having a good wine is 21 % on the contrary, if we have an initial pH value of 2.86 and an initial fixed acidity value, we have a probability of 70.3 % of having a good wine

5.3.3 Partial dependency plots (PDP) and ICE Plots for cancer data set:

With the data we have from the data set, it can be said that predicting whether a tumor will be malignant or benign differs greatly according to the characteristics of the data set, just as The previous data set to arrive at this idea is complex to do and even more so when more characteristics are involved, so we will only base ourselves on the characteristic of the perimeter _mean that had more importance in the Random Forest model. For this, the already adjusted model was used in which the classification of the types of tumors was made saying whether a tomor was benign or malignant as before, what is done is to repeatedly modify the value of the variable (perimeter _mean) to perform a series of classifications. Tumors could be classified if the perimeter only had a percentage, which varies from 0% to X% of presence in the person. let's see. It is necessary to highlight some elements of the graph, for a better understanding. The vertical axis is interpreted as a change in the ranking from what would rank at the baseline or leftmost value. On the other hand, the blue shaded area tells us the level of confidence that one has

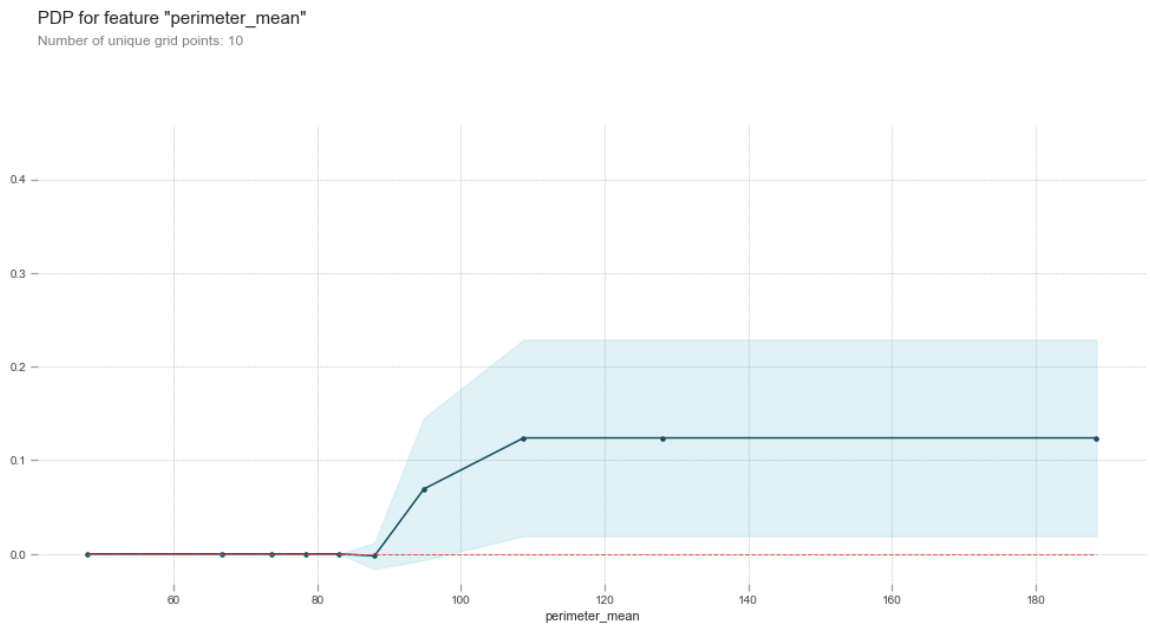

Figure 5.28: PDP's Random Forest Model for cancer Dataset.

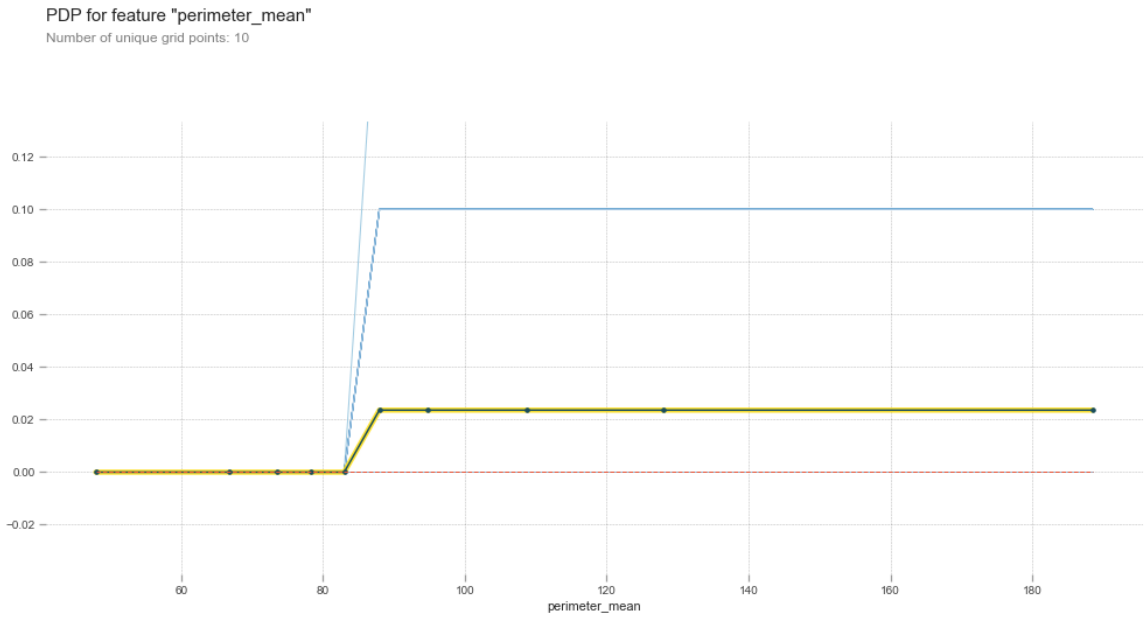

Figure 5.29: ICE Plot Random Forest Model for cancer Dataset.

It is not surprising that the random forest model obtains this type of graph, as we can see the characteristic remains neutral until values between 90 once that threshold is reached, the feature begins to have certain importance until the values of approximately 95 up to 110 there is a growth of importance a greater and already from that value onwards the characteristic remains constant

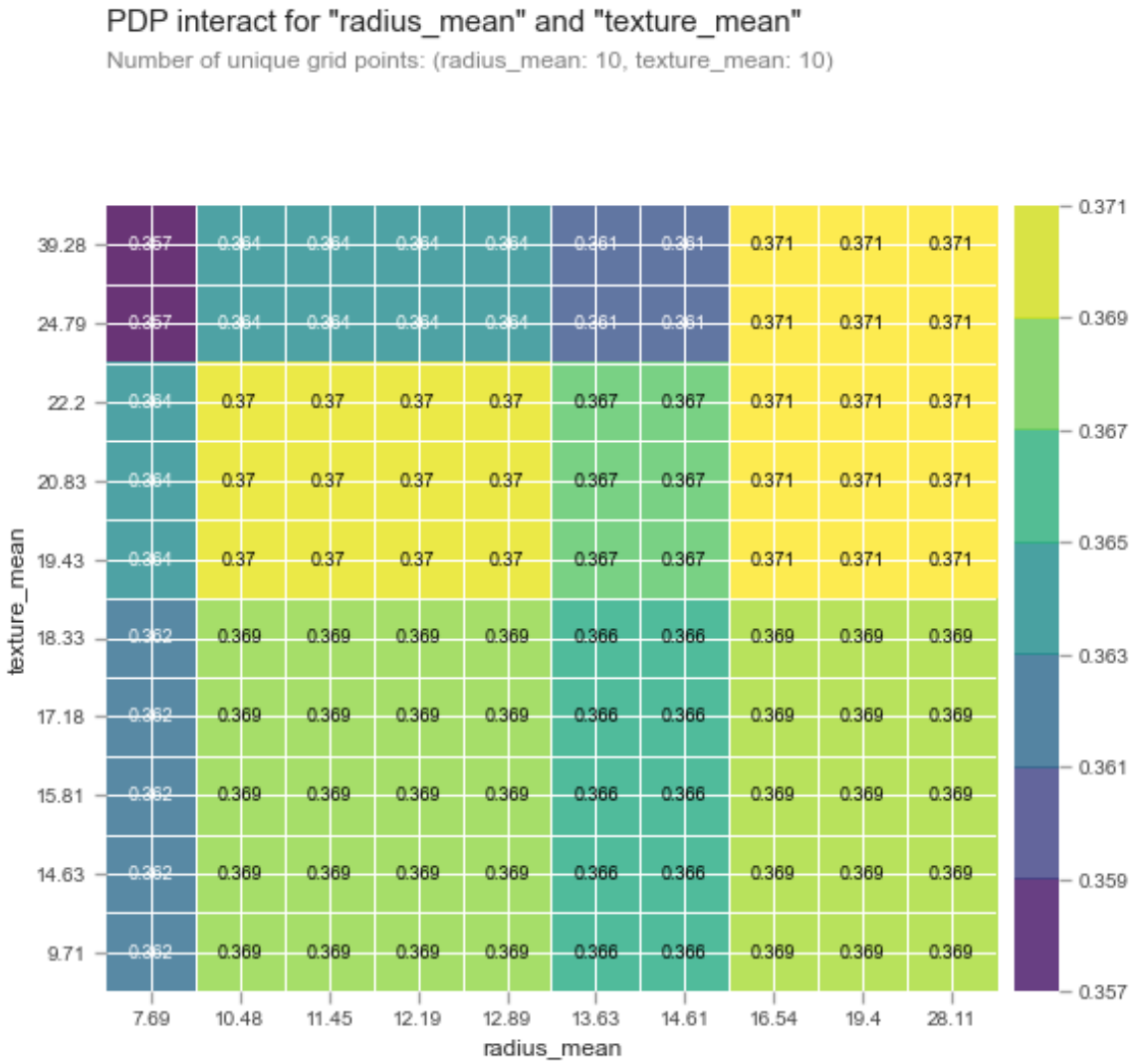

Figure 5.30: Bivariate PD Plot Random Forest Model for cancer Dataset.

As we can see to the right of the graph, we have a bar that indicates the probability of having a good quality wine, the best probability is in dark purple and the best probability of having a good quality wine is in yellow. Now when we have a graph showing the combination of features (on the horizontal axis we have the radius_mean and on the vertical axis we have texture_mean), it is important to do an analysis of what is happening, so let's start with the worst Of the cases, when the radius_mean has a value of 7.69 and the texture_mean has a value of 39.28, the probability that the tumor is malignant is 25.7 %, on the other hand, we would tend that if the radius_mean has a value of 16.54 and the texture_mean has a value of 39.28, we would have the probability that the tumor is malignant of 37.1%. This probability is obtained by this model with those characteristics that we have taken

considering that they are not characteristics that are very bad and do not differ much from each other.

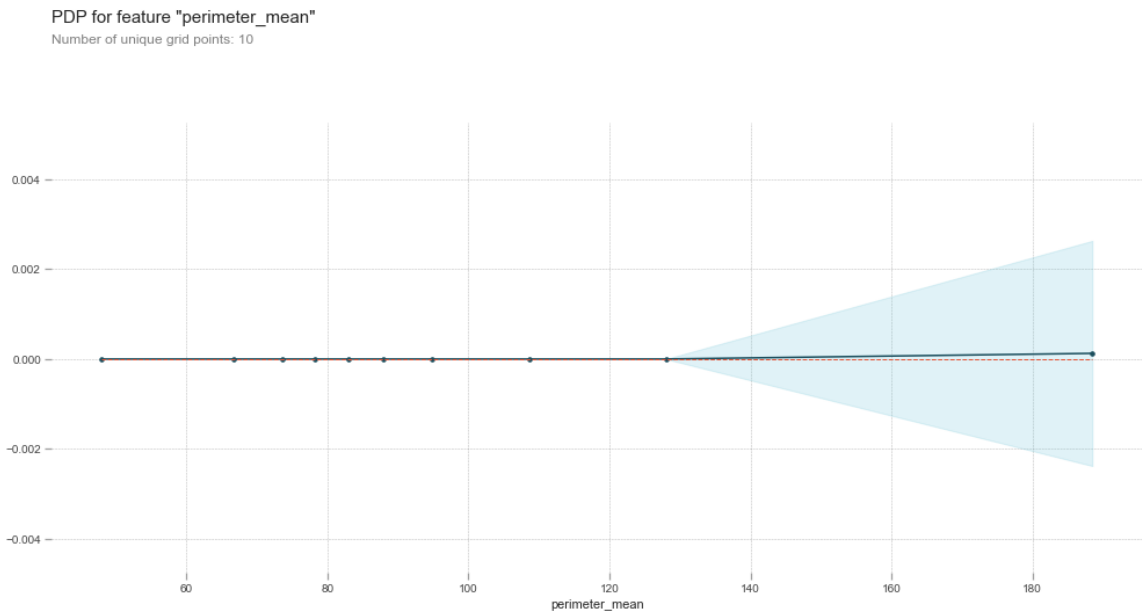

Figure 5.31: PDP's KNN Model for cancer Dataset.

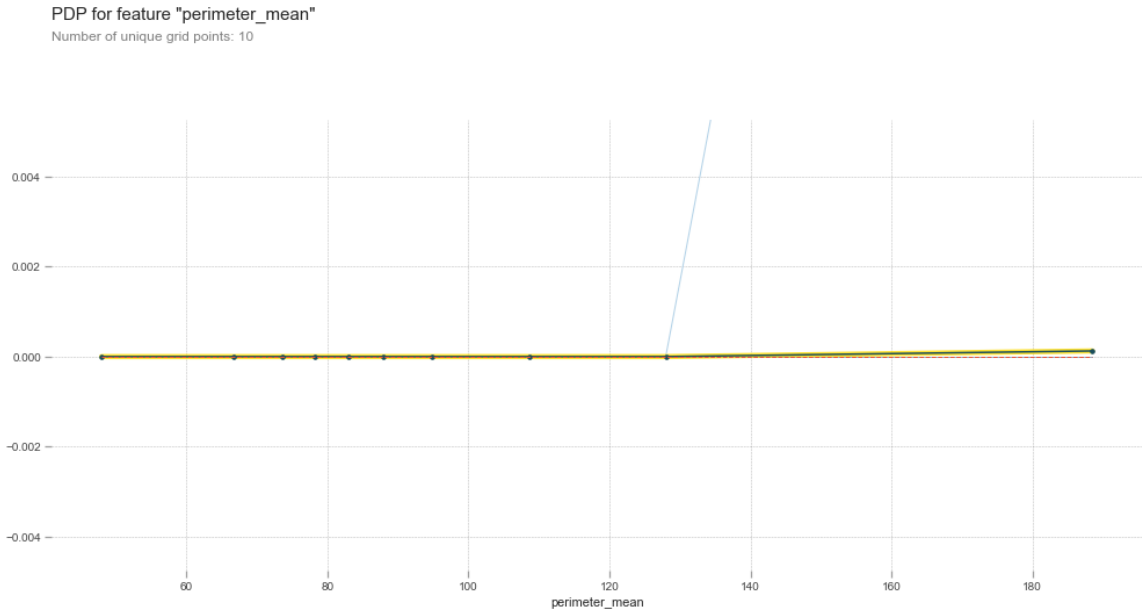

Figure 5.32: ICE Plot KNN Model for cancer Dataset.

In the knn model, we can identify this time that a difference of the graphs that were obtained with the data set came now with the cancer data set because it behaves differently, for now, the mean_perimeter characteristic remains constant in the influence it has on decision making.

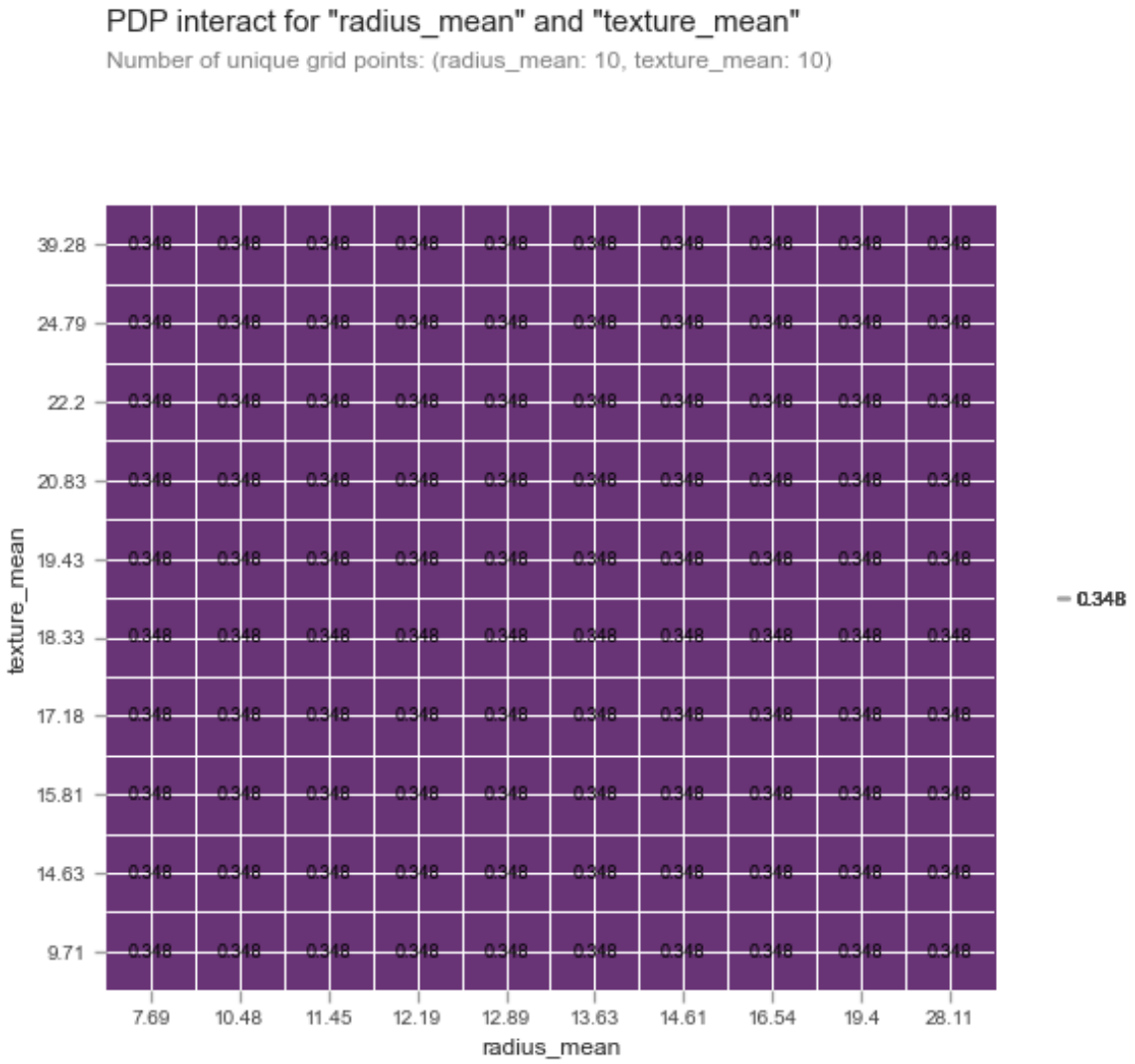

Figure 5.33: Bivariate PD Plot KNN Model for cancer Dataset.

This model, as mentioned before, if you do not have an idea of what is happening, can be misinterpreted, for the model the radius _average and the texture _average are inconsequential, that is, it does not matter when one or the other varies the percentage to have a malignant cancer will always be 34.8 %

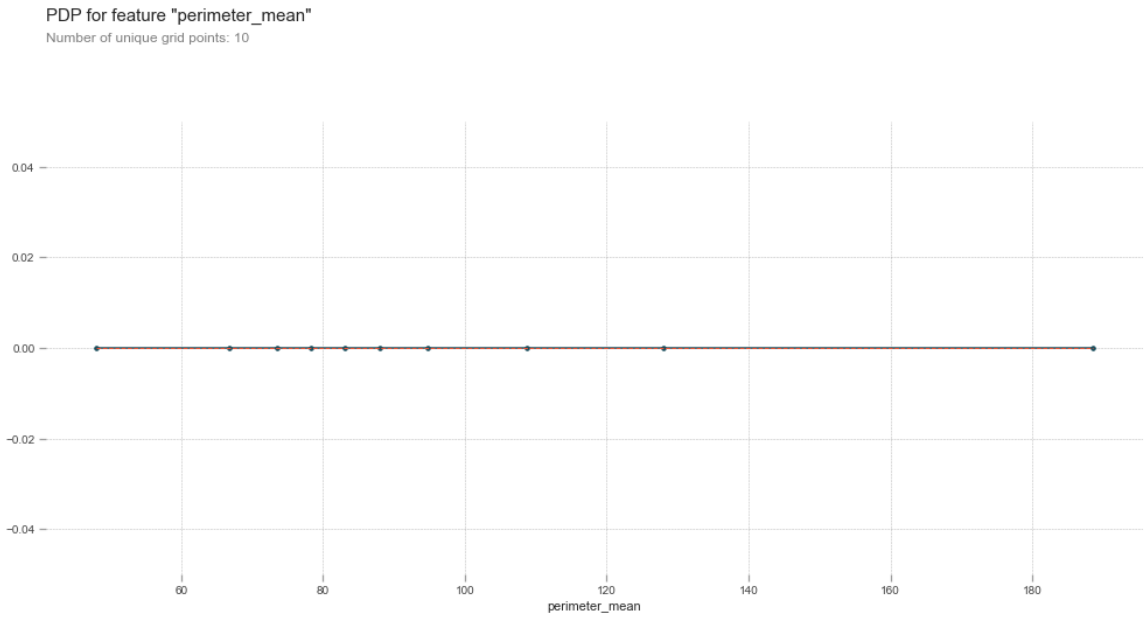

Figure 5.34: PDP's Naevi Bayes Bernoulli Model for cancer Dataset.

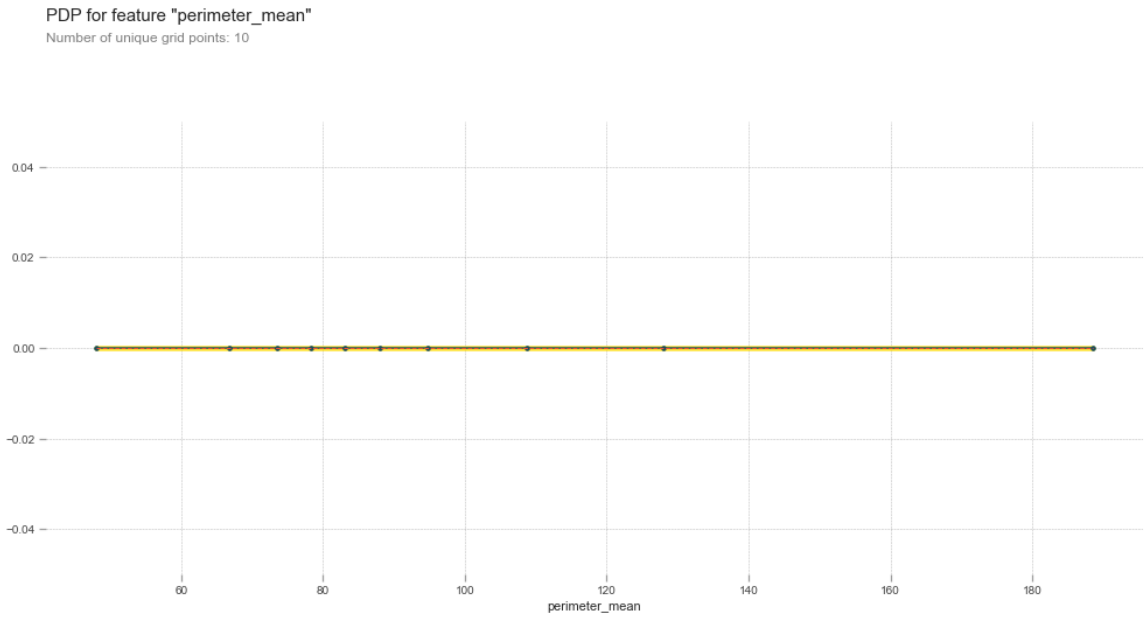

Figure 5.35: ICE Plot Naevi Bayes Bernoulli Model for cancer Dataset.

As we can consider in the two graphs, it is interesting to see how a model can change with some data and how another model can be kept in its classification perspective, but now the two models that show disinterested or are not affected in any way by the selected characteristic for analysis

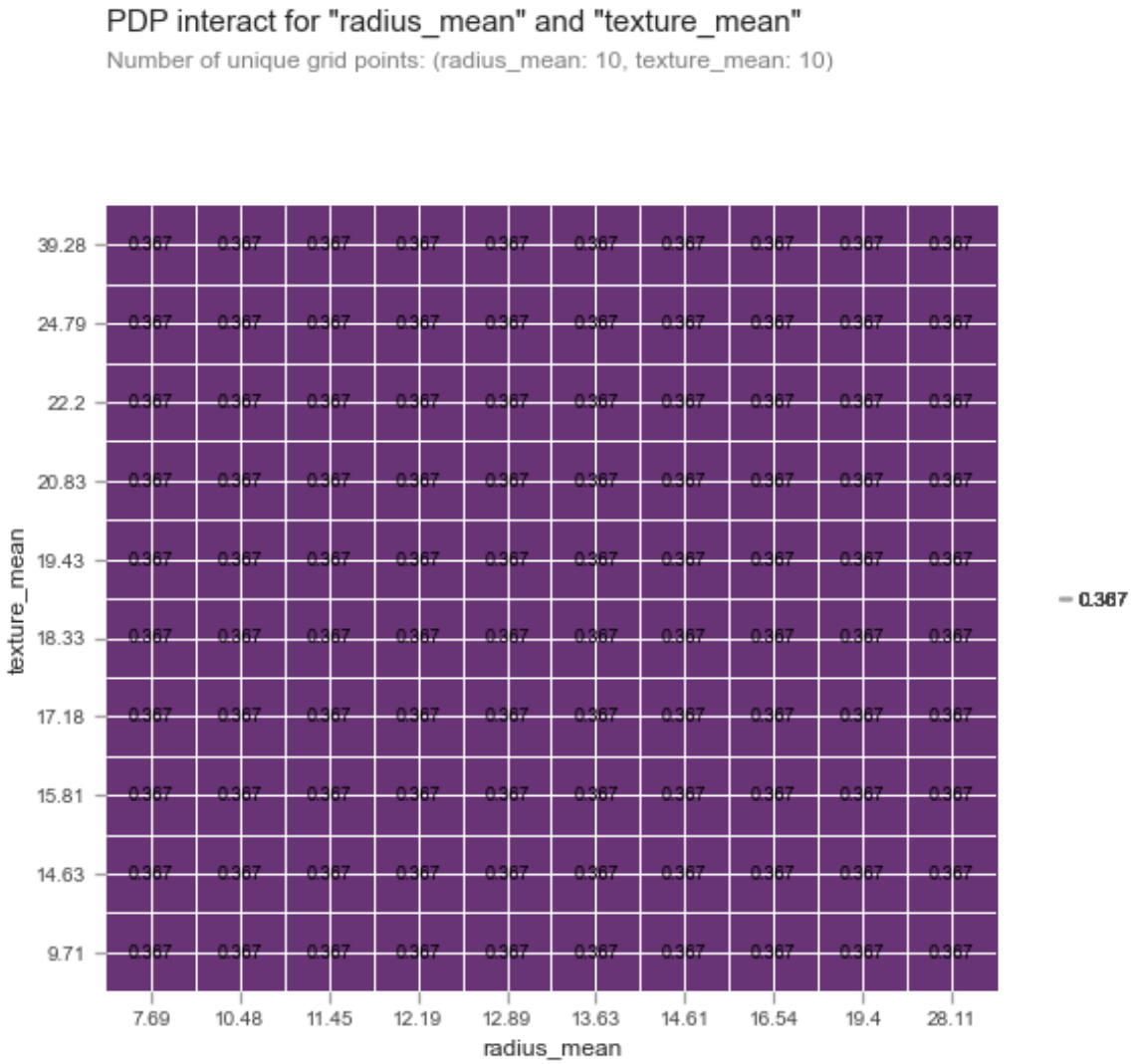

Figure 5.36: Bivariate PD Plot Naevi Bayes Bernoulli Model for cancer Dataset.

likewise with this model, if you do not have an idea of what is happening, it can be misinterpreted, for the model the radius _average and texture _average are inconsequential, that is, it does not matter when one or the other varies the percentage of having a malginant cancer will always be 36.7 % being a better probability than the previous model

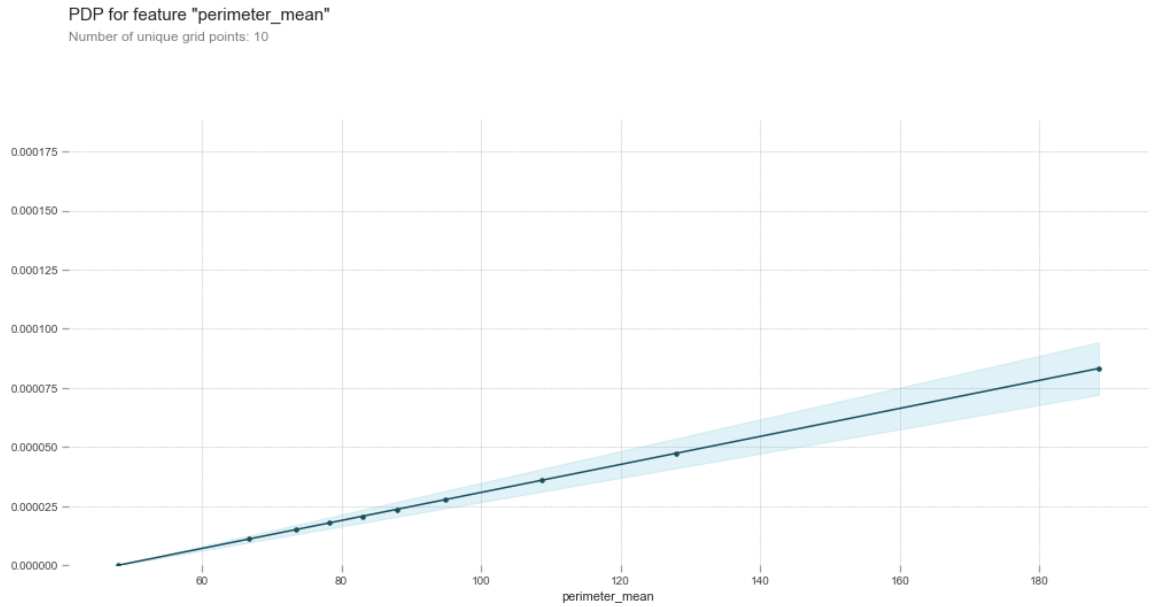

Figure 5.37: PDP's Naevi Bayes Gaussian Model for cancer Dataset.

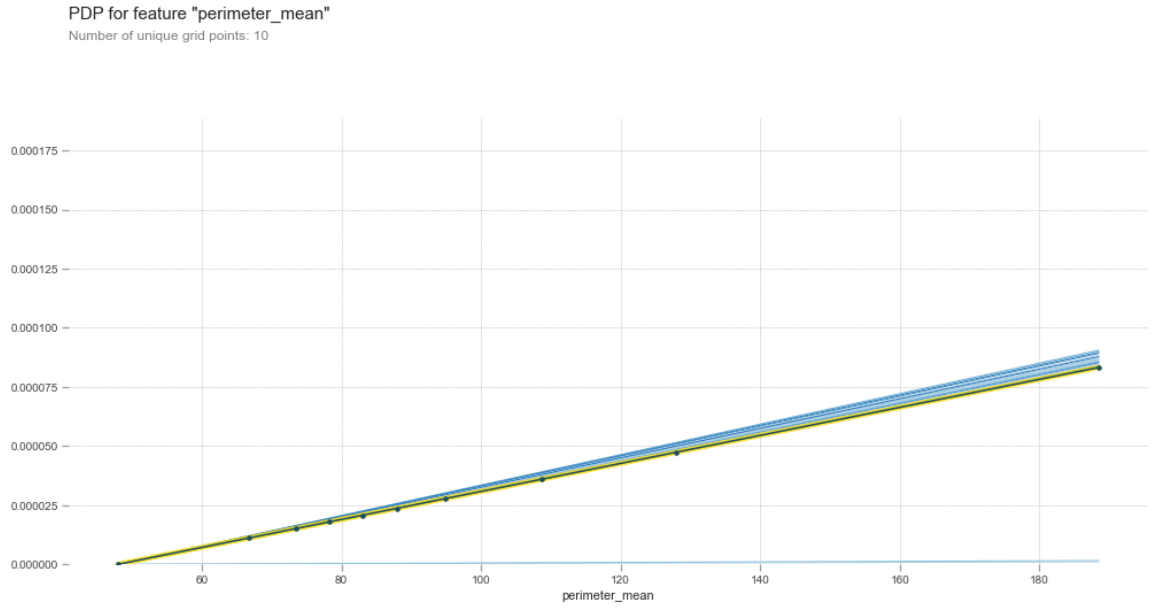

Figure 5.38: ICE Plot Naevi Bayes Gaussian Model for cancer Dataset.

In this case, the model is clearly indicating that the characteristic has a positive influence and linear growth, that is to say, that the more the percentage of the perimeter grows, in the same way resulting from the probability that the tumor is malignant, it will increase

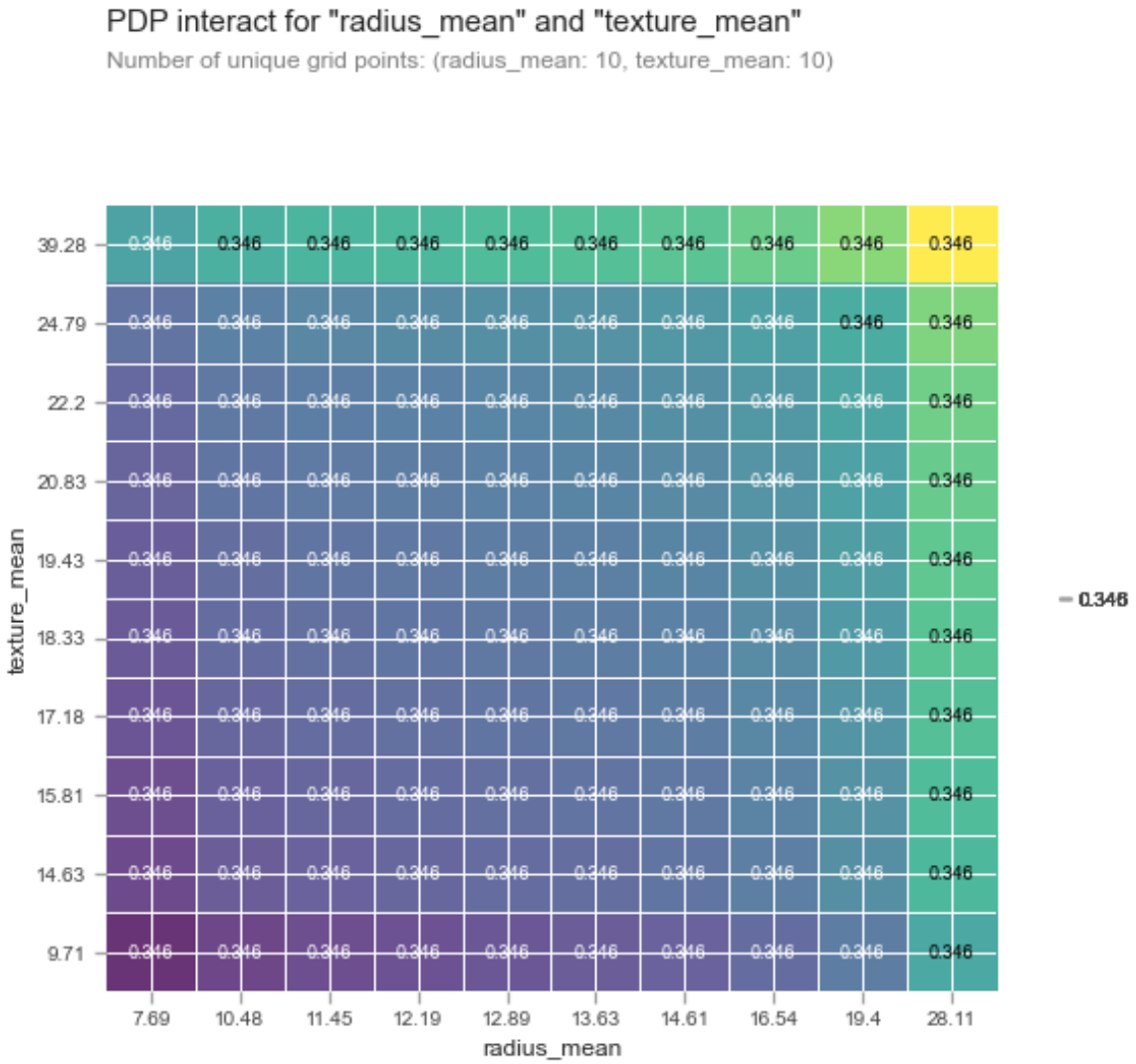

Figure 5.39: Bivariate PD Plot Naevi Bayes Gaussian Model for cancer Dataset.

For the Gaussian model, as can be seen in the figure, it has its worst probability at the beginning of the values for both radius _mean and texture _mean, being that when the radius is equal to 7.69 and texture 9.71, the probability that the tumor it is malignant is 31.6 % and when the radius and texture values reach their highest points the probability of having a malignant tumor is 34.6

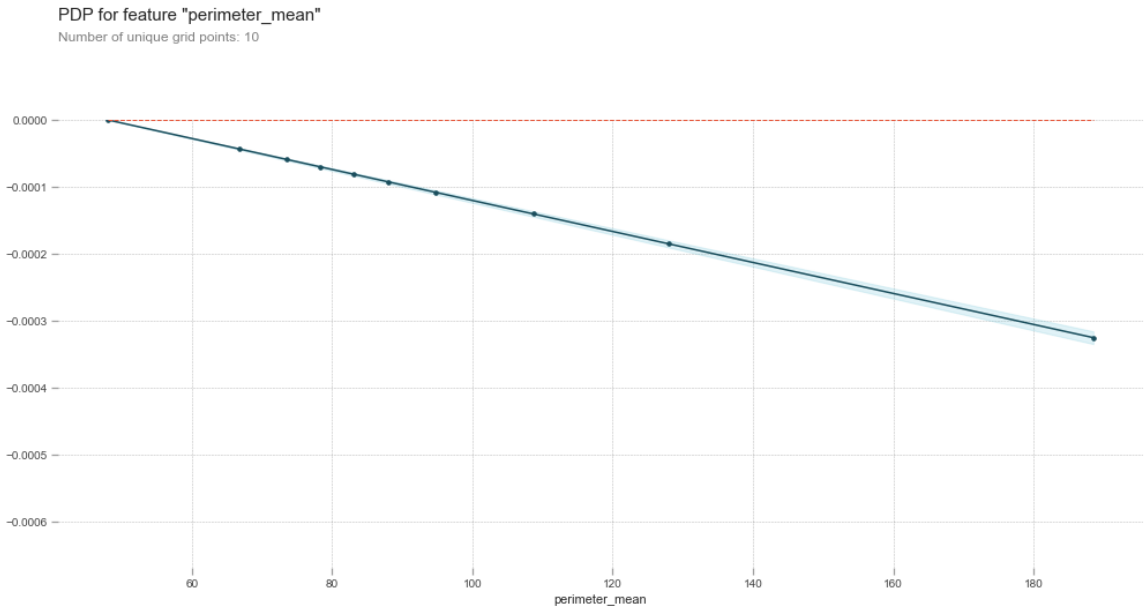

Figure 5.40: PDP's Logistics Regression Model for cancer Dataset.

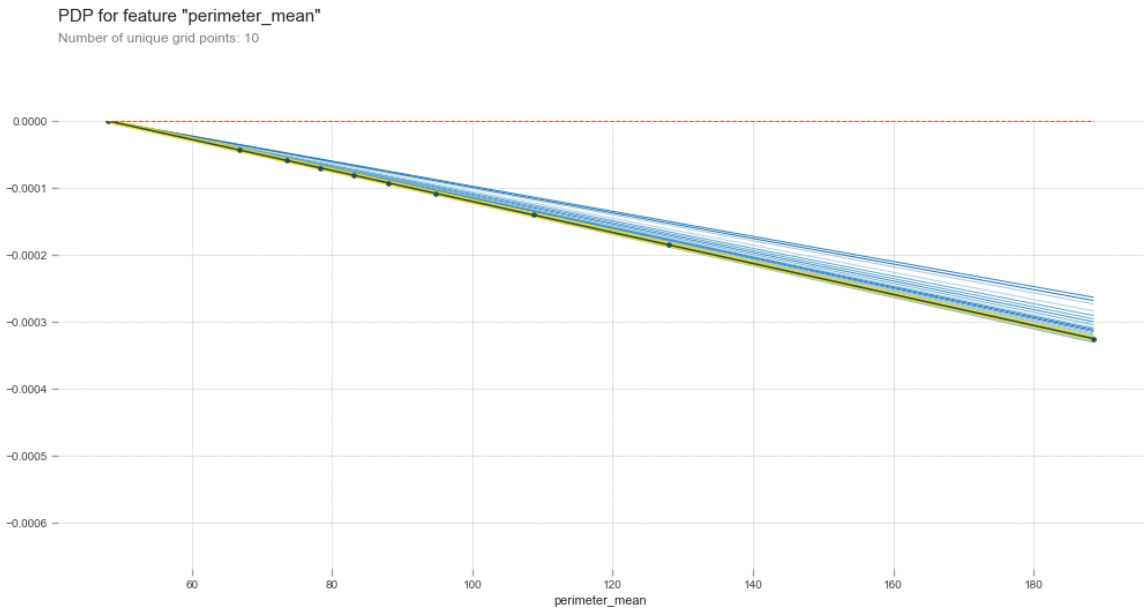

Figure 5.41: ICE Plot Logistics Regression Model for cancer Dataset.

something very interesting, since the graph has given us something different from the previous one that did not It is surprising since it is something that can be expected when resecting a regression, we see a purely linear interaction as we already said. but this does not mean that the model works poorly, on the contrary, it is interesting to know that

for the model if the value of the perimeter decreases, it increases the probability that the tumor is malignant, this is something interesting that medical specialists could give a more explanation deep

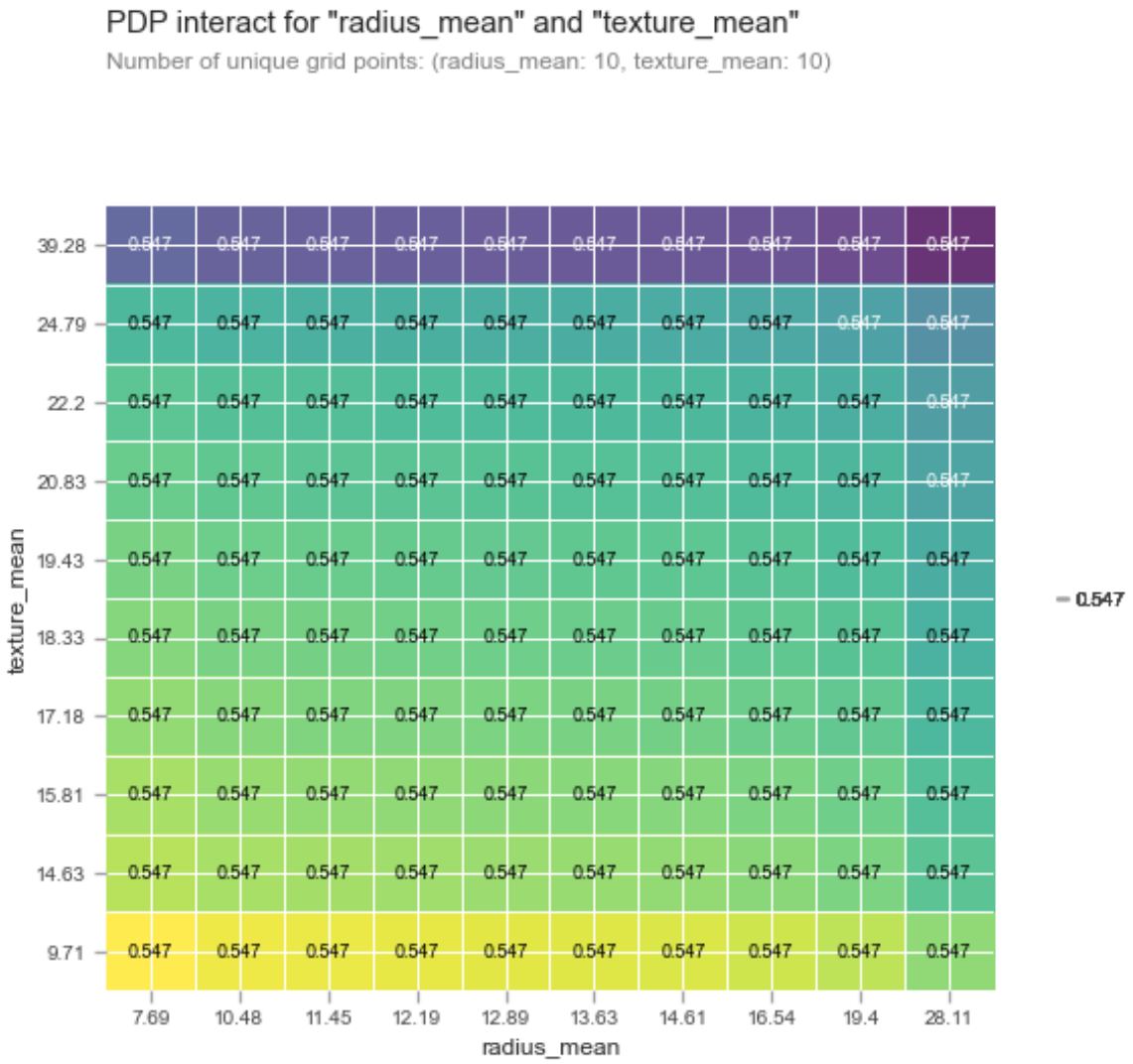

Figure 5.42: Bivariate Plot Plot Logistics Regression Model for cancer Dataset.

When you have graphs like the one in this model in which apparently the values do not change, we can say that they do change but in a very small way, so it seems that there is always a value of 54.7 probability that the tumor is malignant, but you can also see these data are inversely proportional the smaller the value of the radius and the larger the value of the texture the probability will be better

5.4 Comparative Analysis of XAI approaches: Specific Analysis

There are several techniques to explain AI that today give good results and that are at the forefront but that are a little difficult for a common user to understand, in our work we have taken two of them that are analyzed separately from the others since by their nature it is not possible to implement them in all models.

5.4.1 Explaining the classification decisions our models have made with ELI5 for the two datasets

Next, we will use the ELI5 tool that basically uses the same mechanisms for the calculation of native characteristics, using the "gain" parameter that we already have predetermined that we used previously to know which characteristic was the most important or with more relevance for both datasets and that with that we work on the previous point. ELI5 gives us an easy and ergonomic way to display it with the eli5.show_weights method. and from there we will start doing our analysis

Weight	Feature
0.2116 ± 0.0935	x10
0.1095 ± 0.0663	x9
0.1074 ± 0.0900	x1
0.1003 ± 0.0532	x6
0.0866 ± 0.0427	x4
0.0735 ± 0.0442	x0
0.0688 ± 0.0291	x5
0.0664 ± 0.0507	x7
0.0616 ± 0.0309	x8
0.0592 ± 0.0419	x3
0.0551 ± 0.0243	x2

Figure 5.43: Characteristic weights in the wine dataset with Random Forest.

y=1 top features

Weight[?]	Feature
+1.749	x9
+0.949	x10
+0.018	x5
-0.007	x3
-0.017	x6
-0.036	x0
-0.442	x2
-1.142	x7
-1.144	<BIAS>
-1.285	x4
-1.897	x8
-2.261	x1

Figure 5.44: Characteristic weights in the wine dataset with Logistics Regression.

Weight	Feature
0.2170 ± 0.5905	x1
0.1728 ± 0.5863	x25
0.1056 ± 0.4228	x29
0.0748 ± 0.4488	x22
0.0715 ± 0.4112	x24
0.0667 ± 0.3547	x5
0.0667 ± 0.3803	x2
0.0632 ± 0.3513	x28
0.0530 ± 0.1406	x9
0.0352 ± 0.2059	x7
0.0098 ± 0.0398	x8
0.0095 ± 0.0341	x23
0.0079 ± 0.0289	x15
0.0077 ± 0.0252	x31
0.0053 ± 0.0282	x19
0.0046 ± 0.0093	x6
0.0045 ± 0.0270	x17
0.0043 ± 0.0151	x13
0.0039 ± 0.0149	x14
0.0029 ± 0.0124	x16
… 12 more …	

Figure 5.45: Characteristic weights in the cancer dataset with Random Forest.

y=1 top features

Weight[?]	Feature
+0.000	x25
+0.000	x5
+0.000	x15
+0.000	x1
+0.000	x14
+0.000	x28
+0.000	x12
+0.000	x8
… 4 more positive …	
… 9 more negative …	
-0.000	x6
-0.000	x26
-0.000	x10
-0.000	x30
-0.000	<BIAS>
-0.000	x13
-0.000	x22
-0.000	x2
-0.000	x3
-0.000	x24
-0.000	x23
-0.000	x4

Figure 5.46: Characteristic weights in the cancer dataset with Logistics Regression.

With this, as we said, we have obtained the weights that are associated with each characteristic that each model considers important, in addition to that you can see the contribution of each characteristic to be able to classify y = 1 in one case it will be that characteristics predict a good quality wine and in the other case it will be that characteristics predict that a tumor is malignant.

But, the names of the characteristics are not shown, for this you could pass a list with the names of the characteristics to ELI5.

Now so that the models can be more interpretable for both data sets, each prediction made by the model should be considered as a sum of total contributions of characteristics (including bias), which indicates how the characteristics lead us to a particular classification. What ELI5 does is show the weights of each characteristic, and this will tell us how influential one characteristic or another could have been to contribute to the final decision of classification in the models.

Now images of both data sets will be shown but first we will analyze the individual predictions, first when it comes to poor quality wines and benign tumor classes We will put these interpretations in a context of value ”0”

y=0 (probability 0.600) top features

Contribution[?]	Feature	Value
+0.476	<BIAS>	1.000
+0.139	sulphates	0.520
+0.116	alcohol	9.500
+0.090	residual sugar	1.900
+0.067	pH	3.380
-0.015	chlorides	0.075
-0.016	free sulfur dioxide	16.000
-0.018	volatile acidity	0.680
-0.040	total sulfur dioxide	51.000
-0.052	fixed acidity	7.100
-0.066	citric acid	0.070
-0.082	density	0.997

Figure 5.47: Predicting when a particular wine quality will be 'Low Quality' in the with Random Forest.

In this individual prediction we can see that the 4 main and influential characteristics after the bias that are affecting the model to make their classification in this case, we have influences that seem to be, after the bias, alcohol, pH, total sulfur dioxide and chlorides

y=0 (probability 0.779, score -1.257) top features

Contribution[?]	Feature	Value
+6.412	pH	3.380
+1.537	volatile acidity	0.680
+1.144	<BIAS>	1.000
+1.139	density	0.997
+0.846	total sulfur dioxide	51.000
+0.253	fixed acidity	7.100
+0.096	chlorides	0.075
+0.031	citric acid	0.070
+0.013	residual sugar	1.900
-0.285	free sulfur dioxide	16.000
-0.909	sulphates	0.520
-9.019	alcohol	9.500

Figure 5.48: Predicting when a particular wine quality will be 'Low Quality' in the with Logistics Regression.

On the other hand, in this graph we can see that alcohol is no longer relevant to the model, now, however, we have more characteristics that provide value for decision-making, including bias in this opportunity and for the model, the characteristics that affect alcohol. Decision making are pH, volatile acidity, density, total sulfur dioxide, fixed acidity, chlorides, citric acid, and residual sugar

y=1 (probability 1.000) top features

Contribution[?]	Feature	Value
+0.357	<BIAS>	1.000
+0.123	area_worst	1403.000
+0.116	diagnosis	1.000
+0.066	area_mean	1094.000
+0.065	perimeter_worst	139.900
+0.063	concave points_worst	0.149
+0.061	radius_worst	21.310
+0.057	radius_mean	18.610
+0.053	concavity_worst	0.345
+0.031	compactness_mean	0.107
+0.004	area_se	93.540
+0.003	fractal_dimension_worst	0.074
+0.001	smoothness_worst	0.134
+0.001	texture_worst	27.260

Figure 5.49: Predicting when a particular tumor type will be 'benign' in the with Random Forest.

When working with more variables that have a more collaborative influence on the model, we can also have the number of characteristics that collaborate for decision-making, they also increase, in this case we will have 13 characteristics that somewhat influence the decision-making process. of decisions.

y=1 (probability **0.589**, score **0.358**) top features

Contribution[?]	Feature	Value
+0.277	area_worst	1403.000
+0.083	area_mean	1094.000
+0.001	area_se	93.540
+0.000	perimeter_se	5.632
+0.000	diagnosis	1.000
+0.000	radius_se	0.853
+0.000	concavity_worst	0.345
+0.000	concavity_mean	0.149
+0.000	compactness_worst	0.212
+0.000	concave points_worst	0.149
+0.000	concave points_mean	0.077
+0.000	compactness_mean	0.107
-0.000	fractal_dimension_se	0.004
-0.000	concave points_se	0.019
-0.000	smoothness_se	0.011
-0.000	compactness_se	0.027
-0.000	concavity_se	0.051
-0.000	symmetry_se	0.023
-0.000	fractal_dimension_mean	0.057
-0.000	fractal_dimension_worst	0.074
-0.000	smoothness_mean	0.094
-0.000	smoothness_worst	0.134
-0.000	symmetry_mean	0.170
-0.000	symmetry_worst	0.234
-0.000	<BIAS>	1.000
-0.000	texture_se	1.849
-0.000	radius_worst	21.310
-0.000	radius_mean	18.610
-0.000	texture_mean	20.250
-0.000	texture_worst	27.260
-0.001	perimeter_worst	139.900
-0.001	id	852781.000
-0.001	perimeter_mean	122.100

Figure 5.50: Predicting when a particular tumor type will be 'benign' in the with Logistics Regression.

When the model is changed, everything changes in this case, although more features are shown that collaborate, since their collaboration is insignificant.In reality, two characteristics are those that make the most contributions to make the final decision of whether a tumor is benign, in this case the worst area and mean area

Finally, images from both data sets will be shown, but first we will analyze the individual predictions, first when it comes to wines of good quality and classes of malignant tumors. We will put these interpretations in a context of value "1"

y=1 (probability 0.900) top features

Contribution[?]	Feature	Value
+0.524	<BIAS>	1.000
+0.178	chlorides	0.062
+0.074	sulphates	0.650
+0.056	citric acid	0.320
+0.045	fixed acidity	7.300
+0.033	residual sugar	2.100
+0.020	pH	3.300
+0.014	density	0.997
+0.011	volatile acidity	0.480
+0.010	total sulfur dioxide	54.000
+0.001	alcohol	10.000
-0.067	free sulfur dioxide	31.000

Figure 5.51: Predicting when a particular wine quality will be 'High Quality' in the with Random Forest.

When decisions are made about choosing whether a wine is of good quality, we can choose that the collaboration changes drastically when you see a class individually, the model behaves in a way that alcohol did not show before, now it is the characteristic that contributes the least and the top 4 of characteristics without counting the bias would remain with chlorides, sulfates, citric acid, and fixed acidity.

y=1 (probability 0.541, score 0.163) top features

Contribution[?]	Feature	Value
+9.494	alcohol	10.000
+1.137	sulphates	0.650
+0.552	free sulfur dioxide	31.000
-0.014	residual sugar	2.100
-0.080	chlorides	0.062
-0.142	citric acid	0.320
-0.260	fixed acidity	7.300
-0.896	total sulfur dioxide	54.000
-1.085	volatile acidity	0.480
-1.139	density	0.997
-1.144	<BIAS>	1.000
-6.260	pH	3.300

Figure 5.52: Predicting when a particular wine quality will be 'High Quality' in the with Logistics Regression.

Everything changes with the change of model, as we can affect in the figure, with logistic regression, alcohol is the one that makes the most contributions, it is only important for this model when it makes the decision if the wine is of good quality, although again it can See that sulfates appear in the equation.

y=0 (probability 0.900) top features

Contribution[?]	Feature	Value
+0.643	<BIAS>	1.000
+0.231	diagnosis	0.000
+0.036	concave points_worst	0.115
+0.029	radius_worst	17.180
+0.027	radius_mean	14.920
+0.026	compactness_mean	0.085
+0.025	concave points_mean	0.032
+0.024	perimeter_worst	112.000
+0.020	area_mean	686.900
+0.007	concavity_mean	0.055
+0.001	texture_se	0.433
+0.001	texture_worst	18.220
+0.001	fractal_dimension_mean	0.057
+0.001	area_se	23.310
+0.001	fractal_dimension_worst	0.083
+0.000	id	911384.000
-0.052	concavity_worst	0.315
-0.122	area_worst	906.600

Figure 5.53: Predicting when a particular tumor type will be 'malignant' in the with Random Forest.

As we can see when the model is classifying if a tumor is malignant, the diagnosis is also considered although it is irrelevant as bias, which leaves us with characteristics such as concave points, worse, the mean radius and the worst radius that the model contributes to know how to classify. everything.

y=1 (probability **0.557**, score **0.228**) top features

Contribution[?]	Feature	Value
+0.179	area_worst	906.600
+0.052	area_mean	686.900
+0.000	area_se	23.310
+0.000	perimeter_se	1.826
+0.000	concavity_worst	0.315
+0.000	radius_se	0.245
+0.000	compactness_worst	0.279
+0.000	concave points_worst	0.115
+0.000	concavity_mean	0.055
+0.000	concave points_mean	0.032
+0.000	compactness_mean	0.085
-0.000	fractal_dimension_se	0.002
-0.000	smoothness_se	0.003
-0.000	concave points_se	0.008
-0.000	compactness_se	0.018
-0.000	concavity_se	0.023
-0.000	symmetry_se	0.011
-0.000	fractal_dimension_mean	0.057
-0.000	fractal_dimension_worst	0.083
-0.000	smoothness_mean	0.081
-0.000	smoothness_worst	0.106
-0.000	symmetry_mean	0.169
-0.000	symmetry_worst	0.269
-0.000	texture_se	0.433
-0.000	<BIAS>	1.000
-0.000	radius_worst	17.180
-0.000	radius_mean	14.920
-0.000	texture_mean	14.930
-0.000	texture_worst	18.220
-0.001	perimeter_worst	112.000
-0.001	perimeter_mean	96.450
-0.001	id	911384.000

Figure 5.54: Predicting when a particular tumor type will be 'Malignant' in the with Logistics Regression.

in the same way that it did before the model does again each and every one of its features are contributing something of some value however again the worst area and the area means son the best at the time of inflicting the taking of the decision in the classification

5.4.2 Explaining the classification decisions our models have made with SHAP for the two datasets

Now we will explain the decision making process SHAP (explanations of additive SHapley) which is an approach that is based on the theory of games, and does so by giving local explanations, what it does is to merge several methods of attribution of characteristics that add up, they are consistent

SHAP values can explain the output of AI / ML models, but care must be taken with the data that is entered, since if you work with an overly complex data set, shap tends to consume a lot of machine resources and turns into slow work SHAP has implemented C ++ characteristics that support tree models, that is why it will be analyzed in our Random Forest model, what shap does is assign to each characteristic a value of importance to obtain a particular prediction, such as other techniques that are already known do it, the novelty in shap are some components such as:

- Identify a new class of measure of the importance of additive characteristics
- Theoretical results capable of showing that there is a unique solution in a class with a group of characteristics that are desirable.

Specifically, the values displayed by shap try to explain the result of our model or the result that our model has generated, but what is shown as the sum of the effects of each characteristic introduced in the conditional expectation. Importantly, for nonlinear functions, the order of the input characteristics is important. To get the shap value, all you have to do is calculate the average of all possible values. Game theory tests show that this is the only consistent method possible.

To understand the form more easily, the following analogy will be made, assuming that the values of the characteristics are entered in a quarter in random order. There is a game in the room where all the characteristics are involved and involved in one way or another, which means that all the values are useful for prediction. The Shapley value ϕ_{ij} is the average marginal contribution from the value of the feature x_{ij} the feature entered the room before joining,that is.

$$\phi_{ij} = \sum_{all.orderings} val(\{features.before.j\} U x_{ij}) - val(\{features.before.j\})[100] \quad (5.6)$$

We are now going to explain shap for our implemented random forest model and for the two data sets in order to observe some results and be able to explain them in the best possible way.

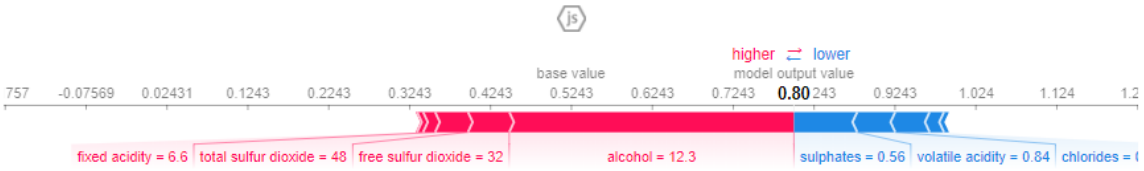

Figure 5.55: SHAP Graph for Random Forest Model in Wine Dataset.

The graphs that we have shown give us the answer when using SHAP, so let's analyze what this graph means. We predicted a value of 0.8, and the base value that we have is 0.5243, as we can see the values of the characteristics that increase the predictions are in a pink color, and the size of each characteristic indicates the magnitude of the effect of the characteristic on the model. on the other hand, the values of the characteristics that reduce the reduction are in blue. and as we can also see alcohol has the greatest impact on the model when its value is 12.3. Although the value of the sulphates has an effect that decreases significantly in prediction. Now if we subtract the lengths of the blue bars from the lengths of the pink bars, then, we know that is equal to the distance from the base value to the model output

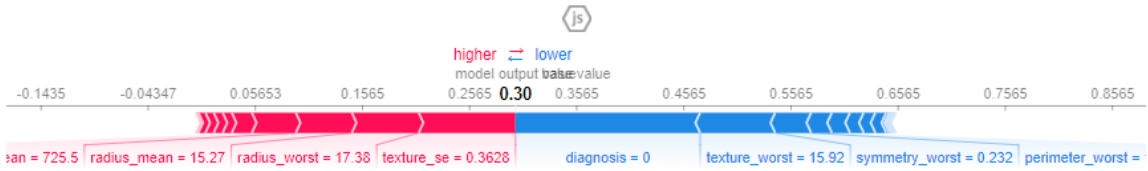

Figure 5.56: SHAP Graph for Random Forest Model in cancer Dataset.

On the other hand, we have the graph that corresponds to the same model using the dataset of the types of cancer, let's analyze what this graph means.

it is predicted that to have a benign tumor we will have a value of 0.3, and the base value we have is 0.1565, just as before we can say that the values of the characteristics that increase the predictions are pink, and the size of each characteristic indicates the magnitude of the effect. of the feature in the model. On the other hand, the values of the characteristics that reduce the reduction are in blue. And as we can also see, this time the feature texture_se is the characteristic that has the most impact on the decision-making of the 0.3628 model. Although the texture_worst value has an effect that significantly decreases the prediction. Now if we subtract the lengths of the blue bars from the lengths of the pink bars, then we know that it is equal to the distance from the base value to the model output

Having the importance of permutation is something that will help us a lot because in this way simple numerical measures have been created, and with them we can see once again the characteristics that mattered in the model. With this it was possible to make comparisons between the characteristics in an easy way and thus it was possible to obtain representations that although we can understand it as engineers, for people who do not have technical knowledge it will not be difficult to interpret, these graphs, in any case, do not They said how important a characteristic was, if one of those characteristics was of medium permutation importance it could be said to have a high effect for certain predictions but no effect in general, or at best a medium effect for all predictions. So here are the SHAP summary graphics that give us a panoramic view of the importance of the characteristics and what they affect. Let's see the following graphs and analyze them

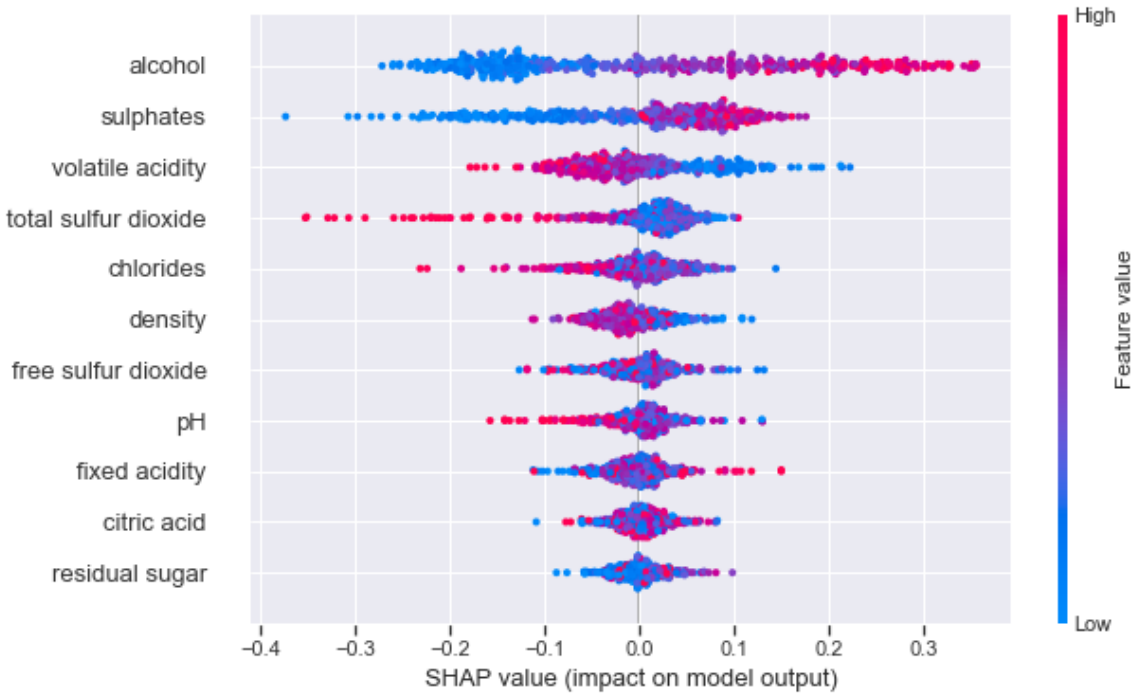

Figure 5.57: SHAP Summary Plot for Random Forest Model in Wine Dataset.

In this plot many points can be observed and each one of them has 3 characteristics:

- on the vertical axis we can see what characteristic is being represented.
- The color as in the previous two figures shows if any characteristic was high or low for that row of the dataset
- the horizontal axis tells us if the effect that this characteristic had caused a higher prediction or a lower prediction

To get an idea of what the image shows us, we can see that the point in the upper right corner of pink color shows us that if a wine that had a greater amount of alcohol in its structure positively affected the prediction by 0.35

Some things that should be seen in addition are: High values of total sulfure dioxide caused lower predictions, and low values caused high predictions.

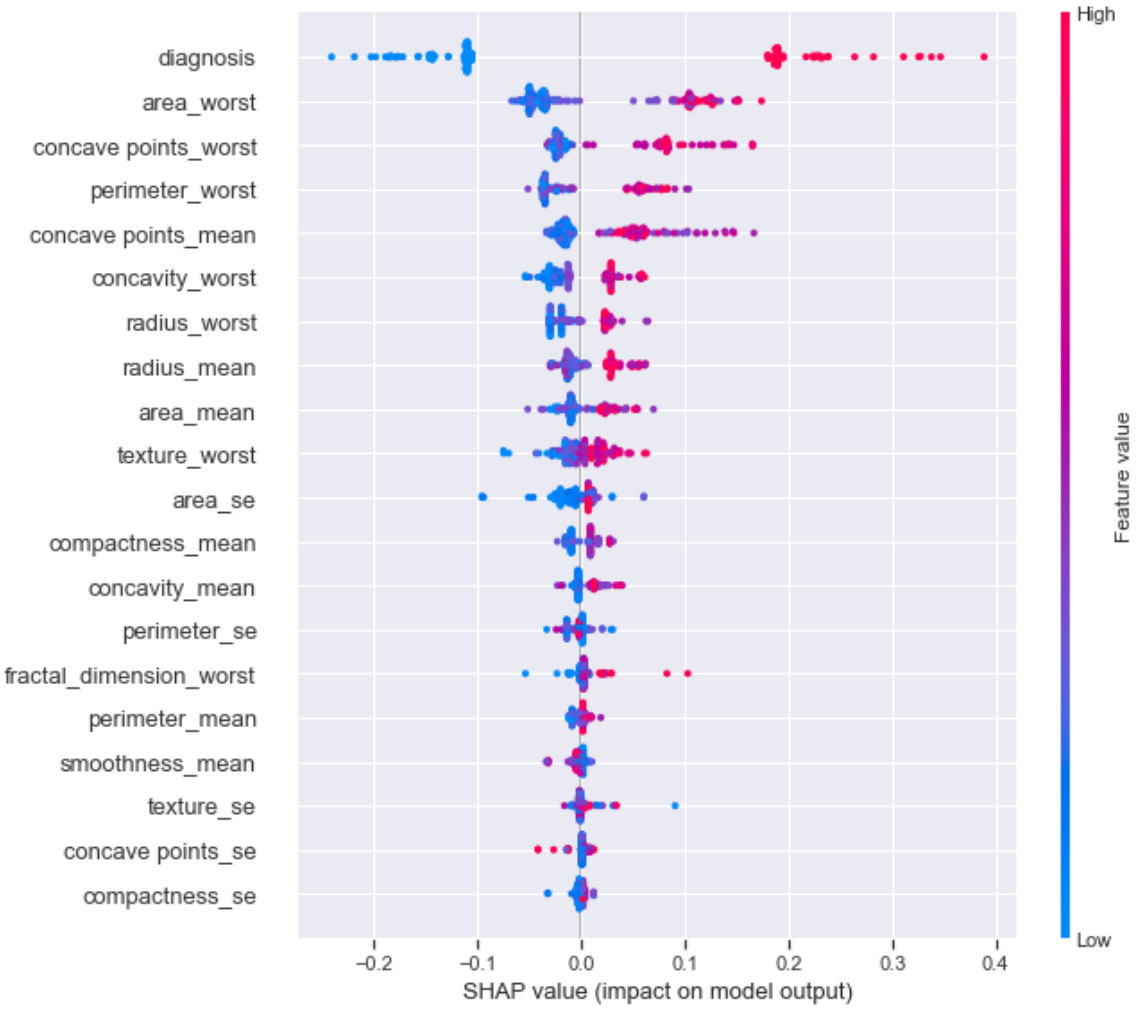

Figure 5.58: SHAP Summary Plot for Random Forest Model in Cancer Dataset.

In this figure and in a deliberate way, the results have been shown both malignant tumors and benign tumors, and that is what we are seeing benign tumors influence their decision making and malignant ones in the same way but it should not be taken into account, let's look rather in the area of characteristics is the blue color point on the left and we can see that without the area there is a lower value, negatively influencing the prediction by 0.1

They can also have SHAP charts in which they can see the effect of two characteristics on the data set, the value of a characteristic must be plotted against the shap value of that characteristic in several samples.

Shap dependency graphs closely resemble PD graphs . The scatter you have vertically is driven by interaction effects, so you can also choose a new feature to color your interactions

Figure 5.59: SHAP Dependence Contribution Plots for Random Forest Model in Wine Dataset.

To begin analyzing the graph, we must consider that each point represents a row of our data sets, on the horizontal axis we have the real value of our data sets, and on the vertical axis, it shows how that value affects the prediction. When I see that the graph is tilted upwards, it indicates that the more alcohol there is in the wine, the greater the model's prediction to say that a wine is of good quality.

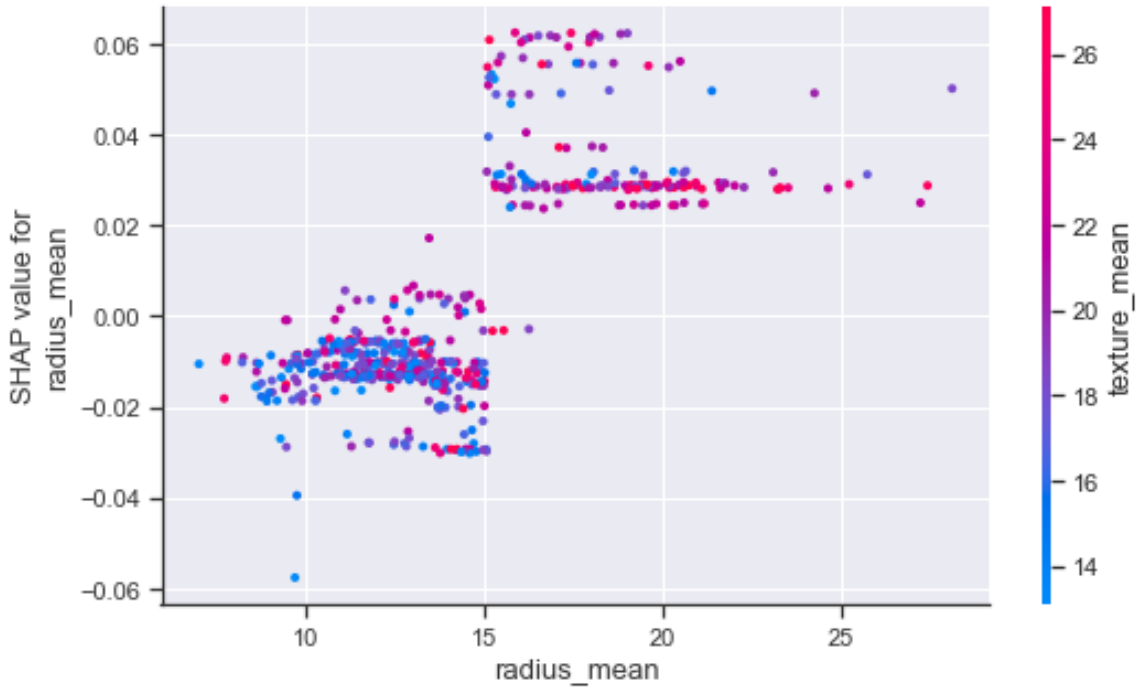

Figure 5.60: SHAP Dependence Contribution Plots for Random Forest Model in Cancer Dataset.

In this case, we can see that when the graph is tilted upwards, it indicates that the larger the area mean of the tumor, the greater the prediction of the model to say that it is a malignant tumor.

CHAPTER 6

Conclusions and Future Work

In this work, we have obtained as a result the ability to explain how certain characteristics influence so that AI/ML models can make their predictions, and thus somehow an XAI can be obtained. Specifically, the characteristics that each of the models have been taken advantage of to be able to make their decisions in such a way that we have not focused on the quality of the predictions that were made but rather, obtain from those values apply the techniques of explanation to understand why these situations occurred, trying to obtain results as visual as possible so that this is a job that in a certain way is easy to understand for people who do not have much technical knowledge but that is also completely complete for People who have knowledge about artificial intelligence. Finally, things that can be developed in the future will be proposed, as it is like waiting for this topic, since it is extensive and of great interest for the future, it still has a long way to go before it can be exploited and thus be useful for the scientific community, as well as For the general public

6.1 Conclusions

As it has been presented in chapter 2 of this thesis, the ancient methods that artificial intelligence used opened the doors for this complex world to develop, in this way from the first expert systems we have been able to demonstrate the technological development that has occurred in With this world, borders have been opened to innovation and not only that, but its expansion has been constant and outstanding at all times, from a machine that played chess a few years ago, to a system that is capable of distinguishing its image.

Some data entered is capable of evidencing, classifying or identifying the presence of cancer in the human body, but this is not yet just the beginning, and in the future artificial intelligence will gain more and more fields of study, more and more sectors of implementation, and for this, the human being must be prepared by his curious nature, he will always look for the why of things and this would not be otherwise, because he is The signs that lead to explainable Artificial Intelligence will be useful with strong AI models and reliable XAI techniques.

Likewise, the progress and advancement of Big Data are fundamental, taking into account that it is precisely there that we will have all the data that will be very useful for the development of artificial intelligence, not only of AI and known, but which also helps the growth of the explanatory techniques for these AI models, these data are so large and so complex that they grow and improve day by day, and that for this, increasingly complex data processing is required, in this way you could have:

- Increased amount of data volume
- The data could be more varied in order to be treated differently and with different motivations.
- The data will have better visibility in order to describe the nature and type of data that is available.
- Improvements in the speed of both sending and receiving data
- More truthful data and that improve your ugliness so that you will have more confidence at the time of your treatment
- With these data, we will be able to generate more value from the information we obtained from them.

With this, what we will only obtain will be to increase the need for AI and Big Data to achieve a dynamic in which work is more efficient wherever it is viewed and needed.

The artificial intelligence models that we have presented in this work have served as a basis for us to show the final results since we have selected traditional and widely used models in the scientific world and in the real world to be able to reach the decision of any question or to let an algorithm make a decision, it must be taken into consideration and as we could see.

In Chapter 4 when the performance graphs that each of the models had with each dataset were shown, that none behaved the same although it has been programmed with the same logics and with the same parameters, and although the previous treatment of the data was the most similar among the datasets, each model is its own world in which the decision and its precision depend a lot of the characteristics.

That is affecting him the most, and not only that but the number of characteristics that are offered to him for training, it is interesting to think about the evolution that AI models have evolved and even more about how they can continue to grow in the future and this for the well-being and development of science and applicability in the real world.

Similarly, when we started applying techniques that would explain the models we used, we could see that they are a world unto themselves, each showing how a feature was important to such a model in a way that could Being, of course, was made both for people who have knowledge of AI and for people who do not have the technical knowledge, since for this work we have selected techniques that are easy to understand, taking into account that these models are not all that exists today. But they have fulfilled the purpose of this thesis so that we could end up saying:

- A very good technique to explain the logic that an algorithm has had to predict are the partial dependency graphs, in this way we can have and extract information from the models that are referred to, these models, as said, can have an amazing complexity but Likewise, a very strong decision-making power that would be as amazing as it can be, these charts are very oriented to explain how they do it so that everyone can understand how a decision-making process was reached.

  In our case, it could be seen in both data sets how the selected characteristic of each dataset contributed to decision making, first individually and then collaboratively when the PDP bivariate graphs were shown, it was very easy to understand the impact that one or another characteristic had in the decision-making of each model, on some occasions we could see that this same characteristic did not affect the decision-making at all, which suggests that although a model or most models can work perfectly with

the characteristics, other models require a more exhaustive treatment and adjustment of their data and of themselves to try to show some more information.

- With ELI5 on the other hand, although its way of explaining seems a little more confusing, we can say that it is a great technique to be able to see what is happening within the model, that is, how each characteristic is contributing to decision-making If, then there are values that give more weight to that contribution, others that are neutral or it seems that they are not contributing a simple view because they affect in a very small way or already flat characteristics that are not considered with anything or that appear to be a contribution. negative towards the model. however, this technique is very easy because it can indicate many characteristics at the same time.

- Shap can be used as an analytical tool and works in such a way that you could say that it is an analyzer of the data, it allows us to show the individual instances that give us some predictions. In the end to explain the result was very easy and in a very intuitive way.

  The analysis can be implemented as part of a validation development, in this way we can provide transparency to the models in our case we only use the Random Forest model and, logically, the model is almost always valid if we look at the general precision that has been had with it, with shap we can detect how this precision happened which characteristics influence the model to have a certain precision value and other characteristics that weighed on the model are not so precise

6.2 Future Work

This thesis is infinitely expandable from any point of view, let's start with data collection, you can consider taking data sets that contain more data, that are of different types and that are applied to other fields, in this work we use classic datasets with data they are not extremely large but they are not a small thing, the treatment of the same data could only be done because it can take into account many aspects working with them, modifying them, creating new data that meets some characteristics of other data, among many other things.

From the point of view of the use of the models, in this work the models have been used which in the same way are some of the most used not only in the scientific field but also in the real world, the models could be trained in different ways, adding more depth, more neighbors or modifying its characteristics, that to mention something about it.

But in the techniques to be able to interpret them, there is greater flexibility since these techniques that we use are techniques that are aimed at a less technical audience, the same

techniques could be implemented with other approaches such as the SKATER interpretation model, in addition to You could modify the libraries that are used to be able to use all the techniques with all the models.

Bibliography

[1] primo.ai. Logistic regression (lr). urlhttp://primo.ai/index.php?title=Logistic_Regression_(LR), 2020. Accedido 2020-06-05.

[2] Computer Science at Cornell University. Bayes classifier and naive bayes. urlhttps://www.cs.cornell.edu/courses/cs4780/2017sp/lectures/lecturenote05.html, 2017. Accedido 2020-06-05.

[3] Adi Pamungkas. k-nearest neighbor (k-nn). urlhttps://pemrogramanmatlab.com/data-mining-menggunakan-matlab/k-nearest-neighbor-knn-menggunakan-matlab/, 2016. Accedido 2020-06-05.

[4] Unknow Author. Random forest – arboles de decisión machine learning. urlhttp://www.soloentendidos.com/random-forest-2007, 2019. Accedido 2020-06-05.

[5] W. Fisher M. van Lent and M. Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. *in Proc. 16th Conf. Innov. Appl. Artif. Intell.*, pages 900–907, 2004.

[6] Defense Advanced Research Projects Agency(DARPA). D.Gunning. Explainable artificial intelligence (xai). `http://www.darpa.mil/program/explainable-artificialintelligence`, 2018. Accessed: 2018-06-06.

[7] M. Hardt J. Kroll S. Venka-Tasubramanian S. Barocas, S. Friedler and H. Wallach. The fat-ml workshop series on fairness, accountability, and transparency in machine learning. `http://www.fatml.org/`, 2018. Accessed: 2018-06-06.

[8] FICO. Explainable machine learning challenge. `https://community.fico.com/s/explainable-machine-learning-challenge`, 2018. Accessed: 2018-06-06.

[9] J. Gehrke P. Koch M. Sturm R. Caruana, Y. Lou and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, pages 1721—1730, 2015.

[10] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813.*, pages 88–100, 2016.

[11] William R. Swartout. Xplain. A system for creating and explaining expert consulting programs. *Artificial Intelligence, 21-3.*, page 55, 1983.

[12] C Paris W Swartout and J Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert, 6-3.*, pages 58–64, 1991.

[13] Owen Rambow Jonathan DeCristofaro Tanya Korelsky Regina Barzilay, Daryl Mccullough and Benoit Lavoie. A new approach to expert system explanations. *In International Workshop on NLG*, pages 67–80, 1998.

[14] C. Lacave and F. J. D´ıez. A review of explanation methods for bayesian networks. *Knowledge Engineering Review, 17*, pages 107–127, 2002.

[15] J Konstan J Herlocker and J Riedl. Explaining collaborative [U+FB01]ltering recommendations. *In Computer Supported Cooperative Work - CSCW.*, pages 78–89, 2000.

[16] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. *In CHI EA*, pages 30–32, 2002.

[17] A. Nanopoulos P. Symeonidis and Y. Manolopoulos. Moviexplain - a recommender system with explanations. *In RecSys.*, page 20, 2009.

[18] P. Symeonidis A. Papadimitriou and Y. Manolopoulos. A generalized taxonomy of explanations styles for traditional and social recommender systems. *DataMin.Knowl.Discov.,24(3)*, pages 555–583, 2012.

[19] Mustafa Bilgic and Raymond J. Mooney. Explaining recommendations: Satisfaction vs. promotion. *In Workshop on the Next Stage of Recommender Systems Research ,San Diego, CA.*, pages 66–78, 2005.

[20] C. Likitvivatanavong E. Freuder and R. Wallace. Deriving explanations and implications for constraint satisfaction problems. *In Principles and Practice of CP.*, pages 585–589, 2001.

[21] Richard J Wallace and Eugene C Freuder. Explanations for whom. in cp01 workshop on user-interaction in constraint satisfaction. *Interaction in Constraint Satisfaction.*, pages 39–54, 2001.

[22] Jason Chalecki Joe Tullio, Anind Dey and James Fogarty. How it works: A [U+FB01]eld study of nontechnical users interacting with an intelligent system. in sigchi human factors in computing. *Systems*, pages 56–67, 2007.

[23] Brian Lim and Anind Dey. Toolkit to support intelligibility in context-aware applications. *In Ubiquitous Computing (Ubicomp)*, page 40, 2010.

[24] A.Reyes and P. deBuen. F.Elizalde, L.E.Sucar. An mdp approach for explanation generation. in explanation-aware computing work shop at aaai. *Vancouver, BC, Canada.*, pages 28–33, 2007.

[25] Pascal Poupart Omar Zia Khan and James P. Black. Minimal suf [U+FB01]cient explanations for factored markov decision processes. *In ICAPS.*, pages 43–46, 2009.

[26] and Judy Goldsmith. Thomas Dodson, Nicholas Mattei. A natural language argumentation interface for explanation generation in markov decision processes. *In Algorithmic Decision Theory.*, pages 67–69, 2011.

[27] and Padraig Cunningham. Conor Nugent, Donal Doyle. Gaining insight through case-based explanation. *J.Intell.Inf.Syst,32-3.*, pages 267–295, 2009.

[28] J Mooij J Peters P Hoyer, D Janzing and BS cholkopf. Non linear causal discovery with additive noise models. *In NIPS.*, pages 78–89, 2009.

[29] Silja Renooij Charlotte S. Vlek, Henry Prakken and Bart Verheij. A method for explaining bayesian networks for legal evidence with scenarios. *Arti[U+FB01]cial Intelligence and Law, 24-3.*, pages 285–324, 2016.

[30] H. Prakken S. Renooij S. Timmer, J. Meyer and B. Verheij. A two-phase method for extracting explanatory arguments from bayesian networks. *Int. J. Approx. Reasoning,80(C).*, pages 475––494, 2017.

[31] A Peysakhovich A Lazaridou and M Baroni. Multi-agent cooperation and the emergence of (natural)language. *In ICLR, Toulon, France.*, pages 50–58, 2017.

[32] and Dan Klein. Jacob Andreas, Anca Dragan. Translating neuralese. *In Proceedings of the Association for Computational Linguistics(ACL).*, pages 78–89, 2017.

[33] Gordon J.Pace and Michael Rosner. Explaining violation traces with finite state natural language generation models. *Springer International Publishing.*, pages 179–189, 2014.

[34] O Lemon D Gkatzia and V Rieser. Natural language generation enhances human decision making with uncertain information. *In ACL.*, pages 76–89, 2016.

[35] Deborah L McGuinness and Alexander Borgida. Explaining subsumption in description logics. *In IJCAI (1).*, pages 816–821, 1995.

[36] U Chajewska and J Y Halpern. Definig explanation in probabilistic systems. *In Uncertainty in artificial intelligence.*, pages 88–106, 1997.

[37] H.Johnson and P. Johnson. Explanation facilities and interactive systems. *In IUI.*, pages 159–166, 1993.

[38] Fahri Yetim. A framework for organizing justifications for strategic use in adaptive interaction contexts. *In ECIS.*, pages 815–825, 2008.

[39] David Corfield. Varieties of justification in machine learning. *Minds and Machines,20(2)*, pages 291–301, 2010.

[40] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). `https://ieeexplore.ieee.org/document`, 2018. Accessed: 2018-09-04.

[41] J. Gehrke P. Koch M. Sturm R. Caruana, Y. Lou and N. Elhadad. "intelligible models for healthcare: Predicting pneumonia risk and hospital30-dayreadmission,". *In Proc.21th ACMSIGKDD Int.Conf.Knowl. Discovery Data Mining.*, pages 1721–1730, 2015.

[42] K. Xu et al. "show, attend and tell: Neural image caption generation with visual attention,". *Int. Conf. Mach. Learn. (ICML).*, pages 1–10, 2015.

[43] B.Ustun and C. Rudin. "super sparse linear integer models for optimized medical scoring systems,". *Mach. Learn., vol. 102, no.3,*, pages 349–391, 2015.

[44] S. Sarkar. "accuracy and interpretability trade-offs in machine learning applied to safer gambling,". *in Proc. CEUR Workshop.*, pages 79–87, 2016.

[45] L.Breiman. "statistical modeling : The two cultures(with comments and a rejoinder by the author),". *Stat. Sci., vol. 16, no. 3In ECIS.*, page 199–231, 2001.

[46] M. van Gerven G. Ras and P. Haselager. "explanation methods in deep learning:users, values, concerns and challenges.". `https://arxiv.org/abs/1803.07517`, 2018. Accessed: 2020-01-15.

[47] A. Santoro. "a simple neural network module for relational reasoning.". `https://arxiv.org/abs/1706.01427`, 2017. Accessed: 2020-01-05.

[48] U. Paquet R. B. Palm and O. Winther. "recurrent relational networks for complex relational reasoning.". `https://arxiv.org/abs/1711.08028`, 2017. Accessed: 2020-01-05.

[49] J. Zhu Y. Dong, H. Su and B. Zhang. "improving interpretability of deep neural networks with semantic information,". *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4306–4314, 2017.

[50] J.M.Mooij D.Sontag R.Zemel andM.Welling. C.Louizos, U.Shalit. "causal effect inference with deeplatent – variable models,". *InProc.Adv. Neural Inf. Process. Syst. (NIPS)*, pages 6446–6456, 2017.

[51] O. Goudet. "learning functional causal models with generative neural networks.". `https://arxiv.org/abs/1709.05321`, 2017. Accessed: 2020-01-08.

[52] A. Rangarajan C. Yang and S. Ranka. "global model interpretation via recursive partitioning."'. `https://arxiv.org/abs/1802`, 2018. Accessed: 2020-01-04.

[53] A. Nagesh M. A. Valenzuela-Escárcega and M. Surdeanu. "lightly-supervised representation learning with global interpretability.". `https://arxiv.org/abs/1805.11545`, 2018. Accessed: 2020-02-08.

[54] J. Yosinski T. Brox A. Nguyen, A. Dosovitskiy and J. Clune. "synthesizing the preferred inputs for neurons in neural networks via deep generator networks,". *in Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pages 3387–3395, 2016.

[55] A. Courville D. Erhan and Y. Bengio. "understanding representations learned in deep architectures,". *Dept. d'Informatique Recherche Operationnelle,Univ.Montreal,Montreal,QC,Canada,Tech.Rep.1355,2010.04253*, pages 230–256, 2010.

[56] S. Harmeling M. Kawanabe K. Hansen D. Baehrens, T. Schroeter and K.-R. Müller. "how to explain individual classification decisions,". *J. Mach. Learn. Res., vol. 11, no.6*, page 1803–1831, 2010.

[57] A. Vedaldi K. Simonyan and A. Zisserman. "deep inside convolutional networks:visual is ingimage classifiation models and saliency maps.". https://arxiv.org/abs/1312.6034, 2017. Accessed: 2020-02-16.

[58] M. D. Zeiler and R. Fergus. "visualizing and understanding convolutional networks,". *in Proc. Eur. Conf. Comput. Vis. Zurich, Switzerland: Springer*, pages 818–833, 2014.

[59] A. Lapedriza A. Oliva B. Zhou, A. Khosla and O. Torralba. "learning deep features for discriminative localization,". *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2921–2929, 2016.

[60] andQ.Yan M.Sundararajan, A.Taly.)."axiomatic at tribution for deep networks.". https://arxiv.org/abs/1703.01365, 2017. Accessed: 2020-02-10.

[61] B. Kim F. Viégas D. Smilkov, N. Thorat and M. Wattenberg. "smoothgrad: Removing noise by adding noise.". https://arxiv.org/abs/1706.03825, 2017. Accessed: 2020-02-08.

[62] M. Robnik-Sikonja and I. Kononenko. "explaining classfications for individual instances,". *IEEE Trans. Knowl. Data Eng., vol. 20, no. 5.*, pages 589–600, 2008.

[63] A. Binder W. Samek G. Montavon, S. Lapuschkin and K.-R. Müller. "explaining nonlinear classfication decisions with deep taylor decomposition,". *Pattern Recognit., vol. 65*, pages 211–222, 2017.

[64] K.-R. Müller S. Bach, A. Binder and W. Samek. "controlling explanatory heatmapre solution and semantics viade composition depth,". *inProc. IEEE Int. Conf. Image Process. (ICIP).*, pages 2271–2275, 2016.

[65] S. M. Lundberg and S. I. Lee. "a unified approach to interpreting model predictions,". *in Proc. Adv. Neural Inf. Process. Syst.*, pages 4768–4777, 2017.

[66] C. Kim O. Bastani and H. Bastani. "interpretability via model extraction.". https://arxiv.org/abs/1706.09773, 2017. Accessed: 2020-03-01.

[67] P. Sattigeri J. J. Thiagarajan, B. Kailkhura and K. N. Ramamurthy. "treeview: Peeking into deep neural networks via feature-space partitioning.". https://arxiv.org/abs/1611.07429, 2016. Accessed: 2020-03-12.

[68] D. P. Green and H. L. Kern. "modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees,". *in Proc. Annu. Summer Meeting Soc. Political Methodol.*, pages 1–40, 2010.

[69] J. Leathwick J. Elith and T. Hastie. "a working guide to boosted regression trees,". *J. Animal Ecol., vol. 77, no. 4.*, pages 802–813, 2008.

[70] R. Berk and J. Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment,. *Criminol. Public Policy, vol. 12.*, pages 513–544, 2013.

[71] P. B. Brockhoff S. H. Welling, H. H. F. Refsgaard and L. H. Clemmensen. Forest floor visualizations of random forests. https://arxiv.org/abs/1605.09196, 2016. Accessed: 2020-02-25.

[72] J. Bleich A. Goldstein, A. Kapelner and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,. *J. Comput. Graph. Statist., vol. 24, no. 1.*, pages 44–65, 2015.

[73] C. Molnar G. Casalicchio and B. Bischl. Visualizing the feature importance for black box models. https://arxiv.org/abs/1804.06620, 2018. Accessed: 2020-02-08.

[74] R. König U. Johansson and I. Niklasson. "the truth is in there - rule extraction from opaque models using genetic programming,". *in Proc.FLAIRS Conf*, pages 658–663, 2004.

[75] M. H. Aung. "comparing analytical decision support models through boolean rule extraction: A case study of ovarian tumour malignancy,". *in Proc. Int. Symp. Neural Netw. Berlin, Germany: Springer*, pages 1177–1186, 2007.

[76] T. Hailesilassie. "rule extraction algorithm for deep neural networks: A review.". https://arxiv.org/abs/1610.05267, 2017. Accessed: 2020-02-23.

[77] M. van Gerven G. Ras and P. Haselager. "explanation methods in deep learning: Users, values, concerns and challenges.". https://arxiv.org/abs/1803.07517, 2018. Accessed: 2020-02-21.

[78] T. A. Etchells and P. J. G. Lisboa. "orthogonal search-based rule extraction (osre) for trained neural networks: A practical and efficient approach,". *IEEE Trans. Neural Netw., vol. 17, no. 2.*, pages 374–384,, 2006.

[79] N. Barakat and J. Diederich. "eclectic rule-extraction from support vector machines,". *Int. J. Comput. Intell., vol. 2, no. 1*, pages 59–62, 2005.

[80] D. Whiteson P. Sadowski, J. Collado and P. Baldi. Deep learning, dark know ledge, and dark matter,. *in Proc.NIPS Work shop High-Energy Phys. Mach. Learn. (PMLR), vol. 42.*, pages 81–87, 2015.

[81] O. Vinyals G. Hinton and J. Dean. "distilling the knowledge in a neural network.". https://arxiv.org/abs/1503.02531, 2015. Accessed: 2020-02-25.

[82] R. Khemani Z. Che, S. Purushotham and Y. Liu. "distilling knowledge from deep networks with applications to healthcare domain.". https://arxiv.org/abs/1512.03542, 2015. Accessed: 2020-02-23.

[83] D. H. Yi K. Xu, D. H. Park and C. Sutton. "interpreting deep classifier by visual distillation of dark knowledge.". https://arxiv.org/abs/1803.04042, 2018. Accessed: 2020-02-27.

[84] Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners guide to) convolutional neural networks for sentence classification. https://arxiv.org/abs/1510.03820, 2016. Accessed: 2020-02-22.

[85] P. Cortez and M.J. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models,. *Inf.Sci.,vol.225.*, pages 1–17, 2013.

[86] P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis,. *in Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM).*, pages 341–348, 2011.

[87] F. Doshi Velez and B. Kim. "towards a rigorous science of interpretable machine learning.". https://arxiv.org/abs/1702.08608, 2018. Accessed: 2020-03-13.

[88] C. Rudin A. Fisher and F. Dominici. "model class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective.". https://arxiv.org/abs/1801.01489, 2018. Accessed: 2020-02-20.

[89] and C.Russell S. Wachter, B. Mittelstadt. "counter factual explanations without opening the black box: Automated decisions and the gdpr.". https://arxiv.org/abs/1711.00399, 2017. Accessed: 2020-02-21.

[90] A.Khosla A.Oliva and A.Torralba D.Bau, B.Zhou. "network dissection: Quantifying interpretability of deep visual representations.". https://arxiv.org/abs/1704.05796, 2017. Accessed: 2020-02-21.

[91] B. Z. Yuan A. Bajwa-M. Specter L. H. Gilpin, D. Bau and L. Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. https://arxiv.org/abs/1806.00069, 2018. Accessed: 2020-03-24.

[92] T.Miller. Explanation in artficial intelligence: Insights from the social sciences. https://arxiv.org/abs/1706.07269, 2017. Accessed: 2020-03-08.

[93] P. Howe T. Miller and L. Sonenberg. Explainable ai: Beware of in mates running the asylum,. *in Proc. IJCAI Workshop Explainable AI (XAI).*, pages 36–42, 2017.

[94] S. Risi R. Bidarra J. Zhu, A.Liapis and G. M. Youngblood. Explainable ai for designers: A human centered perspective on mixed initiative creation,. *in Proc. IEEE Conf. Comput. Intell. Games (CIG)*, pages 458–465, 2018.

[95] M. Burnett S. Yang I. Kwan T. Kulesza, S. Stumpf and W.-K. Wong. Too much, too little, or just right? ways explanations impact end users' mental models,. *in Proc. IEEE Symp. Vis. Lang. Hum.Centric Comput. (VL/HCC).*, pages 3–10, 2013.

[96] and S.Stumpf. D.Holliday, S.Wilson. Usertrust in intelligent systems: A journey over time,. *in Proc. 21st Int. Conf. Intell. User Interfaces.*, pages 164–168, 2016.

[97] R. Fujimaki J. Wang and Y. Motohashi. Trading interpretability for accuracy: Oblique treed sparse additive models,. *in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining.*, pages 1245–1254, 2015.

[98] G. Shmueli. To explain or to predict? *Stat. Sci., vol. 25, no. 3.*, pages 289–310, 2010.

[99] T. Yarkoni and J. Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning,. *Perspect. Psychol. Sci., vol. 12, no. 6.*, pages 1100–1122, 2017.

[100] Gabriel G. Erion Scott M. Lundberg and Su-In Lee. Consistent individualized feature attribution for tree ensemble. *University of Washington*, pages 1–8, 2019.

[101] S. Singh M. T. Ribeiro and C. Guestrin. "why should i trust you?': Explaining the predictions of any classifier,". *in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pages 1135–1144, 2016.

[102] A. Rinaldo R. J. Tibshirani J. Lei, M. G'Sell and L. Wasserman. "distribution-free predictive inference for regression,". http://www.stat.cmu.edu/~ryantibs/papers/conformal.pdf, 2018. Accessed: 2020-02-23.

[103] G. Montavon F. Klauschen K.-R. Müller S. Bach, A. Binder and W. Samek. "on pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,". *PLoS ONE, vol. 10, no. 7*, page e0130140, 2015.

[104] S. Mohseni and E. D. Ragan. "a human grounded evaluation benchmark for local explanations of machine learning.". https://arxiv.org/abs/1801.05075, 2018. Accessed: 2020-02-07.

[105] C. Mues J. Vanthienen J. Huysmans, K. Dejaeger and B. Baesens. "an empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models,". *Decis. Support Syst., vol. 51. no. 1*, pages 141–154, 2011.

[106] A. Backhaus and U. Seiffert. "quantitative measurements of model interpretability for the analysis of spectral data,". *in Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM).*, pages 18–25, 2013.

APPENDIX A

Impact of the project

textit Artificial intelligence has had an impact on society since its inception in the past decades, now we know that almost everything goes through an artificial intelligence filter from the way they suggest us to order food, or the way they show us personal articles in social networks, or in companies to measure their growth, or in health, in short, in everything we have in life we can see that artificial intelligence is present, but human nature is just as complex that is always wondering the why of the situations that is why in this topic the curiosity of the human being also plays a fundamental role, knowing why an artificial intelligence model that for the common user is transparent, has made a decision is a basis that the XAI exists.

A.1 Social Impact

As mentioned above, artificial intelligence models help us help the environment where we handle a large amount of data (BIG DATA) to handle it more quickly, but that is not all that this idea is that it can lead us to create problems if these data is not treated in the best way, respecting the laws and the rights of the human being, there are many examples that are seen on television or on social networks or in the press in which artificial intelligence models have become very serious For example, they identify children as terrorists since we used a lexicon or expressions that these types of people on the web these boys made reference to video games and keywords, in addition to health we also saw it when the models said that an asthmatic patient they are less likely to die than a healthy patient if for some reason they had pneumonia, this is because the model analyzes that one human being is medicated and the other no, therefore, the one with the most drugs has in theory more to survive. Therefore, there are many examples that, although it is true, changing the way we live if we are not careful, we face great and serious problems. So if we know what is being done in a model, we can say that our social system will be able to evolve and make a leap into the future with a good projection of development in which machines and technology really facilitate the life of the human being.

A.2 Economic Impact

Likewise, the economic impact that artificial intelligence has and will have is something that has no precedent, since the industry itself moves through this technology, companies develop or buy artificial intelligence models in order to have more profits by being more efficient and proactive to when generating products to meet the needs of its users as well as knowing the growth projection of their organizations. banks also benefit from these technologies since they automate the entire credit line system and thus remove that burden from themselves, in general, the impact that AI has on our economic situation will depend on how strong the model is and how well Explained how he makes the decisions that will generally be transparent but will be there if needed

A.3 Environmental Impact

Technologies linked to computing as such are already an environmental problem, and this will be the case at the XAI since for a model to interpret, adjust and predict something, a computer is needed that is always ready to collaborate with it, and how is to imagine in

the companies or in the places where these predictions are stronger or cost more to achieve, more resources will be needed to help us do this, applying techniques to explain these models, what is achieved is that the effort of a machine increase, and with it the impact of the ecological footprint that is already in itself, the challenge would be to find alternatives so that this footprint decreases because the tendency would be to increase and that would help in one aspect of our life but it would affect us in other.

A.4 Ethical and Professional Implications

The ethics in explainable artificial intelligence is a fact that has much to be studied yet, since there are still gaps in the laws that ordinary people ignore, such as with data from people, proven data that could be taken and treated in such a way that they are not very attached to the law, and it is that it is a conflict that we will always be able to see and demonstrate in our society, the more progress is made with technology the more careful they must be with the laws, now The people who dedicate themselves to this type of work in which artificial intelligence is a tool for decision-making, must be people who have very high morals and ethics. Explainable artificial intelligence, although it is true, helps that these models are not as transparent to users and in this way it is possible to know what is happening and how a model is making a decision that could affect a user. .

APPENDIX B

Cost of the System

B.1 Physical Resources

First, to develop and test the system, you need a computer. Also, this computer must have enough memory and functionality to run the model used to make predictions, and then use technology to explain that it is done in a graphical form that is difficult to process or difficult to obtain. Therefore, we believe that this computing unit should provide the following resources:

- **Hard Disk**: 500 GB
- **RAM**: 16 GB
- **CPU**: Intel i7 processor, 2.80 GHz

We estimate that the cost of a machine with these capacities would be around € 800 and as a cap we would have around € 1,200.

B.2 Human Resources

In order to carry out this work, a human effort was needed that could be remunerated since the models and techniques that were used to explain them were developed. We will make a brief analysis of the salary that would be required to pay someone to do what has been done in this work.

The cost of the work hours dedicated to the development of the project is estimated at approximately 460 hours. Considering that the academic plan estimates that the cost of the project is approximately 12 ECTS credits (European Credit Transfer System), this prediction is calculated. Since each credit involves 25 to 30 hours of work, the final investment time must be 340 hours; We added 130 hours specifically to document this project. Assuming that a computer engineer salary is approximately € 2000 (gross) and there are 23 business days in a month, the development of this project will take up to three months (considering an eight-hour business day) or related costs of € 7,000. With labor.