Machine Learning Principles

Introduction

In the medical world, understanding and researching the diseases that affect the human body is very important, because it is through this that new medicines or procedures are created in order to reduce the risk of death and/or sequels caused by diseases. With the development of new technologies, research in this field is becoming increasingly advanced since the data that are generated and made available today can be analyzed by advanced tools that manage to find, as is the case of the BigML platform that facilitates and automates the processes of data analysis, which are often not noticed by human beings. As a result, this work will have the main purpose of predicting which patients are more likely to have a heart problem in the future, which is why supervised learning algorithms will be used.

1 - BigML

For the development of this work, a Machine Learning as a Service (MlaaS) platform known as BigML was used. According to information provided by the company's website, this is a cloud-based platform that aims to facilitate the development of Machine Learning activities, such as: Classification, Regression, Forecasting, Clustering... (BigML, na).

2 - Dataset

For the development of this work, a dataset offered free of charge by the Kaggle website, called "Heart Failure Prediction", which can be accessed through the link – https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction, was used. According to the author, this dataset was taken from a research work carried out by the authors Davide Chicco, Giuseppe Jurman in the year 2020 and contains the medical records of 299 Pakistani patients who have heart failure.

2 - Data Pre-Processing

Data pre-processing is considered a crucial step for the application of machine learning algorithms, since it is through this step that data is prepared to be analyzed, in order to bring greater efficiency, thus facilitating decision making. and making predictions using the data. (Javatpoint, n.d).

2.1 - Data pre-processing in BigML

2.1.1 – Importing the Data

In order to use the selected dataset in the development of this work, it was necessary, at first, to create the data source within the platform, which in this case refers to the import of data into the platform.

The platform offers several possibilities for importing and/or creating, which facilitates the use of different types and sources of data, however, I ended up using the "upload of a local file" option, as it is demonstrated in the Figure 1.

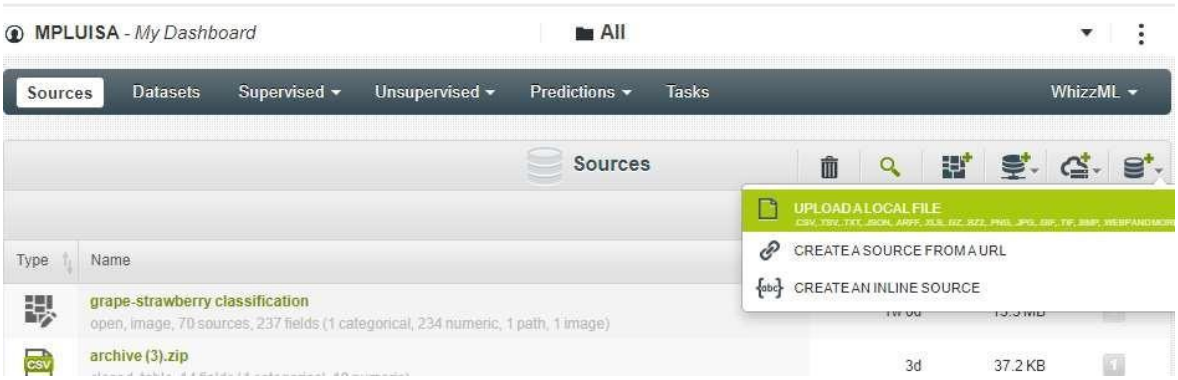

Figure 1 – Importing the Dataset

2.1.2 – Understanding the Data

In order to carry out an analysis and/or understand the best algorithm to be used, it is necessary to understand the data to be used, in the platform information such as data type, amount of data, missing values and errors are easily found in the Dataset field, as shown in the Figure 2, presented below: (BigML Education, 2017)

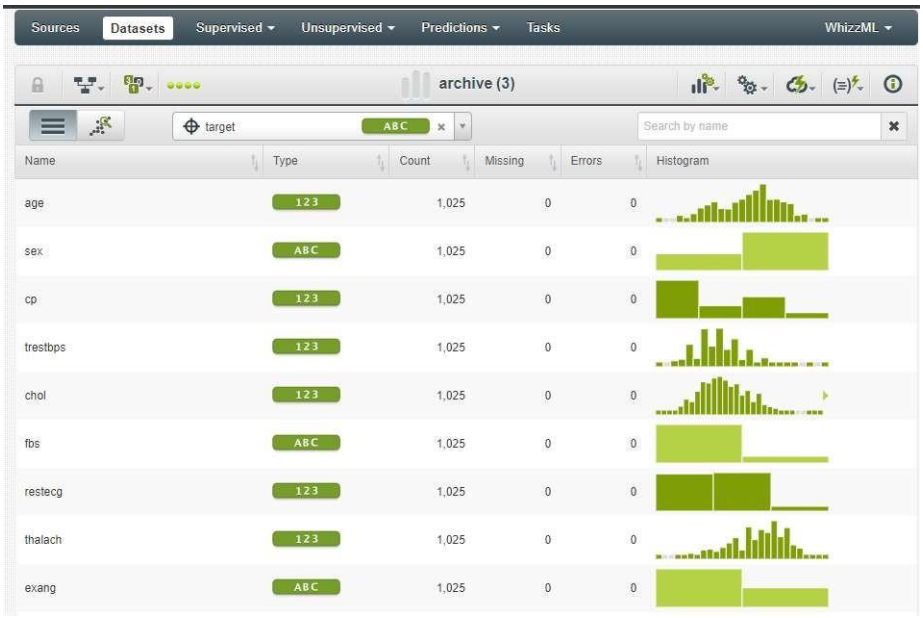

Figure 2 – Dataset

Analyzing the figure above, it can be seen that this dataset is formed by numerical and categorical data, in addition, it is noted that there is no missing value and because of this, the data does not need to be treated, since this has already been previously performed.

Another very important tool presented by BigML education (2017) refers to the ease of analyzing the correlation of two or more variables within the dataset, as is the example of the cholesterol x age variables shown in the figure below.

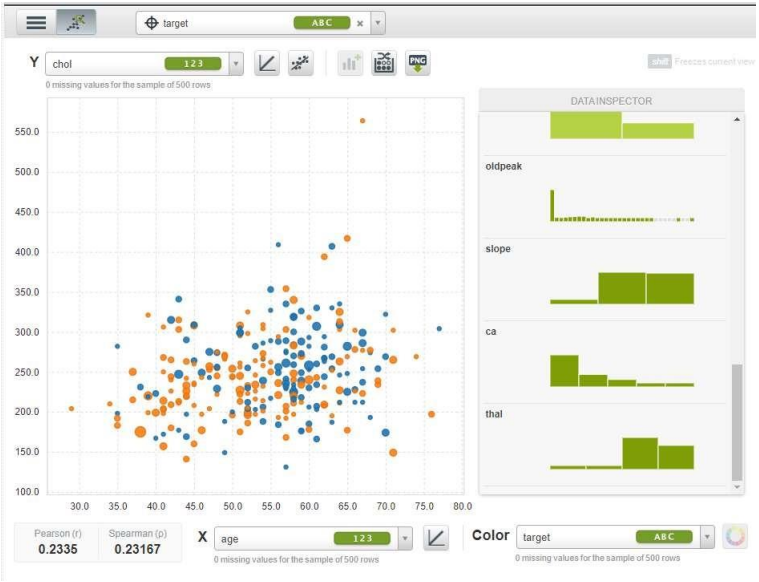

Figure 3 - Cholesterol x Age

In view of the above, it is noted that this platform visually presents extremely important information for the application of machine learning models, thus facilitating the choice of the analysis method that would best fit the data presented, information that will be addressed in the next topic.

3) Processing the Data

Before the data are processed, it is of paramount importance to understand the analysis method that will be used and the algorithms available for this. For this work, we chose to use the supervised analysis method, since we want to understand the probability of a person having heart problems, taking into account the variables presented.

According to Brownlee, 2019, Supervised learning is one of the data analysis methods that uses algorithms that learn through per-defined or known results, using a labeled data set, thus having a reference of what is right and what is wrong.

Batta and Mahesh (2018) also adds that supervised learning algorithms need external assistance, since they use historical data that are divided to be trained and tested in order to reach the desired levels of accuracy, and are then often used to make predictions (regression) or classification.

After understanding the type of method to be used, it began to be applied within the BigML platform, for which it was necessary to use the guide or step-by-step provided by BigML Education (2017), in the related videos. to Models 1 and Models 2.

The use of this tutorial facilitated the understanding of the platform, so at first I generated a model for the entire data set, with that a decision tree was created automatically, using the last variable of the dataset as target (since I did not I selected the target variable).

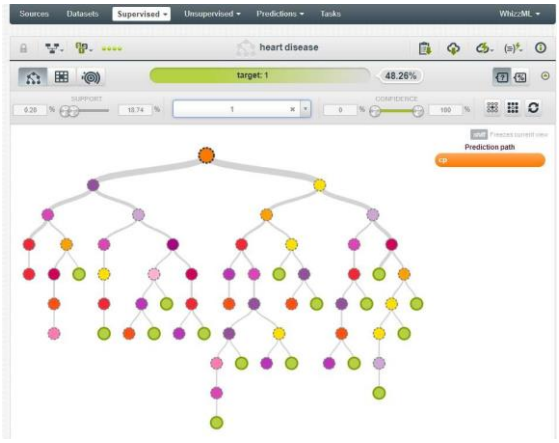

Figure 4 – Model aplication

Analyzing the Figure 4, it can be noticed that I selected the target variable "1" that represents people who have a predisposition to have heart problems and through this Decision Tree, you can analyze the ramifications, also presenting the level of confidence about the results presented by the algorithm at the end of each node.

According to the information provided by Tonye (2021, August 25), it is through this information that we are able to analyze the probability that the information provided by the model is correct and satisfies the information expected by the user.

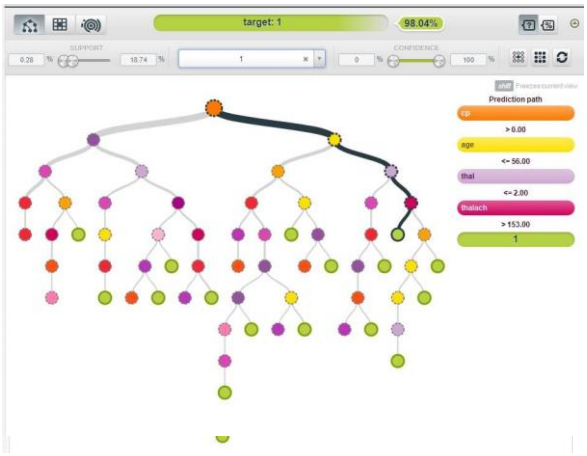

Figure 5 – Nodes Analyses

Figure 6 – Nodes Analyses

In figures 5 and 6 above, you can see two examples where the percentage of confidence changes according to the selected node, in addition, you can analyze the variables in which they are being taken into account so that this percentage is reached, as well as the values that these variables have within this model.

On the other hand, I had a little difficulty understanding the other graphics, such as the PDP and the Sunburst, even with the examples given by the platform, I was unsuccessful in replicating the tutorial within the chosen dataset.

4 - Performance

To carry out the performance analysis of the used dataset, at first it is necessary to divide this dataset into Training and Test. For this work, I chose to use the 80:20 ratio – 80% training and 20% test – with the samples selected randomly.

For this method to be applied within the platform, I used the guide provided by BigML Education (2017), where they clearly explain how to evaluate the algorithm used.

The separation of the data set within the platform is carried out automatically, through the "Training/split" link, there you can select the percentage you want, both for training and for the test and after that you just need to select the "create training" button /test".
After this procedure you can evaluate your model by generating the confusion matrix, this is also another easy resource to be found, where with just one click the matrix is displayed.

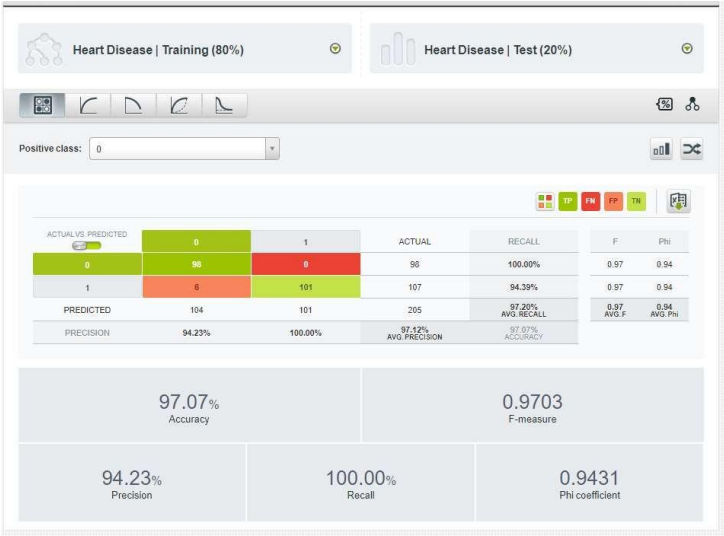

Figure 7 – Confusion Matrix

Analyzing the information presented in Figure 7, it can be noted that in general this model performed well, since the Accuracy is 97.07%, which is relatively high for a machine learning algorithm. In the table, we can see that the numbers of false positives and false negatives are relatively low compared to the sample size, which confirms the performance of this algorithm.

I also took the opportunity to test how to make predictions using the dataset, however I created a prediction using the training data set, for that I used the guide provided by BigML education (2017) where they explain how to make predictions.

Figure 8 – Prediction

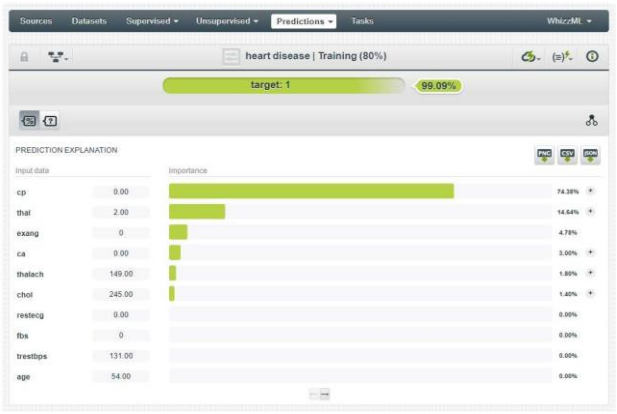

Figure 9 – Variable Importance for Prediction

Analyzing figures 8 and 9, it can be noted that the probability of a person who has the same conditions as the patients presented in the dataset, of having some heart disease is very high, in addition, according to the data presented by the model, the variable "cp" or chest pain type has a greater importance within the prediction.

5 – Useful Features in BigML

We are living in an age of data, where the information provided by them brings different insights that can be used by companies in order to generate a competitive advantage, or even for a more accurate decision making.

As a result of the above, technology companies are developing and providing more and more applications, platforms that facilitate the use of tools that will bring benefits to companies, as is the case of Machine Learning as a Service.

According to information presented by HP development Company (2020), the use of MlaaS brings several benefits to companies, reducing costs of using Machine Learning tools, greater ease of use and application of algorithms, bringing a new wave of innovation within companies.

Gunthrie (2019) also adds that the use of MlaaS helps companies to have access to several Machine Learning functionalities, such as: Speech recognition, sentiment analysis, chatbot enhancement, image and video analysis, and classification and regression. In addition, the author emphasizes that the use of cloud-based platforms helps companies to reduce costs, time and, mainly, to reduce the risks that the use of machine learning tools can bring.

On the other hand, Onose (2021) warns that companies that work with secret data, that need major customizations or work with complex algorithms, should not use MlaaS platforms, since it is cloudbase, the security level is reduced.

Conclusion

Through the development of this work, one can get a sense of the use of one of the MlaaS tools that are available on the market, in this case, BigML. The use of this platform was very user-friendly, providing visual ease in the analysis and processing of data. In addition, this tool has self-explanatory videos and texts, with practical examples of use, which makes it even easier to use.

Complementing the above, it can also be seen that to use this type of service you do not need to have programming knowledge, focused on Machine Learning, but you need to have an analytical profile to be able to analyze the data and draw conclusions and/or take decisions based on results.

On the other hand, in this work basic activities were carried out, but it can be noted that there are several ways to further deepen the use of data in order to bring even more accurate information in a reduced time.

It is concluded, then, that the use of MlaaS platforms tends to gain much more space within the market, since it has a low implementation cost compared to the on-site structures that are necessary to process Machine Learning algorithms.

REFERENCES

Batta, M., & Mahesh, B. (2020). Machine learning algorithms: A review. International Journal of Advanced Intelligence Paradigms, 14(4), 237-252

BigML education (2017) Dataset [Video file]. YouTube. Retrieved June 30, 2023, from https://www.youtube.com/watch?v=FEI0PP5KUAI&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=3

BigML education (2017) Model 1 [Video file]. YouTube. Retrieved June 30, 2023, from https://www.youtube.com/watch?v=hnt7z24wvxs&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=5

BigML education (2017) Model 2 [Video file]. YouTube. Retrieved June 30, 2023, from https://www.youtube.com/watch?v=y4AkvCAa8Ik&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=6

BigML education (2017) Prediction [Video file]. YouTube. Retrieved June 30, 2023, from https://www.youtube.com/watch?v=g2mCLfnKt8E&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=16

BigML education (2017) Evaluations [Video file]. YouTube. Retrieved June 30, 2023, from https://www.youtube.com/watch?v=cPErxYP9CmQ&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=11

BigML (na) About Big ML https://bigml.com/about/#:~:text=BigML%20is%20a%20consumable%2C%20programmable,Discovery%2C%20and%20Topic%20Modeling%20tasks.

Brownlee, J. (2019, December 5). A tour of ML algorithms [Blog post]. Machine Learning Mastery. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Darmala, Rishi (2020) Heart Disiase Prediction. https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction

Guthrie, G. (2019, 3 December). Machine learning as a service (MLaaS) is the next trend no one is talking about. Data Driven Investor. https://medium.datadriveninvestor.com/machinelearning-as-a-service-mlaas-is-the-next-trend-no-one-is-talking-about-e100973121c1.

HP Development Company, L. P. (2020, July). AI for all machine learning as a service. HP Marketing Document Library. https://h20195.www2.hp.com/v2/getpdf.aspx/4AA7-7926ENW.pdf

JavaTpoint. (n.d.). Data Preprocessing in Machine Learning. Retrieved June 30, 2023, from https://www.javatpoint.com/data-preprocessing-machine-learning

Onose, E. (2021, 16 February). Machine learning as a service: What it is, when to use it and what are the best tools out there. Neptune Blog. https://neptune.ai/blog/machine-learning-asa-service-what-it-is-when-to-use-it-and-what-are-the-best-tools-out-there.

Tonye, Guy (2021, August 25) Machine Learning Confidence Scores — All You Need to Know as a Conversation Designer. https://medium.com/voice-tech-global/machine-learning-confidence-scores-all-you-need-to-know-as-a-conversation-designer-8babd39caae7