Università degli Studi dell' Aquila

Machine Learning For Data Analysis Of Dual Trauma Project

Department of Engineering and Computer Sciences and Mathematics

Master's Degree in Computer Science, SEAS

Candidato

Tiziano Santilli

Matricola 261502

Relatore

Prof. Antinisca Di Marco

Anno Accademico 2020/2021

Machine Learning For Data Analysis Of Dual Trauma Project
Tesi di Laurea. Università degli Studi dell' Aquila

Questa tesi è stata composta con LaTeX e la classe uaqthesis.

Email dell'autore: tizianosantilli@gmail.com

Contents

Chapter 1

Introduction

In 2009 in L'Aquila, an earthquake shocked the whole world and caused 309 deaths, 1,600 injured, 80,000 displaced persons, and problems for many people residing in central Italy. Unfortunately, the repercussions of this catastrophe continue still today, in the form of psychiatric issues that affect people and especially children who experienced the earthquake firsthand.

Many studies in literature concern the high incidence of psychiatric pathology on populations who personally experienced natural catastrophic events such as an earthquake [10, 18, 7, 12, 2]. They conduct a questionnaire and compare the results obtained with the healthy population (i.e., the experimental control). In such studies, the analysis is traditionally performed without considering innovative prediction techniques such as machine learning that can be very useful to identify risky situations in advance.

As for Machine Learning applied to psychiatry, we found only two interesting works [11, 9]. In [11], the authors try to predict the Based Suicide Ideation on a sample of military population by using *i)* an index called $BSRS-5$ that provides a score representing the probability of suicide and *ii)* Machine learning, in particular applying Neural networks [6], Decision Tree [16], Random Forest [5] and SVM [13].

To this aim, the authors determined the $BSRS-5$ using 5 factors (i.e., Anxiety, Depression, Hostility, Interpersonal sensitivity, and Insomnia) and compared it to the results of machine learning techniques. In the end, the study shows that machine learning prediction is more accurate than $BSRS-5$ index and among the considered techniques, Neural Networks over perform among the others. Similarly to [11], in this thesis we will apply machine learning techniques to predict the self-harm inventory. Differently from [11], we apply a greater number of Machine Learning techniques on a very different sample-based, i.e. students attending high schools in L'Aquila. Furthermore, we use feature importance approaches to explain the results obtained by machine learning techniques.

Paper [9] uses, instead, analysis a sample coming from prisoners by using machine learning to predict suicidal behaviors without setting out questions regarding suicidal ideation in the questionnaire. This study is interesting in that it focuses beyond prediction, but also on the identification of the minimal set of questions that lead to an accurate prediction of suicidal behavior. Such kind of analysis allows to reduce the number of questionnaire's questions/answers to be considered in the machine learning approaches by eliminating those not useful to the prediction and, hence, reducing the computational resources (such as execution time and memory space) needed for the analysis. The article ends with an accurate forecast even having reduced the questionnaire questions to 29 from a starting base of 51. It is interesting to note that none of the 29 selected questions are directly linked to suicidal ideation. Again, differently from [9], we use a different and wider sample. Furthermore the study in [9] focuses much more on the reduction aspects of the dataset features in order to keep the forecast accuracy high, than to a pure prediction.

To the best of our knowledge, there are no scientific studies targeting people who have faced natural disasters. For this reason, **The Dual Trauma Project** can be considered pioneering as, in addition to standard data analysis, aims to predict

self-harm and suicidal risk to be able to identify those risks in advance and be able to intervene promptly.

To this aim, we implement in this thesis machine learning based approaches to predict 7 Self Harm Inventory (SHI-i) indicators that are considered of extreme importance in the psychiatric panorama, as, in addition to being explanatory indices for self-harm and suicidal ideation, they are often linked to borderline and dissociating syndromes. To predict as much accurate results as possible, we clean and pre-process the data to make them ready for maximum accuracyDealing with data of people with psychiatric risk, the biggest challenge was to minimize the number of false negatives (those people considered sane by the algorithm but who present problems). Moreover, being neural networks natively black boxes methods, another challenge is to be able to render the methods of explainable prediction.

To achieve these prediction goals, we implement, configure the best combination of parameters, and finally apply 13 different machine learning methods (described in Chapter 2. We decide to run all such methods to improve results' accuracy trying to leverage on each method's strengths to reduce their weaknesses. In the implementation step, we use *sk-learn* library that allows an easy and rigorous implementation and configuration of all selected methods.In addition to the prediction of SHI-i indicators, the project also aims to provide an explanation of the obtained results by applying Feature Selection techniques. Feature selection techniques allow us to reason on those psychiatric parameters that can likely lead the onset of suicidal thoughts and self-harm.

We started the research project (i.e., The Dual Trauma project) in January 2020. At the beginning of the project, the idea was to find a method able to predict some indicators that would help prevent harmful behaviors in the examined sample so that to help people with simile profile of the analysed sample, in the future.

Figure 1.1 sketches the process we follow in the study. After a careful analysis

of the conducted survey with the psychiatric research team, we agreed that the best variables to predict were those of the SHI (Self Harm Inventory) indicators represented by 7 questions concerning self-harm:

- SHI1. Taking an excessive amount of drugs, alcohol or drug therapy?
- SHI2. Did you cause any cuts / wounds / burns / scratches?
- SHI3. Did you hit yourself / deliberately hit your head against something?
- SHI4. Did you stop your wounds from healing?
- SHI5. Have you made your medical condition worse?
- SHI6. Have you engaged in sexually promiscuous behavior?
- SHI7. You have entered into a relationship in which you felt rejected or humiliated sexually or psychologically?

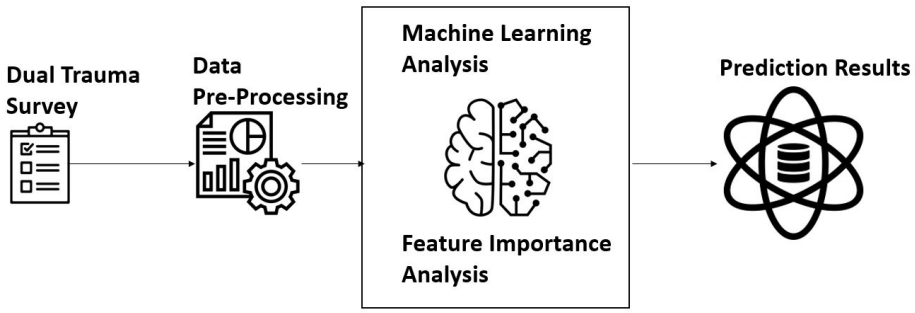

Figure 1.1. Project Workflow

As second step, we process the survey data to ensure that they are in a format suitable for machine learning techniques. Then we implemented and run the selected machine learning techniques and the Feature importance analysis. As soon as we obtained the results, we realized that not only the accuracy of the predictions was of fundamental importance for a psychiatric project but also a sort of explanation

to be able to understand in detail the reason for these results. To be able to do this, we have also implemented a random forest which, together with providing the accuracy of the method's predictions, also provides a way to be able to classify the features that most of all influence the predictions of the SHI. These results of the feature importance are reported in chapter 5. As soon as we had these results, it was a natural logical consequence to apply the Decision Tree method in addition to the Machine Learning Random Forest method, and we started comparing the results with each other. At this point, after studying other papers on similar topics, we also decided to implement other Machine Learning methods to have more concrete results, as each method has its weaknesses and strengths. In the end, also thanks to the sk-learn library, we implemented all the most famous classification methods present in the literature that have now become a de-facto standard. The results of the application of these methods are analysed in order to provide insights to the psychiatric operator (in Chapter 5).

Roadmap

This thesis project is structured as follows:

- Chapter 2: In this chapter, we provide the description of the 13 used Machine Learning methods and of the used Feature Importance approach.
- Chapter 3: In this chapter, we provide the description of the context of the study, a description of the questionnaire administered to the students and the structure of the data were used for the prediction and Feature Importance techniques. The chapter, then, proceeds with the presentation of the structure and of the problems related to raw dataset, and the methods used to solve these problems have also been explained.
- Chapter 4: In this chapter, we show all the setting and the configurations for

the Machine Learning techniques, we also show all the indicators that we use as results of the project.

- Chapter 5: In this chapter, we show the results of both the predictions on the analyzed features, and the Feature Importance results.

- Chapter 6: In this chapter we we have analyzed the implications of the results obtained from the analysis. We also report on possible future scenarios that have as a starting point the work done in this thesis.

Chapter 2

Background

This chapter explains the Machine Learning and Feature Importance techniques used. For the choice of techniques, we based both on the current state of the art and on the study of similar scientific articles, but also on the psychiatric context of our study. The following sections show all 13 Machine Learning techniques used plus the explanation of Feature Importance.

The techniques used were:

- Neural Networks
- Decision Tree
- Random Forest
- Support Vector Machine with linear kernel
- Support Vector Machine with polynomial kernel
- Support Vector Machine with RBF kernel
- Lasso Linear Model
- Linear Bayesian Model

- Gradient Boosting
- Nearest Neighbors Classification
- Gaussian Process Classifier
- AdaBoost
- Stochastic Gradient Descent

As regards the neural network, the Pytorch library was used, while the scikit-learn library was used for all the other analyzes.

2.1 Neural Network

Neural networks [6], also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm.

Think of each individual node as its own linear regression model, composed of input data, weights, a bias (or threshold), and an output. Once an input layer is determined, weights are assigned. These weights help determine the importance of any given variable, with larger ones contributing more significantly to the output compared to other inputs. All inputs are then multiplied by their respective weights and then summed. Afterward, the output is passed through an activation function, which determines the output. If that output exceeds a given threshold, it "fires" (or activates) the node, passing data to the next layer in the network. This results in the output of one node becoming in the input of the next node. This process of passing data from one layer to the next layer defines this neural network as a feedforward network.

2.2 Decision Tree

A decision tree [16] is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules.

Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification

And Regression Tree (CART) is general term for this.

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the features, x1, x2, x3 etc., that are used for that task.

2.3 Random Forest

A random forest [5] is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like scikit-learn).

2.4 Support Vector Machine

The Support Vector Machine (SVM)[13] is a supervised machine learning algorithm that can be used for both classification and regression purposes. It is popular in applications such as natural language processing, speech and image recognition, and computer vision.

The SVM algorithm obtains maximum effectiveness in binary classification

problems. Although it is used for multiclass classification problems, in this post we will mainly focus on binary classification, mainly looking at how such an algorithm works. The Support Vector Machine aims to identify the hyperplane that best divides the support vectors into classes. To do this, perform the following steps:

- Look for a linearly separable hyperplane or a decision limit that separates the values of one class from the other. If there is more than one, look for the one that has the highest margin with the support vectors, to improve the accuracy of the model.

- If such a hyperplane does not exist, SVM uses a non-linear mapping to transform the training data into a higher dimension (if we are in two dimensions, it will evaluate the data in 3 dimensions). In this way, the data of two classes can always be separated by a hyperplane, which will be chosen for splitting the data.

SVM algorithms use a set of mathematical functions defined as a kernel. Its purpose is to take the data as input and transform it into the required form if it is not possible to determine a linearly separable hyperplane, as is the case in most cases.

We analyzed the Dual Trauma dataset using three different kernels: linear, polynomial and RBF (radial basis function).

2.5 Linear Models

In the analysis of the Dual Trauma data we also used two types of linear models[19]: Lasso and Bayesian.

- Lasso regression [15] is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The

lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

- Bayesian linear regression [3] is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

2.6 Gradient Boosting and AdaBoost

Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm.The AdaBoost Algorithm [17] begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2. We then compute

the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models. Gradient Boosting [8] trains many models in a gradual, additive and sequential manner. The major difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners (eg. decision trees). While the AdaBoost model identifies the shortcomings by using high weight data points, gradient boosting performs the same by using gradients in the loss function (y=ax+b+e , e needs a special mention as it is the error term). The loss function is a measure indicating how good are model's coefficients are at fitting the underlying data. A logical understanding of loss function would depend on what we are trying to optimise. For example, if we are trying to predict the sales prices by using a regression, then the loss function would be based off the error between true and predicted house prices. Similarly, if our goal is to classify credit defaults, then the loss function would be a measure of how good our predictive model is at classifying bad loans. One of the biggest motivations of using gradient boosting is that it allows one to optimise a user specified cost function, instead of a loss function that usually offers less control and does not essentially correspond with real world applications.

2.7 K-Nearest Neighbor Classification

K-Nearest Neighbor classifier [14] is one of the introductory supervised classifiers, which every data science learner should be aware of. This algorithm was first used for a pattern classification task which was first used by Fix & Hodges in 1951. To be similar the name was given as KNN classifier. KNN aims for pattern recognition

tasks.

K-Nearest Neighbor also known as KNN is a supervised learning algorithm that can be used for regression as well as classification problems. Generally, it is used for classification problems in machine learning.

KNN works on a principle assuming every data point falling in near to each other is falling in the same class. In other words, it classifies a new data point based on similarity.

KNN algorithms decide a number k which is the nearest Neighbor to that data point that is to be classified. If the value of k is 5 it will look for 5 nearest Neighbors to that data point.

The simple version of the K-nearest neighbour classifier algorithms is to predict the target label by finding the nearest neighbour class. The closest class to the point which is to be classified is calculated using Euclidean distance.

In our case we found out that a K-value equals to 3 gave us the best accuracy.

2.8 Gaussian Process Classifier

A Gaussian Process [1] is the generalization of the distribution over functions with finite domains, in the infinite domain.

This is achieved by sampling mean functions and covariance functions that return the mean to be used to generate the Gaussian distribution to sample the first element and also the covariance function between every pair of variables.

The interesting thing is that, while any function is a valid mean function, not every function is a valid covariance.

For classification problems, one simple way to adapt gaussian processes is to choose a 0-1 loss (i.e. punish false positives and false negatives equally), normalize the target into a 0-1 interval (e.g. using the logistic function) so that it can be

viewed as a probability and choosing some threshold value for actual classification.

2.9 Stochastic Gradient Descent

Stochastic Gradient Descent [4] is the extension of Gradient Descent.

Any Machine Learning/ Deep Learning function works on the same objective function f(x) to reduce the error and generalize when a new data comes in.

To overcome the challenges in Gradient Descent we are taking a small set of samples, specifically on each step of the algorithm, we can sample a minibatch drawn uniformly from the training set. The minibatch size is typically chosen to be a relatively small number of examples; it could be from one to few hundred.

Using the examples from the minibatch. The SGD algorithm then follows the expected gradient downhill:

The Gradient Descent has often been regarded as slow or unreliable, it was not feasible to deal with non-convex optimization problems. Now with Stochastic Gradient Descent, machine learning algorithms work very well when trained, though it reaches the local minimum in the reasonable amount of time.

A crucial parameter for SGD is the learning rate, it is necessary to decrease the learning rate over time, so we now denote the learning rate at iteration k as Ek.

2.10 Feature Importance

When we talk about feature importance we are referring to a set of techniques that aim to identify features in a dataset that have a greater importance than the others for the prediction of the final result. The use of these techniques has practical implications that are very important both in terms of the quality of predictions and in terms of performance. Specifically, a feature importance can bring the following advantages:

- Reduction of the size of the dataset keeping only the features that are actually useful for the final prediction
- Better interpretation of the results thanks to the knowledge of the most important features that led to the final choice
- Understanding the relationships between the features that are deemed important by the algorithm and the features that you want to predict.

The most common and most used approach for feature importance, namely the feature importance given by random forest.

Random Forest Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

Chapter 3

Dual Trauma Survey And Data Pre-Processing

In this chapter we find a detailed analysis of the context of the Dual Trauma questionnaire, we also find an accurate description of the sample and the methods of administering the questionnaire. In the survey section, we find a description of psychiatric indicators and indices that form the questions present in the questionnaire. A full description of each question administered is present along with its explanation in the acronyms section. Finally, we find in the Data Pre-Processing section an accurate description of the problems of the raw dataset and the methods used to solve them.

3.1 Context

The target population comprises adult students enrolled in the last year of higher education institutions (IIS) in L'Aquila and Avezzano. We draw the sample from the institutes mentioned above. According to the latest ISTAT data, 2642 people aged 18 and 2613 aged 17 lived in L'Aquila's province in2018. In L' Aquila's province and

Avezzano, there are 10 IIS, 4 in L'Aquila's territory and 6 in the district of Avezzano. We surveyed the IIS, the first agreements with school managers, a sampling of 50% and 75% of the population under examination appears likely for each IIS. It was not possible to estimate the number of minors present in the classes that will be recruited in the study. At the time of data collection, minors excluded from the study are counted. This type of sampling introduces a selection bias by excluding subjects who dropped out of school after age 16. The sampling of 52 classes covers 50% of the population in that high school age group.The sampling of all high schools leads to the minimization of selection bias and increases the generalizability of the results. No subjects under the age of 18 areincluded in the study. From each of the 10 IIS present in the school districts of L'Aquila and Avezzano, after contacting and involving the school management, we have drawn between 75% and 100% of the last classes for each IIS, trying to respect the heterogeneity of different school addresses.

We administered the questionnaires, for logistical reasons, in paper format to all the subjects recruited who will agree to participate in the study and who will have signed the informed consent. The logistics of the administration has been agreed upon from time to time with the school management. The UNIVAQ personnel supervised the administration of the questionnaires. We designed the questionnaire so that it does not take more than 60 minutes to complete. The questionnaire consists of the first part of the information sheet and relatively informed consent, specific for volunteers of the school population and the information relating to the processing of personal data (according to Legislative Decree no. 196of 30 June 2003). After reading the information sheet, the subject gives his consent to participate in the study, and then he is asked to fill in the informed consent form. All questionnaires are pre-filled with a unique alphanumeric code. Attached was provided with a form showing the corresponding alphanumeric code of the questionnaire, on which the subject entered his / her details. In the same form, the subject indicated whether or not he wishes to

be contacted if the tests he will undergo are positive to receive psychological support. The identification cards were collected and stored separately from the questionnaires in a closed and sealed envelope. The information sheets are kept at the DISCAB premises in a special locked locker under the responsibility of the PI. This study involves the examination of different domains of functioning, penological and clinical. Each domain is investigated using standardized tools recognized in international literature. The questionnaire is divided into areas as follows:

3.2 Survey Description

For the study of addictions, the use of alcohol and cannabis, pathological gambling and internet addiction are examined.

- The problematic use of alcohol is investigated through AUDIT-C, a reduced version of 3 items of the original 10-item questionnaire.
- The use of cannabis will be investigated through CAST-6.
- Gambling is identified via the Brief Adolescent Gambling Screen (BAGS-3) (6), abridged version of the Canadian Adolescent Gambling Inventory (CAGI-10). This tool is not validated in Italian
- A 13-item multiple choice questionnaire investigating past family stress experiences (Risky Family Questionnaire, RFQ).
- A questionnaire for measuring exposure to traumatic events (International Trauma Exposure Measure), added with 3 specific items for the 2009 earthquake.
- Bullying and cyberbullying.

- A questionnaire for the evaluation of dissociative symptoms (Dissociative Experiences Scale).
- A 33-item response questionnaire on personal and interpersonal resilience (Resilience Scale for Adult, RSA).
- A 29-item questionnaire for the study of attachment styles (Attachment style questionnaire).
- Brief TEMPS: self-administered questionnaire that measures affective temperaments.
- A self-assessment questionnaire of self-harm to suicidal and non-suicidal behaviors (Self-harm inventory).
- An Emotional Regulation Strategies Assessment Questionnaire (DERS).
- An 18-item questionnaire to assess post-traumatic symptoms (International Trauma Questionnaire, ITQ).
- A 25-item questionnaire for evaluating anxious and depressive symptoms (HSCL25).
- A 16-item questionnaire for assessing pre-psychotic symptoms (PQ16).

In this way, from an IT point of view, we obtained a Dataset consisting of 1062 rows corresponding to the individual questionnaires for 268 columns corresponding to the questions of the complete questionnaire.

3.3 Data Pre-Processing

When we started studying the Dataset, we immediately realized that we had to resolve two problems as soon as possible.

The first was to make numeric all the values that were saved as a string. Otherwise, it would have been impossible to create a neural network based on this data. To solve this problem, we used a technique called “one-hot encoding”.

One-hot encoding is a method to quantify categorical data. In short, this method produces a vector with a length equal to the number of categories in the data set. If a data point belongs to the ith category, then components of this vector are assigned the value 0 except for the ith component, which is assigned a value of 1. In this way, one can keep track of the categories in a numerically meaningful way.

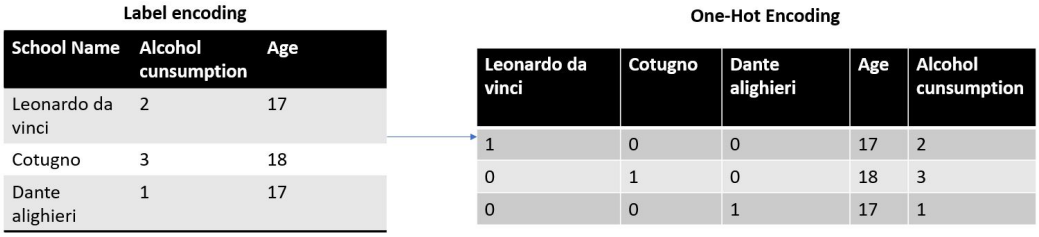

Figure 3.1. One-Hot Encoder At Work

The second problem to be solved, to make the dataset better for the analysis techniques, was to balance the available data. This problem arose because the data showed a large discrepancy between people who actually had self-harm problems (i.e. SHI-i th parameter equal to one) and people who hadn’t had these problems. The ratio between subjects who did not have self-harm problems compared to the others is 9 to 1. If we had left the dataset so unbalanced, any analysis technique that had predicted all negative values would still have had an accuracy of 90%. But in this case the data would have been completely unusable, as the purpose of this research is precisely to identify the true positives, that is, to reduce false negatives. Two main techniques were used to overcome this problem:

- Rebalancing of positives.
- Assigning a higher weight to the SHI prediction class with a positive value.

For the rebalancing of the positives We used the Python Numpy and Pandas libraries, We selected the subjects with SHI equal to 1 and copied them, thus creating a new dataset that is more balanced, with a number of positives equal to 20For the attribution of a greater predictive weight to SHI greater than zero we used the Crossentropyloss () function available on Pytorch which as an input parameter allows you to define various weights to be given to the classes. The weight to be attributed was chosen by testing various values, in the end a weight of 4 to 1 for the positives was the best for minimizing false negatives.

Chapter 4

Experimental Settings

In this chapter, we show the configurations tested and then approved for the various Machine Learning techniques. Most of the configurations have been tested on the neural network, as this technique allows for very high parameters customization. As regards the other analysis techniques used, the configurations were limited to the definition of the parameters of the objective function and to the search for the trade-off parameter that could give the greatest accuracy. In the second part of the chapter, we find a description of all the indices we used to give the results of the research project. These indices were chosen together with the team of psychiatrists as they represent the most used metrics and have the most value in the psychiatric field.

4.1 Machine Learning Configurations

After processing the data we created a first neural network with the library *sk-learn*, but after various tests and configurations, we realized that the neural network creation methods contained in *sk-learn* were not sufficiently powerful and customizable. For example, the *sk-learn* library does not allow the definition of different functions for forward and backward between the various layers during

the creation of the neural network. So we moved to *PyTorch* library which allows configuration of any possible parameter for the neural network.

In order to create a neural network that provides the best performance on our dataset, we tried various parameter configurations for it. In table 4.1 we report all the tested combinations and we highlighted the final setting that has maximized the performance of the Neural Network.

Hiden Layers Configuration	Functions	Epochs	Learning Rate
(210)(64)(20)(5)	Relu	50	0.1
(210)(150)(100)(70)(50)	Linear	100	0.01
(210)(128)(64)(32)	Sigmoid	250	0.001
(210)(128)(5)	Convolutional	500	0.0001
(210)(64)(5)	SoftMax	1000	

Figure 4.1. Neural Net Configuration

Furthermore, we used a 10-fold Cross-Validation to make the quality of the predictions statistically relevant. 10-fold Cross-Validation is a technique that consists in changing the dataset section at each run of the machine learning method, in this way we can be sure that each entry of the dataset was both predictor and predicted variable. The final result is given by the average of the single runs. We used the training set at 90% of the data and the test set at 10% of the data.

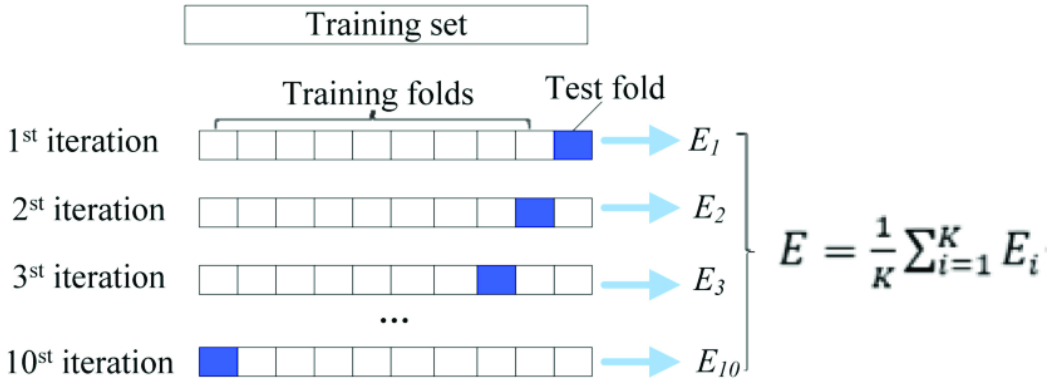

Figure 4.2. How the cross validation work

4.2 Analysis And Quality Indicators

For the analysis of the predictions of the indicators we used various metrics that succeed in the best possible way results that are relevant from a psychiatric point of view, for this reason we have focused on 7 indicators:

- Accuracy: With Accuracy we indicate the probability of successful prediction of the neural network, this value is given by the number of successfully predicted samples divided by the total number of predicted samples.
- False Positives: With False positives we indicate all those subjects that the analysis techniques classify as self-injurious and with suicidal ideation (therefore positive), but who in reality do not present any problems.
- False Negatives: With False Negatives we indicate all those subjects that the analysis techniques classify as not at risk (therefore negative), but who actually exhibit self-injurious behaviors and suicidal ideation.
- True Positives: With True Positives we indicate those subjects who are correctly classified as at risk both by the analysis techniques and in reality.

- True Negatives:: With True Negatives we indicate those subjects who are correctly classified as not at risk both by the analysis techniques and in reality.

- Sensitivity: The term sensitivity, in statistics, more precisely in the field of epidemiology, indicates the intrinsic capacity of a screening test to identify sick subjects in a reference population. It is defined in equation 4.1 as the ratio of true positive (V_+) and the entire patient population ($V_+ + F_-$).

$$S = \frac{V_+}{V_+ + F_-} \tag{4.1}$$

- Specificity: The term specificity, in medicine, indicates the ability of a test to give a normal ("negative") result in healthy subjects. The specificity of a test is the probability of a negative result in surely healthy subjects. It is defined in equation 4.2 as the ratio between the true negatives(V_-) and the total of healthy ones($V_- + F_+$).

$$S = \frac{V_-}{V_- + F_+} \tag{4.2}$$

Chapter 5

Results Analysis

In this chapter, we report the final results of the research project together with the explanation and discussion about them.

In the first section, we discuss the results of Feature Importance Analysis, where we list the features (questions of the survey) that most impacted the prediction. This analysis is fundamental from the psychiatric point of view because, although it is important to have a system that predicts with the utmost accuracy, it is also essential to be able to provide explanations of the events that led to that prediction. In this way, psychiatrists can try to preventively intervene on those factors that are most at risk.

In the second section, we discuss all the results obtained from Machine Learning techniques. The results show the percentage accuracy of the predictions and also the results obtained for the psychiatric indicators described in Chapter 4.

5.1 Features Importance Analysis

We conducted the Features Importance Analysis for each SHI indicators selected. In conducting the analysis we take into account the 30 most important features that alone contribute to about 80% of the final prediction result.

Below we can find the charts with the Feature Importance for each SHI, at the subscript of the document we find an acronyms section where all the variables are explained.

SHI1. Taking an excessive amount of drugs, alcohol or drug therapy?

Regarding the SHI-1 variable, we can see how the most relevant factors for predicting the excessive use of drugs, medications or alcohol are those that derive from the thought of taking drugs, the lack of organization of one's time and propensity to play. of gambling. A more detailed view is presented in the following chart.

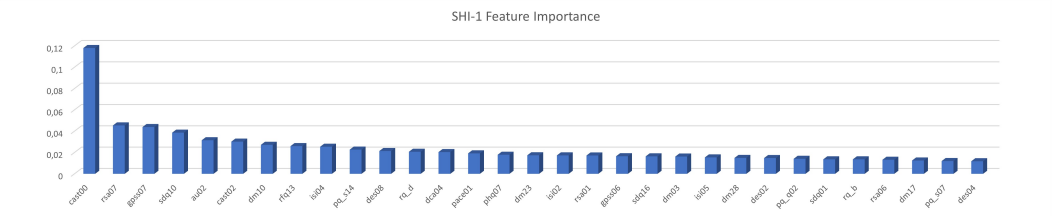

Figure 5.1. SHI-1 Features Importance

SHI2. Did you cause any cuts / wounds / burns / scratches?

Regarding to the SHI-2 variable, we can see how the most relevant factors for self-harm are the thought of harming oneself and loss of control, as well as the fact of not having a best friend. A more detailed view is presented in the following chart.

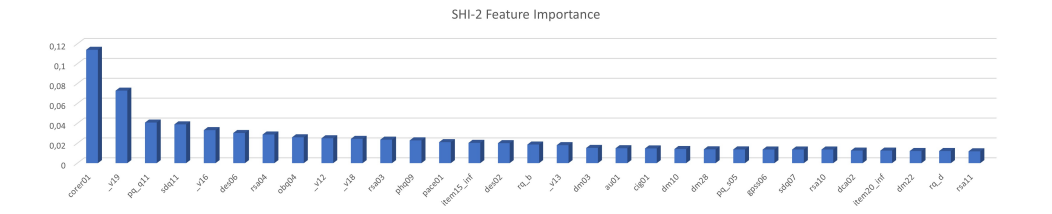

Figure 5.2. SHI-2 Features Importance

SHI3. Did you hit yourself / deliberately hit your head against something?

Regarding the SHI-3 variable, we can see how the most relevant factors to predict the fact of hitting oneself are for the most part demographic factors, i.e. related to the family environment or to one's body, such as the weight of the subject, the fact of having met a psychologist, but also the fact of not sharing food. It is interesting to note that although SHI 2-3 are similar, they actually have triggering causes that are almost completely different. A more detailed view is presented in the following chart.

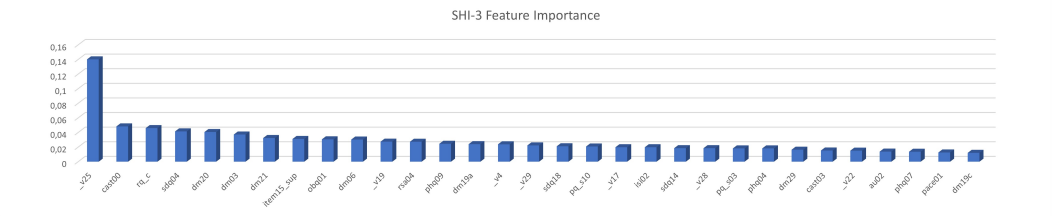

Figure 5.3. SHI-3 Features Importance

SHI4. Did you stop your wounds from healing?

Regarding the SHI-4 variable, we can see how the most relevant factors to predict the situation in which the subject prevented the wounds from healing were the fact of having voices in the head, having suffered a theft and having thought about the game. of gambling. A more detailed view is presented in the following chart.

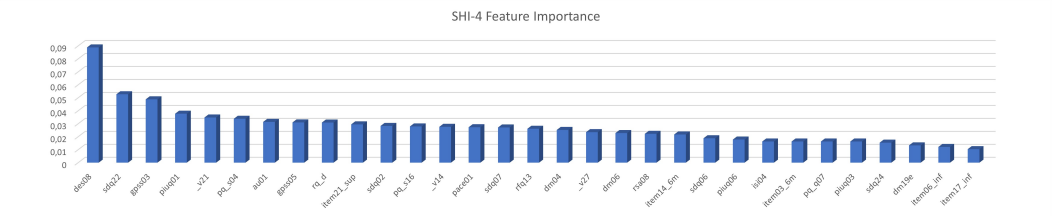

Figure 5.4. SHI-4 Features Importance

SHI5. Have you made your medical condition worse?

Regarding the SHI-5 variable, we can see how the most relevant factors for predicting the fact of having worsened one's medical condition concern the sphere of food, such as having a poor or excessive appetite and binge eating, moreover also the abandonment is an important fact for prediction. A more detailed view is presented in the following chart.

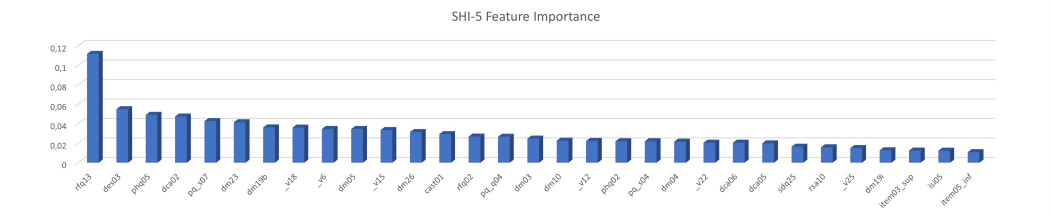

Figure 5.5. SHI-5 Features Importance

SHI6. Have you engaged in sexually promiscuous behavior?

Regarding to the SHI-6 variable, we can see how the most relevant factors for predicting the fact of having sexually promiscuous behaviors are the thought of abusing drugs, seeing imaginary objects and even relapsing into gambling. Surprisingly, even if in a more marginal way, even the fact of owning a car or not affects this parameter. A more detailed view is presented in the following chart.

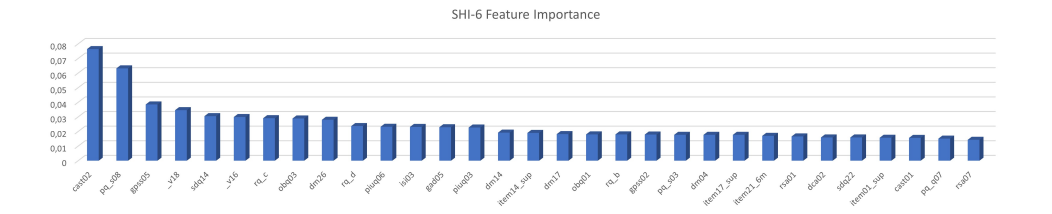

Figure 5.6. SHI-6 Features Importance

SHI7. You have entered into a relationship in which you felt rejected or humiliated sexually or psychologically?

Regarding the SHI-7 variable, we can see how the most relevant factors to predict the fact of having spontaneously undertaken a harmful relationship are having offensive nicknames, the fact of not completing one's goals and also the diagnosis of a fatal disease to one or more relatives. A more detailed view is presented in the following chart.

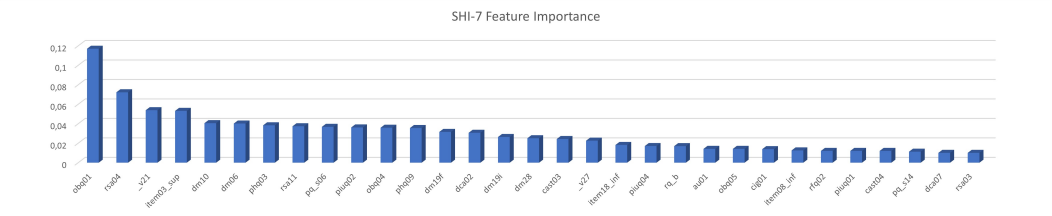

Figure 5.7. SHI-7 Features Importance

SHI-TOT. At least one of the conditions listed above.

Regarding the SHI-TOT variable we can see how the most relevant factors for the prediction of at least one of the previous SHI variables is the fact of having thoughts about drug and alcohol abuse, gambling, and the perception that the world is not real. A more detailed view is presented in the following chart.

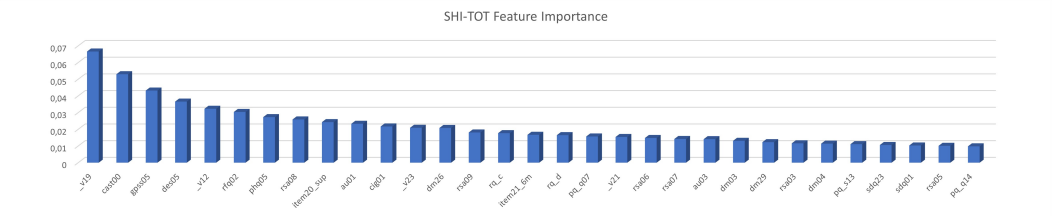

Figure 5.8. SHI-TOT Features Importance

5.2 Machine Learning Results

In this section, we find the results of the predictions in graphic format, in total 16 structures have been created that represent the 8 predicted variables, both in original format and in format with doubled positives. The doubled positives, as well as the technique of increasing the weight of the objective function, has the purpose of being able to minimize false negatives, as the original dataset has a very unbalanced percentage of positives. The minimization of false positives is a fundamental goal from a psychiatric point of view since being the false positive a subject who is ill, but who is recognized as healthy by the system, we could have a situation in which this subject cannot be helped and therefore could incur self-injurious and/or suicidal behaviors that are not treated. On the contrary, a high number of false positives is negligible, as having a healthy subject who is treated as a potentially sick person is not a problem, since a psychiatric session has no contraindications. For completeness of the results, we reported o the true positives and true negatives. The other two indicators, specificity, and sensitivity are very useful from the physiological point of view. It is rare to find a test that has high specificity and at the same time high sensitivity, for this reason during this research project we have tried to find the best trade-off between the two indices. Furthermore, the values described in the graphs are an average given by the 10 runs of the system due to the 10-fold cross-validation. The results indicators are explained in details in Chapter 4.

SHI1. Taking an excessive amount of drugs, alcohol or drug therapy?

The percentage of positives present for this variable in the dataset is 14%, which becomes 28% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, even if the Neural Network is to be preferred as it has a lower number of false negatives. Linear methods, on the other

hand, have a number practically equal to zero of true positives, this particularity shows us how in such unbalanced datasets these methods tend to always and only give the majority value, to obtain a high accuracy percentage. As regards the dataset with doubled positives, we note comparable performances.

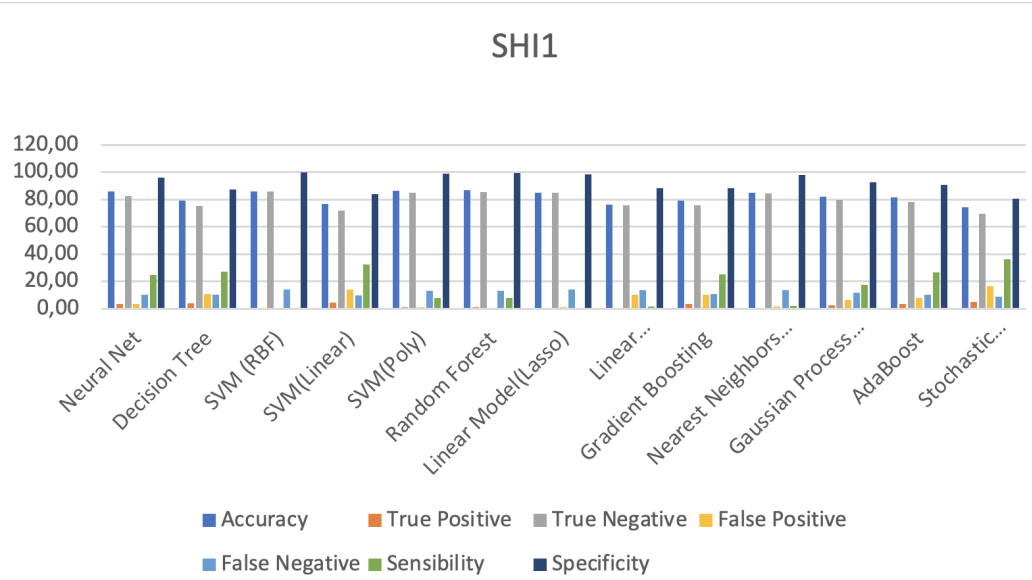

Figure 5.9. Predictions For SHI1 Feature

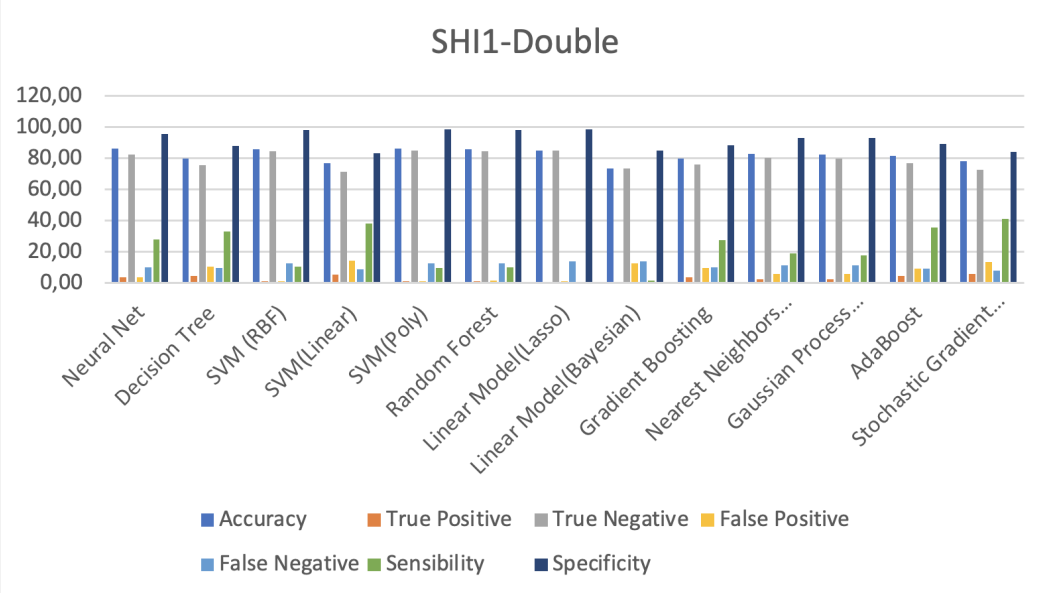

Figure 5.10. Predictions For SHI1-Double Feature

SHI2. Did you cause any cuts / wounds / burns / scratches?

The percentage of positives present for this variable in the dataset is 13%, which becomes 26% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, even if the Neural Network is to be preferred as it has a lower number of false negatives. Linear methods, on the other hand, have a number practically equal to zero of true positives. As regards the dataset with doubled positives, we note slightly better performances, especially on the false-negative indicator.

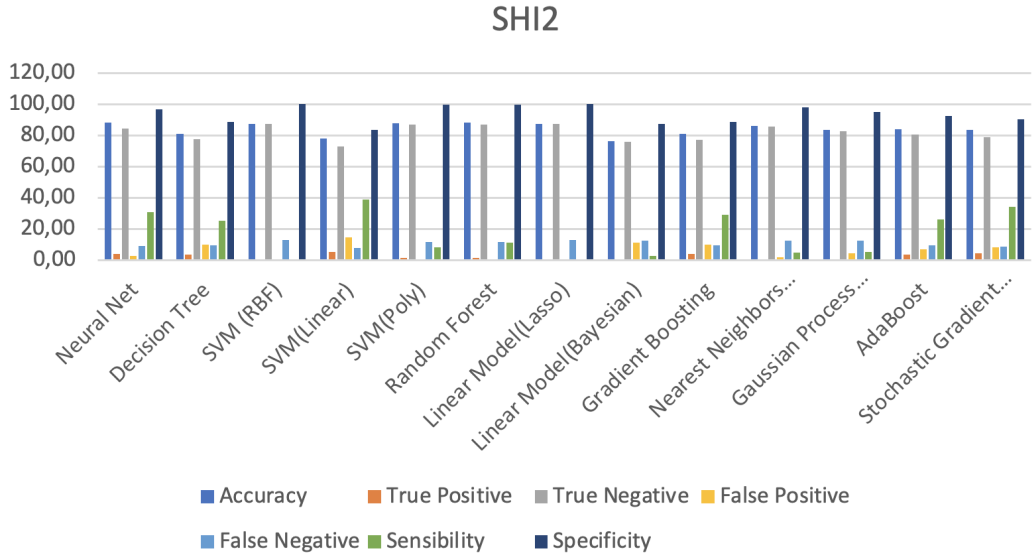

Figure 5.11. Predictions For SHI2 Feature

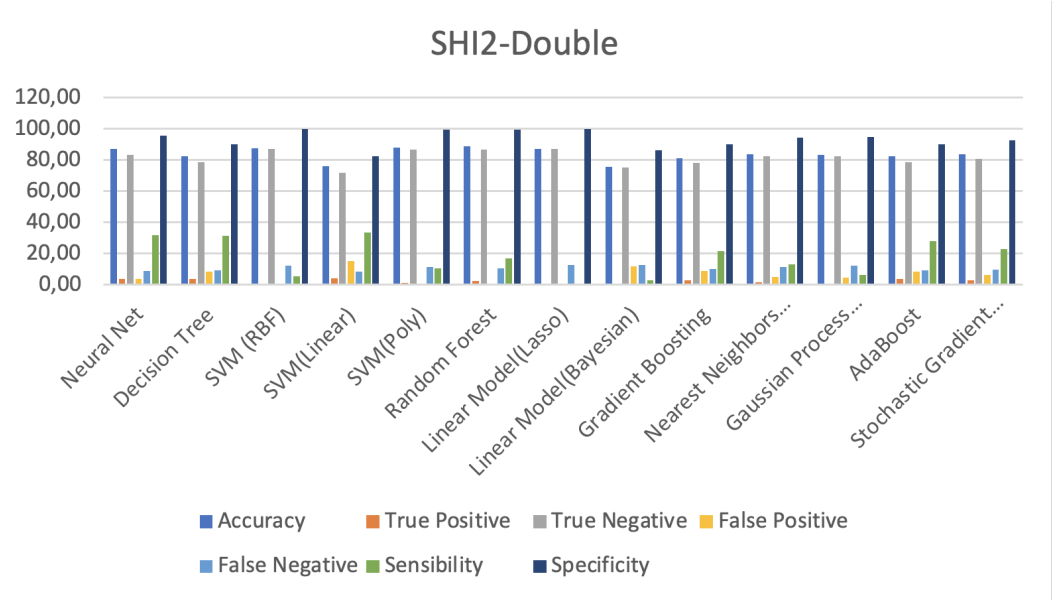

Figure 5.12. Predictions For SHI2-Double Feature

SHI3. Did you hit yourself / deliberately hit your head against something?

The percentage of positives present for this variable in the dataset is 9%, which becomes 18% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the SVM and Random Forest methods. Linear methods, on the other hand, have a practically zero number of true positives. As for the dataset with doubled positives, we notice a clear improvement in performance for the Decision Tree.

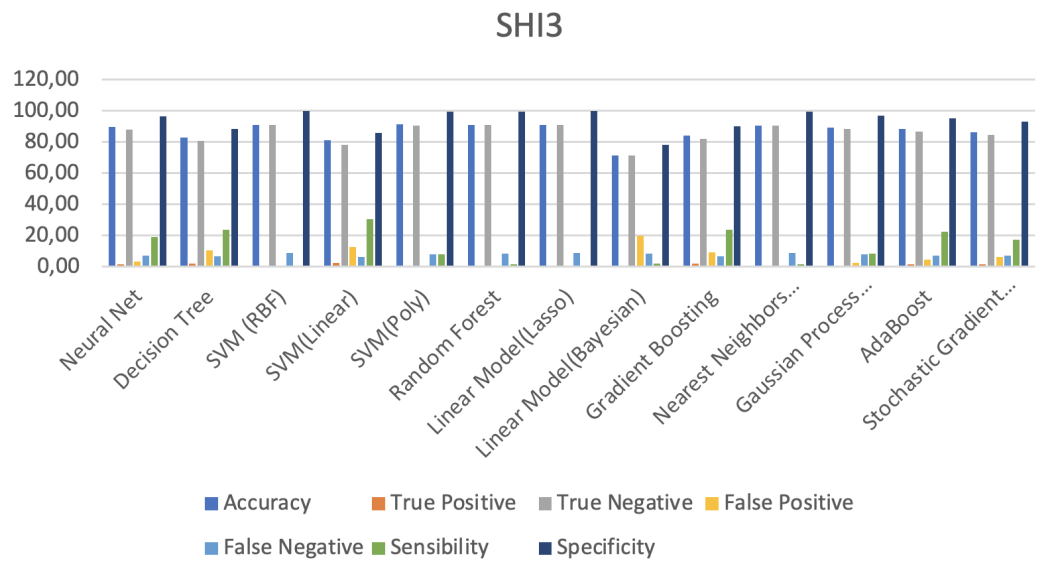

Figure 5.13. Predictions For SHI3 Feature

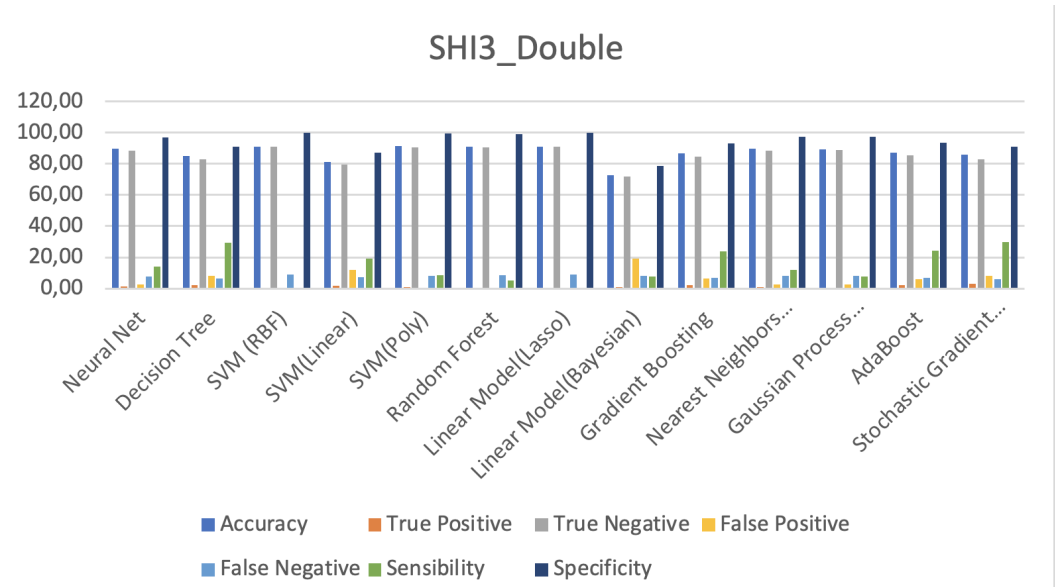

Figure 5.14. Predictions For SHI3-Double Feature

SHI4. Did you stop your wounds from healing? The percentage of positives present for this variable in the dataset is 7%, which becomes 14% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, in this case the Neural Network is to be preferred as it has a smaller number of false negatives. Regarding the dataset with doubled positives, we do not notice a change in performance.

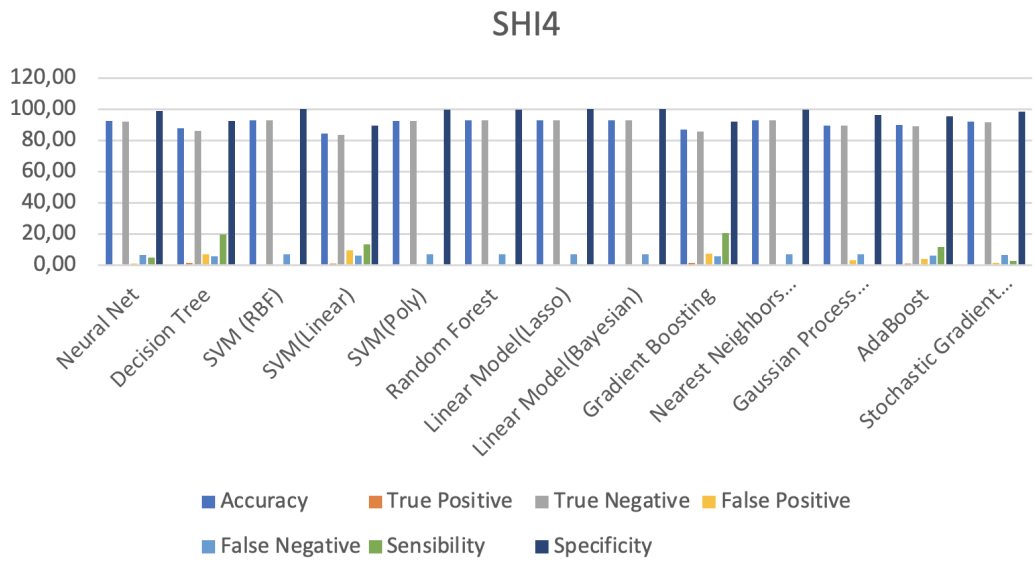

Figure 5.15. Predictions For SHI4 Feature

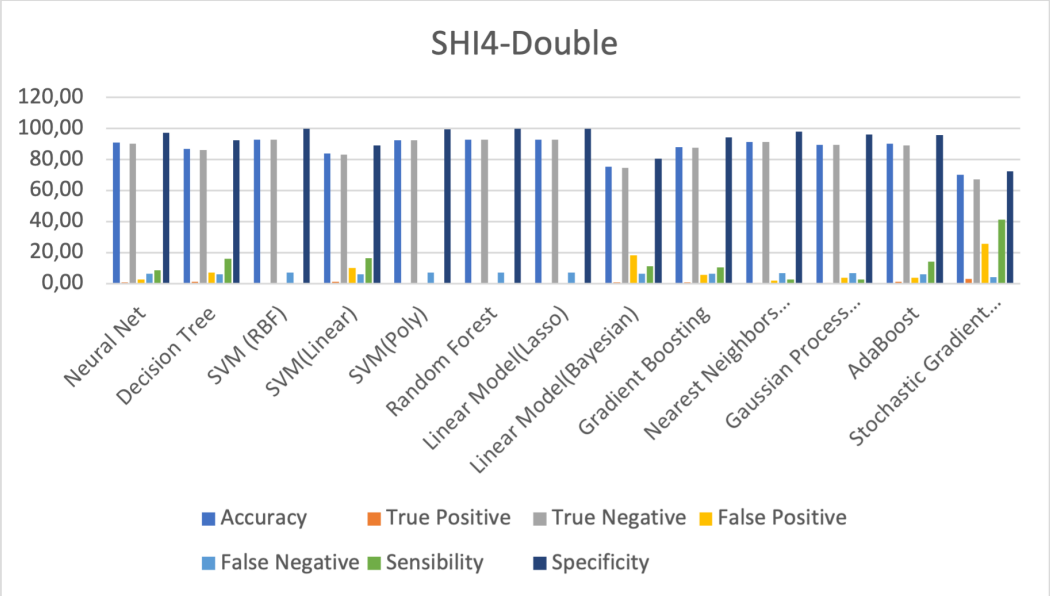

Figure 5.16. Predictions For SHI4-Double Feature

SHI5. Have you made your medical condition worse?

The percentage of positives present for this variable in the dataset is 5%, which becomes 10% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, in this case, the Random Forest is to be preferred as it has a smaller number of false negatives. As regards the dataset with doubled positives, we note an improvement in performance in the Boosting methods

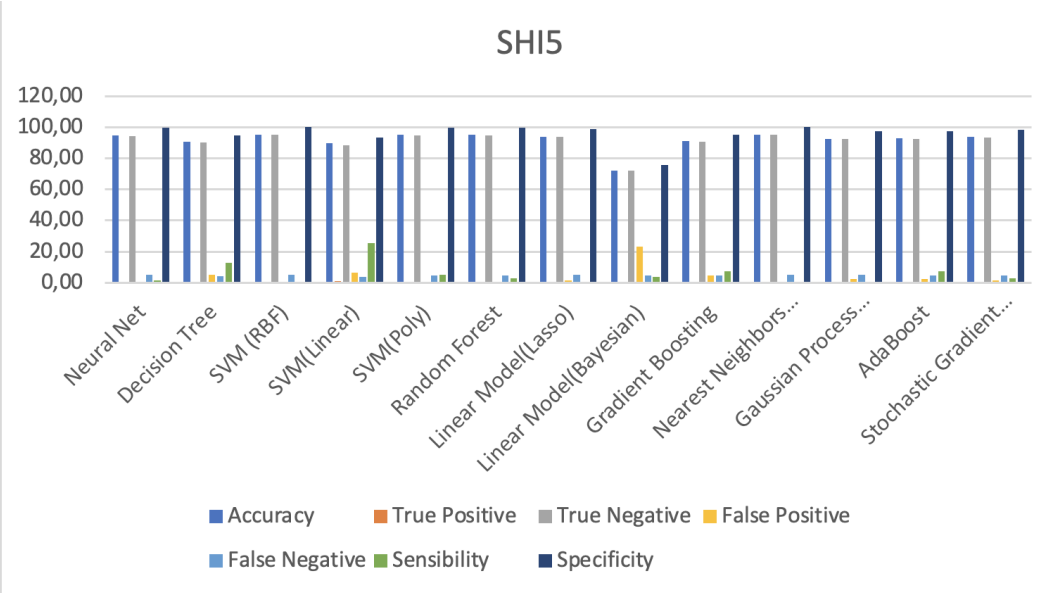

Figure 5.17. Predictions For SHI5 Feature

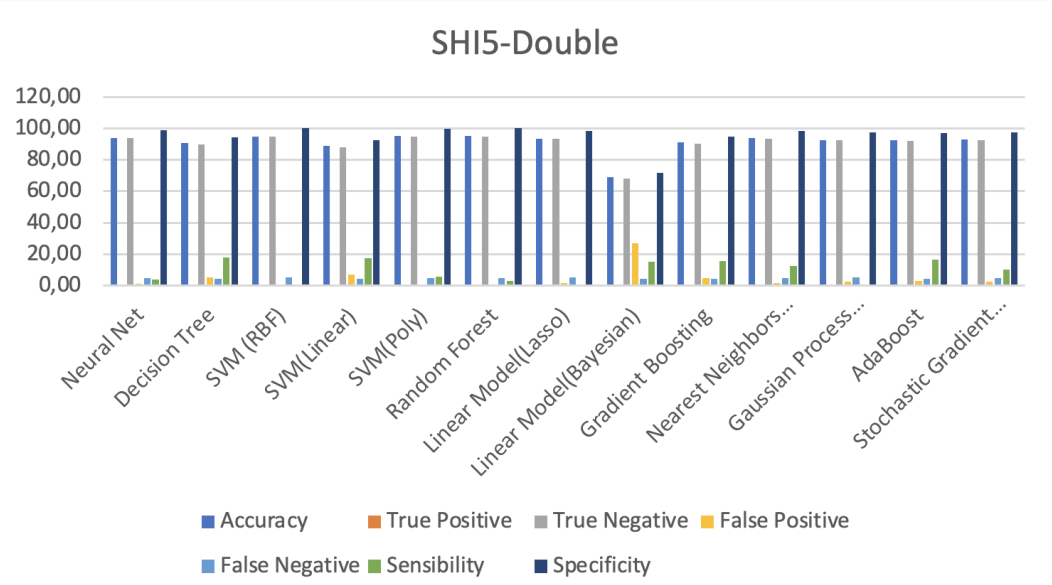

Figure 5.18. Predictions For SHI5-Double Feature

SHI6. Have you engaged in sexually promiscuous behavior?

The percentage of positives present for this variable in the dataset is 7.5%, which becomes 15% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, in this case, the Neural Network is to be preferred as it has a smaller number of false negatives. As regards the dataset with doubled positives, we note a generalized improvement in performance.

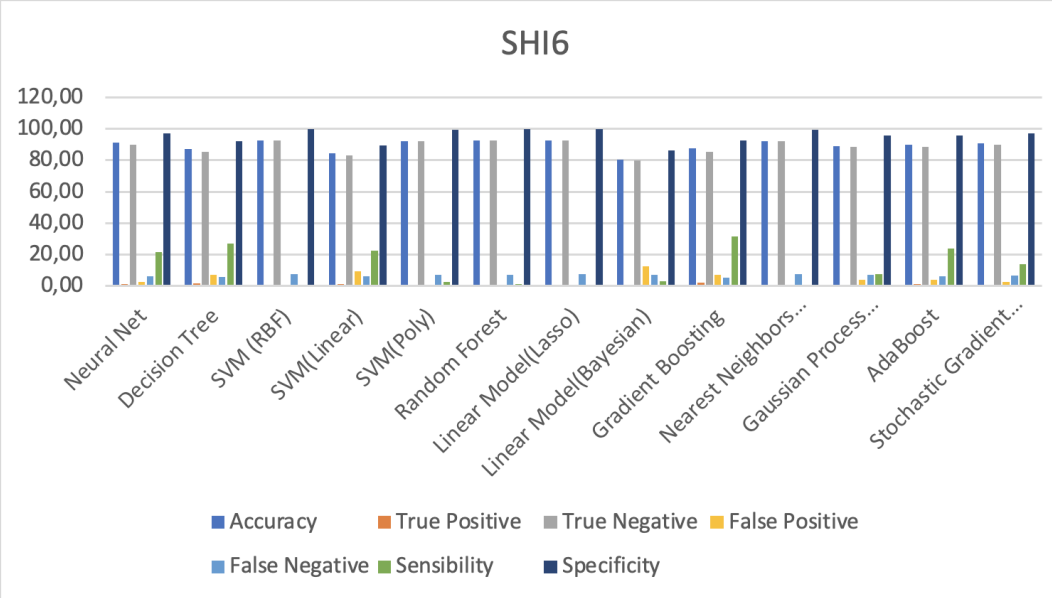

Figure 5.19. Predictions For SHI6 Feature

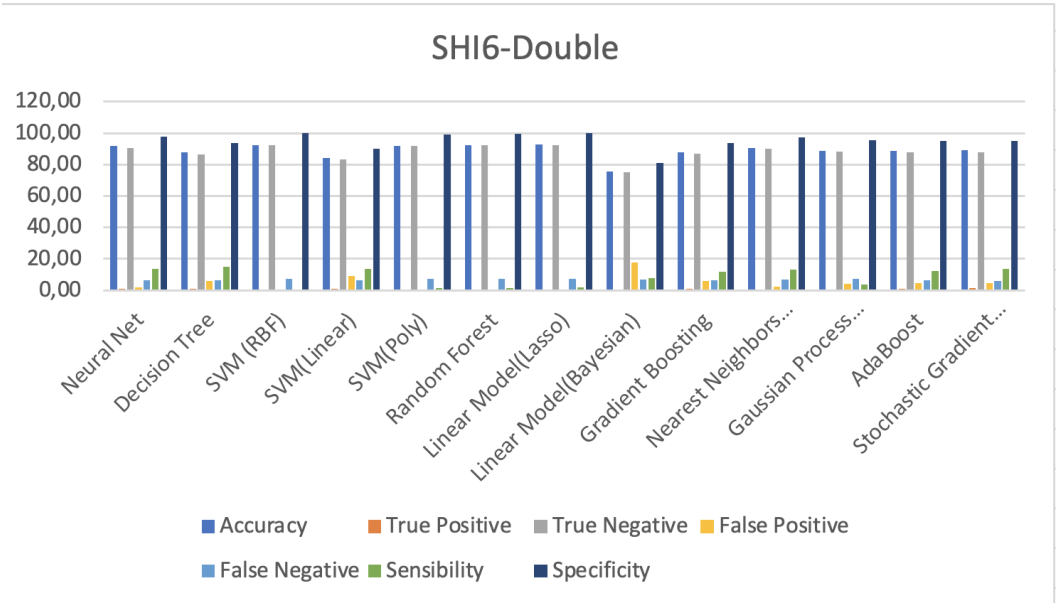

Figure 5.20. Predictions For SHI6-Double Feature

SHI7. You have entered into a relationship in which you felt rejected or humiliated sexually or psychologically?

The percentage of positives present for this variable in the dataset is 5%, which becomes 10% in the runs performed for the dataset with doubled positives. For the prediction of this variable, we can see how the greatest accuracy is achieved with the Neural Network and Random Forest methods, in this case, the Random Forest is to be preferred as it has a smaller number of false negatives. Regarding the dataset with doubled positives, we do not notice a significant change in performance.

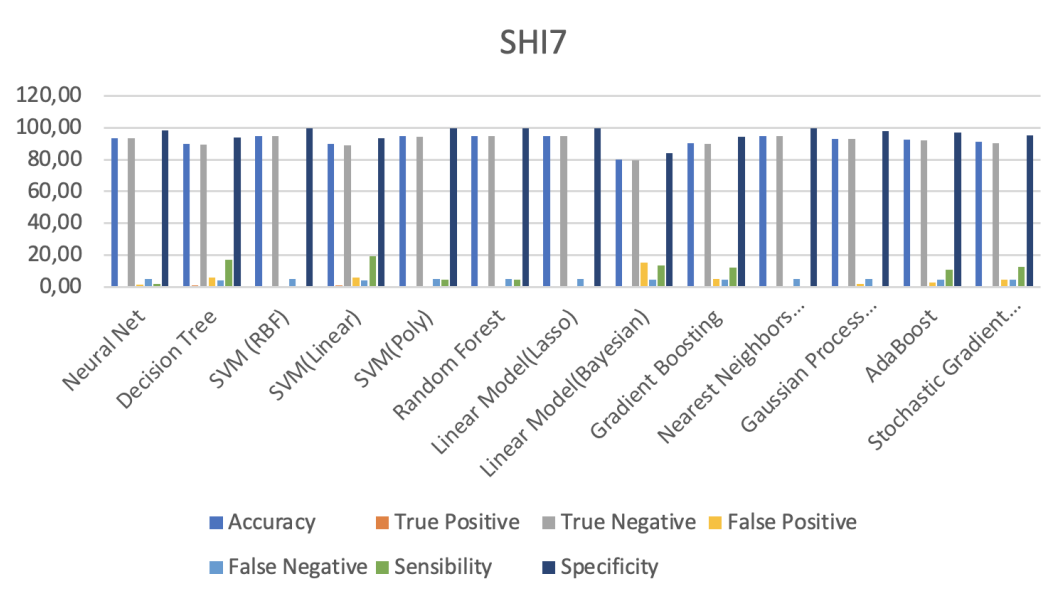

Figure 5.21. Predictions For SHI7 Feature

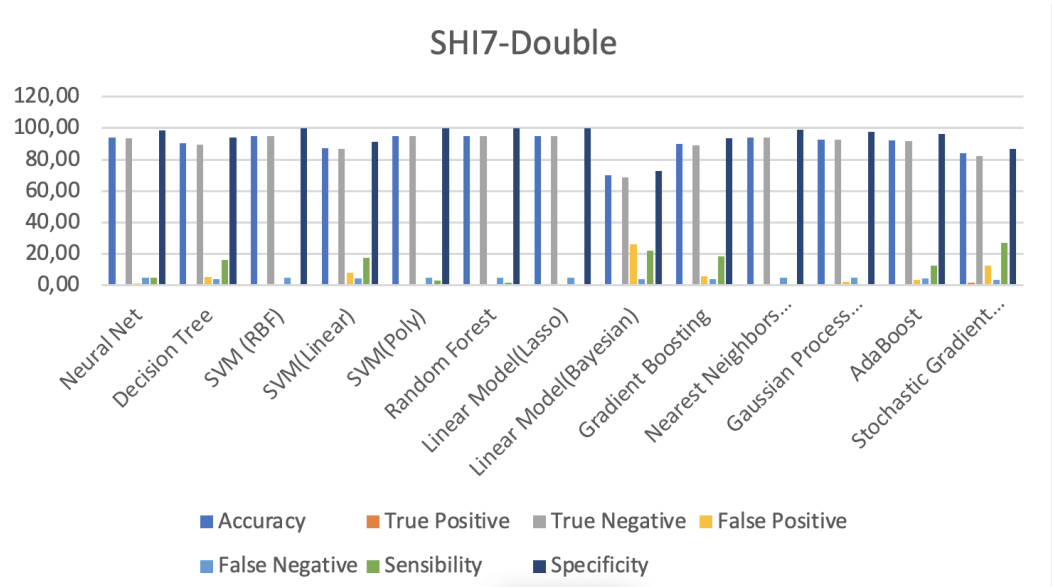

Figure 5.22. Predictions For SHI7-Double Feature

SHI-TOT. At least one of the conditions listed above.

The percentage of positives present for this variable in the dataset is 35%, which becomes 70% in the runs performed for the dataset with doubled positives. We can consider this indicator as to the most important for two fundamental reasons.

The first is that being an amalgamation of the previous indicators, it presents in itself a unique way to recognize various characteristics of self-harm and suicidal ideation. The second is that being an amalgamation, it has a higher percentage of false positives than the previous indicators. This means that Machine Learning techniques are in a position to operate on a perfectly balanced dataset. Maximum accuracy, combined with the lowest number of false negatives, is obtained with the Neural Network, which offers significantly better performance than all other methods. This difference is even more marked in the dataset which presents doubled positives.

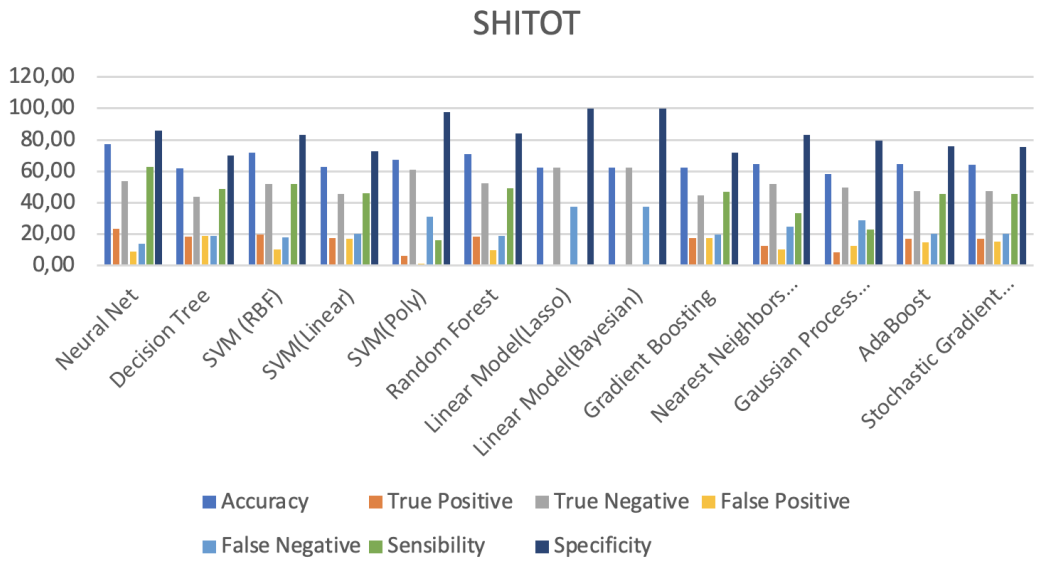

Figure 5.23. Predictions For SHI-TOT Feature

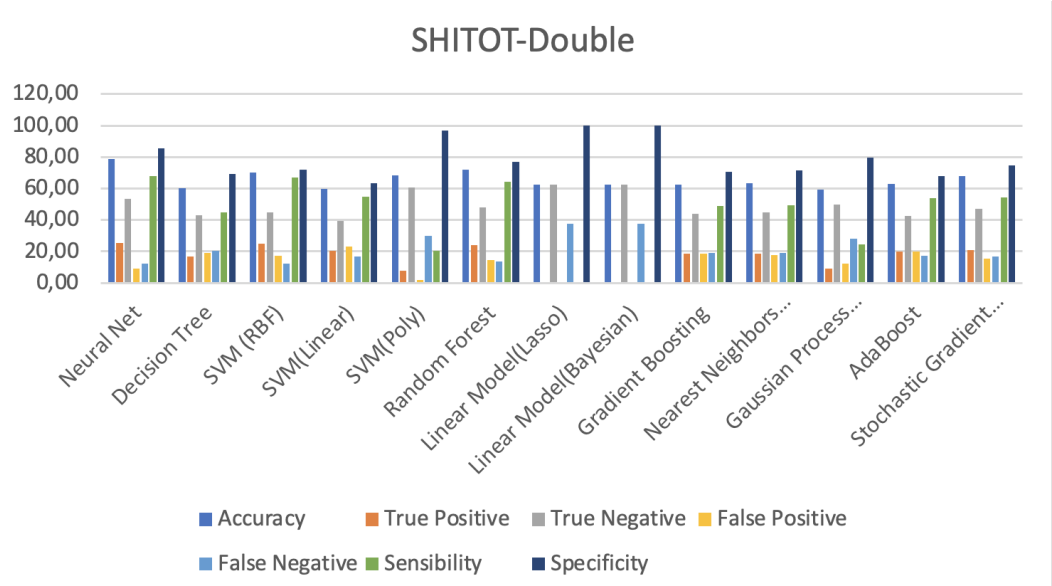

Figure 5.24. Predictions For SHI-TOT-Double Feature

Chapter 6

Conclusions

In this final chapter, we conclude the thesis by reporting the achieved objectives, the encountered difficulties, and some directions for future work.

We achieved all the objectives of the thesis, in fact, we implemented up to 13 machine learning approaches to analysis SHI indicators obtaining satisfactory results. We also implemented Feature Importance approach to better explain the obtained results with the realized neural network.

While studying the analysis techniques we were able to notice that all the analyzes had pros and cons. The main discriminating factor for the performance of these analysis was the structure of the data. We noted that in the specific case of psychiatric data, linear techniques, boosting and K-Neighbor-Classification gave lower results than the alternatives. Despite this, these techniques have the great advantage of being extremely light, therefore usable on large datasets with hardware with minimum performance.

On the contrary, Machine Learning techniques such as the Neural Network, the Random Forest, and Support Vector Machine have provided much more accurate results but required more computational resources. For the size of our dataset, however, the performance of the machine has never been a problem, as a 268 X 1020

matrix is not able to put the above-mentioned techniques in difficulty.

It was also interesting to work on the data pre-processing part 3, as it allowed us to thoroughly explore the structure of the dataset but also the features of the used psychiatric questions. Indeed, we were able to enter more into the domain, allowing us to customize and use Machine Learning and Feature Importance techniques based on the context.

Thanks to the knowledge of the domain during the customization of Machine Learning methods 4, we were able to go straight on the use of the Rectified Linear Unit (ReLU) activation function for the Neural Network (which generally works well on this type of medical data). In this way, we were able to reduce the number of configurations to test, not having to analyze configurations that we know a priori do not work well with this type of data. From the point of view of understanding the domain, it was useful to discuss with psychiatrists about the features to predict and the problem under analysis in order to better understand the used indices. This provided a more accurate understanding of the results, allowing psychiatrists to have an additional useful tool in treating and preventing self-harm and suicidal ideation.

As for the results in Chapter 5, it was very useful to compare the results obtained by Feature Importance with those obtained by Machine Learning methods. We have combined the two techniques managing to derive a system that combines the accuracy of Machine Learning methods with the explanation given by Feature Importance. Thanks to this double tool, any psychiatrist can decide to rely on both the accuracy of the predictions and the affirmative response to the questions that the Feature Importance methods have highlighted as being more related to self-harm and suicidal ideation. With regard to Machine Learning methods, the Neural Network is the method that on average provides both the highest prediction accuracy and the minimum number of false negatives. Another note of merit goes to the Random Forest method, as it approaches the performance of the Neural Network

but uses much fewer resources; it is, therefore, conceivable to use it on hardware systems that do not have access to high resources.

Obviously, some difficulties arose during the course of the research project. Some of these difficulties concerned the IT part as, thanks to the study of technologies, we have managed to avoid blocking problems. For example, we had to face the problem of proprietary formats used to store the data generated by psychiatrists. Such formats can be managed only by some medical programs that poorly interface with the formats required by computer analysis techniques and, more in the specific, of Machine Learning.

Other difficulties came from the multidisciplinary characteristic of the research projects. In fact, we have faced the problem to reduce the knowledge gap between computer scientists and the psychiatric ones. At the beginning of the project, the psychiatric team had research ideas that were not feasible in practice or that, on the contrary, were not of interest to the computer scientist team. Thanks to numerous meetings we were able to find a common point that managed to satisfy both the requirements of the psychiatric team and the requirements of computer scientists. Notice that, knowledge gap filling can also solve technological problems. For example, the data format problem we reported earlier can be easily solved showing the psychiatrists that they can export, from their systems, data having a standard format such as .csv.

As future work of the Dual Trauma research project we reported in this thesis we identify at least two research directions. The first study that could develop from Dual Trauma is about further expanding the Feature Analysis. In the future, we could have a questionnaire that contains less than half of the questions while maintaining the same forecast accuracy and allowing a considerable time saving for psychiatrists as well as less stress for the subjects who have to fill it out. Another possible development is to try to make Dual Trauma techniques more general so

that the results can be expanded not only to subjects who have experienced natural disasters but also to other types of subjects.

.1 Acronyms

dm01 Dob

dm02 Gender

dm03 Weight

dm04 Height

dm05 Italian

dm06 Sent Report

dm07 Mother Dob

dm08 Mother Job

dm09 Mother Nationality

dm10 Mother Edu

dm11 Father Dob

dm12 Father Job

dm13 Father Nationality

dm14 Father Edu

dm15 Common Home

dm16 Housing

dm17 No Occupants Husehold

dm18 Marital Status Parents

dm19a Lives With Father

dm19b Lives With Mother

dm19c Lives With How Many Brothers

dm19d Lives With Mother'S Partner

dm19e Ilves With Father Partner

dm19f Lives With Other Relatives

dm19g Lives With Grandparents

dm19h Lives With Own Partner

dm19i Lives With Others

dm20 In The Past You Meet With Psi

dm21 In The Last Month You Meet Psy

dm22 Previous Ailments

dm23 Medicianle For More Than 3 Months

__v1 Which Medicine Dm23

dm24 Drugs Last 15 Days

dm25 Familiar Psy

__v2 Relative In Care

__v3 Reason Care Relative

dm26 Owns Vehicle

dm27 Bedroom Alone

dm28 Computer Or Tablet In The Family

dm29 Travel Abroad Last Year

dm30 Average School Grades

pace01 Last Week Physical Activity> 1 H

pace02 Physical Activity Typical Week Last 6 Months

__v4 Initial Insomnia

__v5 Average Insomnia

__v6 Late Insomnia

isi02 Current Sleep Satisfaction

isi03 Interference Insomnia Daytime Activity

isi04 Insomnia Eident To Others

isi05 Insomnia Concern

cig01 Cigarettes Die

cast00 I Use Cannabis

cast01 Approx In The Morning

cast02 Ca Alone

cast03 Memory Problems

cast04 They Told You To Stop

cast05 Tried To Quit

cast06 Problems For Approx

au01 Alcohol Consumption Frequency

au02 On The Days You Drink, How Many Au

au03 Frq. > 6 Glasses At A Time

piuq01 Tense Irritable If Not Internet

piuq02 Want To Reduce T Online

piuq03 Online Instead Of Sleeping

piuq04 Hide T Online

piuq05 People Complain T Online

piuq06 Depressed If Offline

gpss01 Skipped Activity For Game

gpss02 Not Dated With Non-Players

gpss03 Fre Planned Game

gpss04 Felt Bad For Fun

gpss05 Return To Play To Win Back

gpss06 Hide Game

gpss07 Heard Problem With Game

gpss08 Spent More Money

gpss09 Stolen

rfq02 Adults Have Insulted Or Humiliated You

rfq05 Lived With An Alcoholic Or Drug Addict

rfq07 Violent Family Member With Another Family Member

rfq12 Chaotic Environment

rfq13 Neglect

obq01 Offensive Nicknames

obq02 Excluded Or Ignored

obq03 Physically Assaulted

obq04 Gossip

obq05 They Stole Things From You

obq06 Threatened

item01_inf Diagnosed Disease

item02_inf Horrible Death

item03_inf A Loved One Was Diagnosed With Lethality

item04_inf Threatened

item05_inf Attacked By Parent

item06_inf Assaulted By Non-Parent

item07_inf Sexually Assaulted By A Parent

item08_inf Sexually Assaulted By Non-Parent

item09_inf Harassment

item10_inf War

item11_inf Torture

item12_inf You Have Inflicted Suffering On Others

item13_inf Witness Of Death

item14_inf Accident

item15_inf Natural Disaster

_v7 Earthquake Aq

_v8 Injured Or Trapped

_v9 Uninhabitable House

_v10 Someone Close Dead Or Rubble

item16_inf Man-Made Disaster

item17_inf Stalking

item18_inf Bullying

item19_inf Humiliated Or Mortified

item20_inf Felt Unloved Or Unwelcome

item21_inf Diagnosed Disease

item22 Other

item23 More Serious Item

item24 How Many Times In Life

item25 How Long Ago

item26 Predominant Emotion

rq_cat Category Style

rq_a Secure Attachment

rq_b Worried

rq_c Fearful

rq_d Avoidant Attachment

rsa01 I Feel My Future

rsa02 I Feel ... With My Family

rsa03 They Are Good At Encouraging Me

rsa04 My Goals

rsa05 New Friendships Are Something

rsa06 My Family Is Characterized By ... Union

rsa07 I'M Good At ... Organizing Time

rsa08 Believe In Myself

rsa09 Making New Acquaintances Is ... Difficult

rsa10 In Difficult Times I Tend To ...

rsa11 I Get Support From

sdq01 I Try To Be Kind To Others

sdq02 I Am Restless

sdq03 I Often Suffer From Headaches

sdq04 I Gladly Share With Others

sdq05 Fits Of Temper

sdq06 I Prefer To Be Alone

sdq07 I Am Willing To Do What Others Want

sdq08 I Have A Lot Of Worries

sdq09 They Are Helpful If Someone Gets Hurt

sdq10 I Am Constantly Restless

sdq11 I Have At Least One Good Friend

sdq12 I Often Quarrel

sdq13 Often Unhappy Or Sad

sdq14 I Am Accepted By Others

sdq15 I Get Distracted Easily

sdq16 Nervous In New Situations

sdq17 Kind To Children

sdq18 Accused Of Being A Liar

sdq19 Targeted By Other People

sdq20 I Often Offer To Help Others

sdq21 I Think Before I Do Something

sdq22 I Stole

sdq23 Better Relationships With Adults

sdq24 Many Fears, I Get Scared Easily

sdq25 I Finish What I Start

des01 Being In Places Without Remembering How

des02 Find New Items

des03 Feel Behind Yourself

des04 I Don'T Recognize Others

des05 The World Is Not Real

des06 Body Does Not Belong To Me

des07 Feel Like Two Different People

des08 Voices Inside The Head

shi01 Alcohol Drug

shi02 Cuts Wounds

shi03 Banged My Head

shi04 Wounds Heal

shi05 Worsen Medical Condition

shi06 Promiscuous Behavior

shi07 Abusive Relationship

shi08 Other

itq00 Worst Experience

__v11 When It Happened

itq__re1 Dreams

itq__re2 Flashback

itq__av1 Internal Avoidance

itq__av2 External Avoidance

itq__hy1 Hyperarousal

itq__hy2 Startle

itq__dist1 Influence On Social Life

itq__dist2 Influence On School

itq__dist3 Influences Other Aspect Of Life

itq__ad1 A Long Time To Calm Down

itq__ad2 Numbing

itq_nsc1 Feel Failed

itq_nsc2 I Feel Worthless

itq_dr1 I Feel Distant Or Cut Off From People

itq_dr2 Difficult To Be Emotionally Close To Other People

itq_dist4 Effect On Relationships

itq_dist5 Effect On School

itq_dist6 Effect On Other

phq01 Little Interest Or Pleasure

phq02 Feeling Depressed

phq03 Insomnia

phq04 Anergy

phq05 Appetite

phq06 Guilt

phq07 Concentration

phq08 Slowdown / Agitation

phq09 Better To Be Dead

corer01 I Hurt Myself

corer02 Better To Be Dead

corer03 Suicide Project

gad01 Nervous Anxious

gad02 Restless

gad03 Tired Out

gad04 Voltage

gad05 Insomnia

gad06 Concentration

gad07 Irritability

dca01 I Don'T Control Food

dca02 Binge

dca03 At Least 2 Times A Week In 3 Months

dca04 He Retched

dca05 Laxatives

dca06 Fast

dca07 Physical Activity

pq_q01 Lost Interest

pq_q02 Dejavu

pq_q03 Olfactory Hallucinations

pq_q04 Elementary Auditory Hallucinations

pq_q05 Confused About The Reality Of The Experience

pq_q06 Seen Face Change

pq_q07 Anxious When I Meet Someone

pq_q08 Visual Hallucinations

pq_q09 Audible Thought

pq_q10 Special Meaning Around Me

pq_q11 Control Over Ideas And Thoughts

pq_q12 Distracted By Distant Sounds

pq_q13 Complex Auditory Hallucinations

pq_q14 Paranoia

pq_q15 Person Or Force Stand Beside Me

pq_q16 Somatic Transformation

pq_s01 Lost Interest

pq_s02 Dejavu

pq_s03 Olfactory Hallucinations

pq_s04 Elementary Auditory Hallucinations

pq_s05 Confused About The Reality Of The Experience

pq_s06 Seen Face Change

pq_s07 Anxious When I Meet Someone

pq_s08 Visual Hallucinations

pq_s09 Audible Thought

pq_s10 Special Meaning Around Me

pq_s11 Control Over Ideas And Thoughts

pq_s12 Distracted By Distant Sounds

pq_s13 Complex Auditory Hallucinations

pq_s14 Paranoia

pq_s15 Person Or Force Stand Beside Me

pq_s16 Somatic Transformation

List of Figures

Bibliography

[1] Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, and Ethan Fetaya. *GP-Tree: A Gaussian Process Classifier for Few-Shot Incremental Learning.* 2021.

[2] Francesco Altamore, Iolanda Grappasonni, Neelam Laxhman, Stefania Scuri, Fabio Petrelli, Giuliana Grifantini, Pamela Accaramboni, and Stefan Priebe. Psychological symptoms and quality of life after repeated exposure to earthquake: A cohort study in italy. *PLOS ONE*, 15(5):1–6, 05 2020.

[3] David Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, USA, 2012.

[4] Léon Bottou. *Large-scale machine learning with stochastic gradient descent.* 2010.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Ivan Nunes da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino, Luisa Helena Bartocci Liboni, and Silas Franco dos Reis Alves. *Artificial Neural Networks: A Practical Course.* Springer Publishing Company, Incorporated, 1st edition, 2016.

[7] Esma Duncan, Martin Dorahy, Donncha Hanna, Sue Bagshaw, and Neville Blampied. Psychological responses after a major, fatal earthquake: The effect

of peritraumatic dissociation and posttraumatic stress symptoms on anxiety and depression. *Journal of trauma & dissociation : the official journal of the International Society for the Study of Dissociation (ISSD)*, 14:501–18, 10 2013.

[8] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[9] Adam Horvath, Mark Dras, Catie C.W. Lai, and Simon Boag. Predicting suicidal behavior without asking about suicidal ideation: Machine learning and the role of borderline personality disorder criteria. *Suicide and Life-Threatening Behavior*, 51(3):455–466, 2021.

[10] Hiroko Kukihara, Niwako Yamawaki, Kumi Uchiyama, Shoichi Arai, and Etsuo Horikawa. Trauma, depression, and resilience of earthquake/tsunami/nuclear disaster survivors of hirono, fukushima, japan. *Psychiatry and Clinical Neurosciences*, 68(7):524–533, 2014.

[11] Gen-Min Lin, Masanori Nagamine, Szu-Nian Yang, Yueh-Ming Tai, Chin Lin, and Hiroshi Sato. Machine learning based suicide ideation prediction for military personnel. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1907–1916, 2020.

[12] Kazunori Matsumoto, Atsushi Sakuma, Ikki Ueda, Ayami Nagao, and Yoko Takahashi. Psychological trauma after the great east japan earthquake. *Psychiatry and Clinical Neurosciences*, 70(8):318–331, 2016.

[13] David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. *Neurocomputing*, 55(1â2):169 – 186, 2003. <ce:title>Support Vector Machines</ce:title>.

[14] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY, 2009.

[15] R Muthukrishnan and R Rohini. *LASSO: A feature selection technique in predictive modeling for machine learning.* 2016.

[16] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[17] Raúl Rojas. *AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting.* 2009.

[18] Li Wang, Zhanbiao Shi, Yuqing Zhang, and Zhen Zhang. Psychometric properties of the 10-item connor–davidson resilience scale in chinese earthquake victims. *Psychiatry and Clinical Neurosciences*, 64(5):499–504, 2010.

[19] Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Gunluk. *Generalized Linear Rule Models*, volume 97 of *Proceedings of Machine Learning Research.* PMLR, 09–15 Jun 2019.