UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Performance comparison between random forest and multilayer perceptron prediction models in a classification problem on substance consumption based on psychological indicators.

Mateo Alejandro Ayala Tola

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 30 de abril de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Performance comparison between random forest and multilayer perceptron prediction models in a classification problem on substance consumption based on psychological indicators.

Mateo Alejandro Ayala Tola

Nombre del profesor, Título académico **Noel Pérez Pérez, PhD in Computer Science**

Quito, 30 de abril de 2021

© DERECHOS DE AUTOR

Nombres y apellidos: Mateo Alejandro Ayala Tola

Código: 00136404

Cédula de identidad: 1722219795

Lugar y fecha: Quito, 30 de abril de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El ambiente universitario, tanto a escala global y ecuatoriana, presenta una situación de alto estrés psicológico que causa abuso de substancias en estudiantes. Una posible solución es usar algoritmos de predicción de minería de datos para detectar este abuso. *Multilayer Perceptron* son usualmente usados para estas predicciones, pero reentrenarlos es usualmente costoso. Se propone los *Random Forest* como una alternativa, y su desempeño es evaluado y comparado al desempeño de un *MLP* en un problema multiclase de clasificación de consumo de alcohol usando indicadores psicológicos y metadatos. El set de datos usados pasó por un proceso de selección de atributos para descartar los altamente correlacionados. Pasó también por un proceso de re-muestreo para compensar desbalance de clases. Para optimizar el *MLP*, 100 configuraciones aleatorias se generaron y entrenaron en el set de datos y aquella con la accuracy más alta fue usado en el problema de clasificación. Dos métodos fueron usados para optimizar el *RF*. El primero exploró todas las configuraciones posibles exhaustivamente, el segundo lo hizo parcialmente. Ambos métodos llegaron a la misma configuración óptima, que fue usado en el problema de clasificación. Para ambos modelos, el *accuracy*, *precision*, y *recall* promedios fueron calculados. El tiempo de experimentación, clasificación, y calificación fue contado. Las curvas de *Precision-Recall* y *Receiver Operator Characteristics* fueron dibujadas. En desempeño, ambos modelos fueron iguales y exitosos. En tiempo, ambos métodos de experimentación del *RF* fueron más rápidos el del *MLP*. El método de exploración parcial fue un orden de magnitud más rápido que cualquier otro método. Por lo tanto, se puede concluir que los *RF* pueden ser un modelo alternativo viable en problemas de predicción de uso de substancias en base a indicadores psicológicos

Palabras clave: Foresta Aleatoria, Perceptrón Multicapa, desempeño, multiclase, indicadores psicológicos, consumo de substancias

ABSTRACT

The university environment, both on a global and Ecuadorian scale, presents a high-stress psychological situation that causes substance abuse in students. A possible solution is using data mining prediction algorithms for early detection of such consumption. Usually, Multilayer Perceptron models are used for such prediction but retraining them is usually costly. Random Forests are proposed as an alternative, and their performance is evaluated and compared to the performance of an MLP in an alcohol consumption multi-class classification problem using psychological indicators and metadata. The used dataset underwent a feature selection process to discard those highly correlated. It also underwent a resampling process to account for class imbalance. To optimize the MLP, 100 random configurations were generated and trained on the dataset and the one with the highest accuracy was used in the classification problem. Two methods were used to optimize the RF. The first one explored all possible configurations exhaustively, the second one did so partially. Both methods arrived at the same optimal configuration, which was used in the classification problem. For both models, the average accuracy, precision, and recall scores were calculated. The runtime of the experimentation, classification, and scoring stages was tallied. Precision-Recall and Receiver Operator Characteristics curves were plotted. Performance wise, both models were equal and successful. Timewise, both experimentation methods of the RF were faster than the MLP experimentation. The partial exploration method was an order of magnitude faster than any other method. As such, it can be concluded that RF may be a viable alternative model for substance use prediction problems based on psychological indicators.

Key words: Random Forest, Multilayer Perceptron, performance, multi-class, psychological indicators, substance consumption

TABLE OF CONTENTS

TABLE INDEX

Figure Index

INTRODUCTION

A recent study has linked psychological stress in all its forms with 75% to 90% of all human diseases (Liu, Wang & Jiang, 2017). These problems are more common on teenagers and young adults; between 13% and 17% of this age group presents symptoms, which is higher than the 10% of adult population worldwide (Kashani & Orvaschel, 1990). In this regard, university and college students are particularly at risks. Studies have revealed that University students exhibit more mental health afflictions than any other social group of the same age, surpassing even working young adults (Saleh, Camart & Romo, 2017) (Blanco *et al,* 2008). It has been discovered that traditional evaluation methods, such as midterms, final exams, and tests, incurs in a severe rise of stress level in students (Ávila-Toscano, Hoyos Pacheco, González, & Cabrales Polo, 2011). This stress, in turn, can lead to emotional pathologies such as depression and anxiety, which have an adverse effect in academic performance, since patients exhibit decreases in memory, attention, and concentration (Borges & Angeli Dos Santos, 2016). The traditional university social and academic environment, paired with these additional psychological stressors, has adverse effects in the well-being of students, pushing them towards the consumption of dangerous substances such as alcohol, tobacco, and other detrimental drugs (Ortiz *et al,* 2008). In Ecuador, in 2012, the rate of young adults that attended public or private Universities was around 39.6% of the population, with an expected attendance ratio of 50% on 2017 (Secretaría nacional de planificación y Desarrollo, 2016). Many of the psychological investigations carried out in Ecuadorian universities have sought to validate a series of psychological tools in this context, including the PSS (Ruisoto, López-Guerra, Paladines, Vaca & Cacho, 2020)., the CAPS (Vicent *et al,* 2020), and the GADS (Reivan-Ortiz, Pineda-Garcia, & León Parias, 2019), with successful results. These studies show that most common psychological tools are valid in an Ecuadorian context, which also

implies that most psychological problems are also prevalent in the Ecuadorian context (Torres *et al,* 2017). The growth in Ecuador's university student population, as well as the similarities this context shows with other contexts worldwide, means that Ecuadorian students also face the stressors and situations that may push them towards the use of dangerous substances.

Simultaneously, data mining has had successful applications in psychology. It was used to successfully develop an alternative treatment for the long-term effects of depression with better results than orthodox methods (Li & Ding, 2019). On the development of psychological tools that employ data mining, a three-layered model (data and data mining layer, prediction layer, and user interface layer) has been proposed with adequate performance (Ratnaparkhi, Katore & Umale, 2015). Additionally, SVM machines have been successful at predicting certain disorders with therapy-collected data (Correia, Trancoso & Raj, 2016). On data mining at the university environment, unsupervised clustering techniques were successful in identifying common social issues in freshmen students (Yuan, 2014), which are the ones most at risk of substance abuse (Ortiz *et al,* 2007). On a more psychological application, a study combined *a priori* algorithms with association rules to try to uncover possible relationships between psychological parameters and positive academic results (Burman, Som, Hossain & Sharma, 2019). A study on supervised models used an MLP network and the three-layered model with adequate results on the prediction of a binary well-unwell well-being scale (Qinghua, 2016). Another study used complex data gathering surveys and MLP variations to predict well-being on a 1 to 4 scale with great results, but with a very resource intensive data gathering method (a long and complex survey) and prediction model (Tyulyupo, Andrakhanov, Dashieva& Tyryshkin, 2018). It should be noted that the study was successful in verifying that the data gathering method could be simplified, but said

survey is still not publicly well known nor available, which makes it both difficult to replicate the study and to apply it practically. In general, these studies show the possibility to employ data mining in prediction models based on psychological data. One such model, for example could attempt to address the substance abuse issue, previously discussed, in universities by gathering psychological data from students in order to predict frequency of substance use. This, in turn, would allow competent professionals and authorities to help students that struggle with these problems in a timely manner.

Yet, the nature of this kind of work raises a complication. Psychological datasets used in these data mining applications are on a state of constant growth by nature, as more data is gathered from new students or events. This, in turn, means that the prediction models used in the applications must be periodically retrained to stay updated to the available data. Most of these studies use artificial neural networks, which have a high training complexity. As such, periodically retraining these models could become costly both in time and computational resources. It is then necessary to evaluate alternative prediction models with lesser training time and resource costs, in order to determine if their results could be comparable to those obtained by these more complex models. One such prediction model are random forests, which are very simple to implement and have low training times and costs. This project will, then, compare the performance of an MLP neural network prediction model (the most commonly used model in the studies) with that of a random forest prediction model, when applied to a multi-class classification problem using psychological indicators in the prediction of frequency of use of detrimental substances. First, the data will be treated to properly fit the models and minimize external error sources. Then, random MLP configurations will be trained in the data, and the most accurate one will be chosen as a baseline for the best possible prediction results of the problem. Later, an unoptimized random

forest classifier will be trained on the data and evaluated. The configuration of the random forest model will then be deterministically optimized to maximize accuracy. Finally, using both the best found MLP artificial network configuration and random forest configuration, performance metrics will be calculated independently. A final comparison will be carried out between the results of both training models, and conclusions about their comparative performance will be drawn.

MATERIALS AND METHODS

Dataset used

For investigative purposes, this project uses a dataset donated by investigators at Cornell University after a study, that is maintained and publicly made available through the UCI (University of California Irvine) Machine Learning Repository, at the following url: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29. The dataset counts with 1885 instances and 32 attributes per instance. The attributes are made up of metadata of the patient such as age, gender, country of residence, and ethnicity, psychological indicators for personality factors, impulsivity, and sensation seeking, and nineteen label attributes that specified the last time the participant had consumed a series of legal and illegal drugs. A detailed description of the attributes of the dataset is shown in Table 1.

Table 1: Description of the attributes found in the instances of the experimental dataset.

Attribute	Column numbers	Description
Identifier	1	Unique sequential identifier number, assigned incrementally to participants as the data was gathered.
Metadata	2-6	General data about the participants of the study. This included, in order, age, gender, education level, country of residence, and ethnicity. These values are numerically encoded.
Personality indicators	8-11	Personality indicators following the Big Five personality traits, gathered through the NEO-FFI-R personality inventory. These traits are, in the order they appear in the dataset, neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness.
Impulsive indicator	12	A measure of the participant's impulsivity as determined by the Baratt Impulsiveness Scale 11 (BIS-11).

Sensation seeking indicator	13	A measure of the participant's drive to seek new and novel experiences, as determined by the Impulsive Sensation Seeking scale (ImpSS)
Drug consumption label	14-32	A series of attributes that specify the last time the participant consumed a certain drug. Both legal and illegal drugs are included. The drugs recorded are as following: alcohol, amphetamines, amyl nitrite, benzodiazepine, caffeine, cannabis, chocolate, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine, semeron (fictional drug), and volatile substances.

Every attribute except the drug consumption labels was numerically encoded by the original dataset owners before it was donated. The drug consumption labels listed one of seven possible classes for the last time the drug was consumed. These clases were non-numeric and followed the legend of Table 2.

Table 2: Class label legend for the drug consumption label attributes

Class label	Last time the drug was used
CL0	Never used
CL1	Used over a decade ago
CL2	Used in the last decade
CL3	Used in the last year
CL4	Used in the last month
CL5	Used in the last week
CL6	Used in the last day

Multilayer Perceptron

Multi-layer perceptrons, often referred to as MLP are the most well-known and widely used type of artificial intelligence neural network structure. It is a fully connected feed-forward model that uses back propagation for training purposes. The MLP consists of an entry layer, which has a number of nodes equal to the number of features in each feature vector in the input data, an output layer, which consists of a single neuron in binary classification

problems or as many neurons as possible classes in multi-class classification problems, and an arbitrary number of hidden layers, which in turn can have an arbitrary number of neurons per layer. Figure 1 diagrams the general architecture of an MLP. Multi-layer perceptrons have been widely used and documented in data mining due to their simplicity and effectiveness and are particularly well suited for the classification of tabular data.

Figure 1: General diagram of an MLP

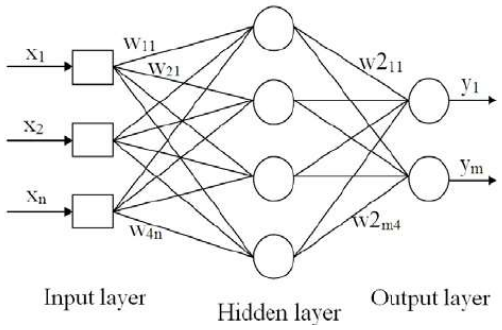


Image taken from Zainal-Mokhtar & Mohamad-Saleh, 2013

In this project, MLP classification models are used as the control group, in order to determine a baseline performance in the prediction problem at hand.

Random Forest

Random Forests are supervised learning algorithms that may be used to create both classification and regression models. For prediction, Random forests randomly select a series of attributes from the attribute vector of a dataset, and proceeds to build a decision tree with said attributes. The process is repeated until a determined number of trees are built. After this, the Random Forest utilizes a sampling through bagging method to generate a random subset of instances from the original dataset for each tree, and each tree is given a vote.

Figure 2 diagrams the general architecture of a Random Forest. The Random Forest predicts the category that has the most votes from its trees. Random forests are widely used due to the simplicity and reliability, as well as the low number of hyper-parameters that require tuning. They are also highly resistant to overfitting. A Random Forest's computational cost is entirely dependent on its number of member trees, and as such can vary greatly from case to case.

Figure 2: General diagram of a Random Forest

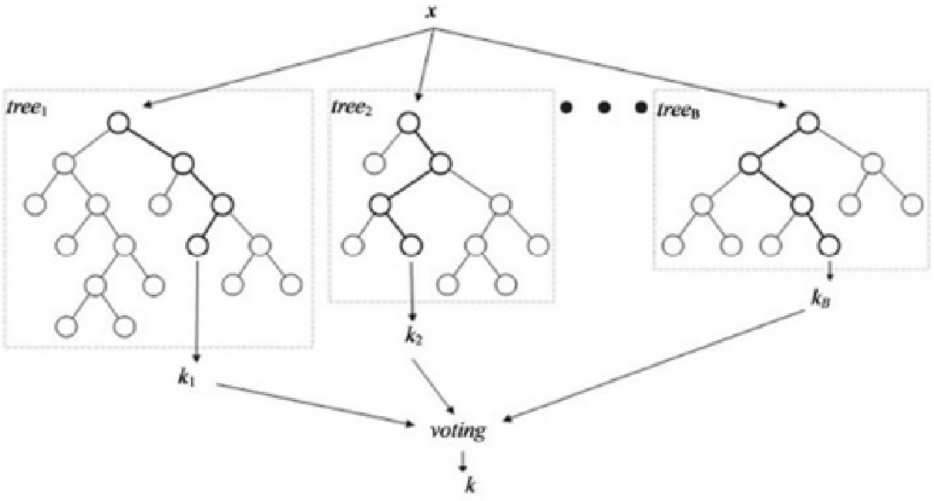


Image taken from Great Learning team, 2020.

In this project, Random Forest classifiers are considered the experimental group.

Experimental setup

Data processing.

Due to its prevalence and frequency of use in the University environment, Alcohol was chosen as the target column for the project and, as such, all other drug labels were ignored.

Additionally, all non-label attributes were normalized into the $[\mathbf{0}, \mathbf{1}]$ numeric range. The dataset then underwent a feature selection process in order to facilitate the training of the prediction models before their use. First, the total information gain ratio of every attribute against the classification label was calculated, and the attributes were sorted from highest information gain ratio to lowest. This was done to prioritize attributes with high information gain ratio in the selection of features. After this, the Pearson correlation index between every two attributes was calculated. Attributes with a high correlation index were considered undesirable. As such, any features found to have a Pearson correlation index $\boldsymbol{P > 0.3}$ with a previously evaluated feature were considered undesirable and discarded. The value of 0.3 was determined empirically to maximize the models' accuracy. The features that were found to be undesirable, as well as the feature they were found to be highly correlated with and the Pearson correlation index between the two, is shown on Table 3. The columns listed refer to the original dataset, not the filtered dataset.

Table 3: Features filtered out of the dataset.

Discarded feature column	**Discarded feature attribute**	**Highly correlated feature column**	**Highly correlated feature attribute**	**Pearson correlation index**
11	Personality Indicator: conscientiousness	8	Personality Indicator: neuroticism	0.308
2	Metadata: age	5	Metadata: country of residence	0.354
13	Sensation seeking indicator	9	Personality: extraversion	0.421

The resulting dataset was highly unbalanced. An unbalanced classification problem is defined as a classification problem where the number of instances of each class is vastly different

from one another. This may introduce errors in the training process. As such, the dataset was resampled. The resampling technique used was the SMOTEEN algorithm, which combines oversampling and undersampling (Batista, Bazzan, & Monard, 2003) (Batista, Prati, & Monard, 2004). Through oversampling and undersampling, the resulting dataset is much more balanced, and likely to produce more accurate prediction results. The number of instances of each class before and after resampling is presented in table 4.

Table 4: Number of member instances per class before and after SMOTEEN resampling

	CL0	**CL1**	**CL2**	**CL3**	**CL4**	**CL5**	**CL6**
Before resampling	34	34	68	198	287	759	505
After resampling	697	689	637	524	362	36	122

Training and test partitions.

Throughout the project, the dataset must be used both for training and testing purposes for the prediction models being evaluated. This is achieved by partitioning the data through stratified k-fold method, utilizing ten folds. Through the use of the stratified k-fold method instead of the traditional k-fold method (Hargreaves, 2021), it is guaranteed that each group will have a percentage division between classes that is somewhat similar to that of the original dataset. Additionally, a static seed of 0 is used throughout the project, to ensure that the partitions are always the same. This is done to ensure stability and replicability throughout the experimentation.

Model configuration.

A hundred MLP models were randomly generated by the variation of the parameters shown in table 5. Table 5, additionally, displays the possible values that each of the parameters

could take, as well as a short explanation of the parameter in question. All MLP random models generated utilized stochastic gradient descent as the solver for weight optimization in training, and a batch size of 200 instances. For a complete list of each of the models generated, as well as their accuracy achieved over the problem, refer to Appendix A.

Table 5: Possible values in the generation of random MLP configurations

Variable	Possible Values	Description
Number of hidden layers	1, 2, 3	The number of hidden layers present in the MLP configuration
Number of neurons	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	The number of neurons present in every single hidden layer in the configuration
Activation function	identity, logistic, tanh, relu	Whether the MLP uses, in respective order, the identity function, the logistic sigmoid function, the hyperbolic tangent function, or the rectified linear unit function as the activation function for its neurons
Learning rate	0.1, 0.3, 0.5	The constant learning rate used in the stochastic gradient descent training of the configuration

For the Random Forest classifier, two configuration methods were proposed. The first method, referred to as the exhaustive method, explored every possible permutation of the available parameters, and calculated their corresponding accuracy. This guaranteed the highest accuracy model in the possible permutation space. For a complete list of the accuracy achieved by every possible combination, refer to Appendix B. The second method, referred to as the partial method, started with a fixed 100 trees in the forest, which is usually considered the default, and the number of features with the highest accuracy for said number of trees was found. Next, the number of features was fixed at the previously found best, and

the number of trees with the highest accuracy was found. This method did not guarantee the highest accuracy model but was more time efficient. Table 7 describes the possible parameters explored for the Random Forest classifier, as well as the possible values it could take, and a short description of the parameter itself.

Table 6: Parameters of the Random Forest classifier and their possible values

Parameter	Possible values	Description
Number of features	1, 2, 3, 4, 5, 6, 7, 8, 9	Number of randomly selected features from the original dataset used in the construction of each individual tree in the random forest
Number of trees	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	Number of trees in the Random Forest classifier

Additionally, an unoptimized model that used the values considered standard (the square root of the total number of features as the number of features and 100 trees in the forest) was also trained in the data, and its metrics recorded.

Assessment metrics.

The following metrics were calculated, tallied, or traced in order to compare the models between each other. To compare the two models, an independent-sample T-Test assuming unequal variance was used (Welch's t-test). Additionally, the null hypothesis chosen was that the average accuracy of both models was the same, with 95% certainty.

Global Accuracy.

Global accuracy refers to average of the total accuracy obtained in each fold of the dataset. Total accuracy refers to the sum of all correctly classified instances divided by the total of

instances in the fold. Global accuracy is described simply as accuracy throughout the investigation.

Average accuracy across classes.

Average accuracy across classes refers to the average of the seven class accuracies obtained from the combined results across the ten folds of the dataset. Class accuracy refers to the number of correct predictions of a particular class, both positive and negative, divided by the total number of instances in the dataset.

Average precision.

Average precision refers to the average of the seven class precisions obtained from the combined results across the ten folds of the dataset. Class precision refers to the number of correctly classified instances of a particular class (true positives), divided by the sum of the number of correctly classified instances of said class (true positives) and the number of instances incorrectly classified as said class (false positives).

Average recall.

Average recall refers to the average of the seven class recalls obtained from the combined results across the then folds of the dataset. Class recall refers to the number of correctly classified instances of a particular class (true positives), divided by the total number of said class in the dataset (true positives plus false negatives).

Experimentation time.

Experimentation time refers to the total runtime in the experimentation of each experimentation method. As such, three values were recorded: MLP experimentation, Random Forest exhaustive method experimentation, and Random Forest partial method experimentation.

Classification time.

Classification time refers to the total runtime of each selected model configuration in its training and testing process across the ten folds of the dataset.

Precision-Recall curve.

The precision recall curve traces the relationship across all their possible values. As such, both axes in this graph range from 0 to 1.

Receiving operator characteristics (ROC) curve.

The ROC curve traces the relationship between the true positive rate and the false positive rate of a binary classification problem. Once again, the axes range from 0 to 1.

Selection model.

The model selection in this investigation was carried out in two stages. First, the model configuration (both for the MLP as well as for the two experimentation methods of the Random Forest) with the highest accuracy was selected. Second, both architectures were compared and the one with the highest accuracy was selected. In case of a tie, the model with the shortest experimentation runtime was selected.

RESULTS AND DISCUSSIONS

Model configuration selection

For the MLP configurations, the average classification accuracy across the 100 models was 0.67615, with a standard deviation of 0.22687. The worst configuration found had three hidden layers, sixty neurons per layer, used the hyperbolic tangent function, and had a learning rate of 0.5. It achieved an accuracy of 0.1809499. The best configuration found also had three hidden layers, sixty neurons per layer, and used the hyperbolic tangent function as well, but had a learning rate of 0.3 instead of 0.5. It achieved an accuracy of 0.9549978. The best configuration, ((60, 60, 60), 'tanh', 0.3) was selected to compare with the Random Forest. Table 7 describes the accuracy obtained for said configuration across each fold.

Table 7: Accuracy of the best MLP configuration across every fold of the dataset

Fold	Accuracy score	Accuracy percentage
1	0.9543974	95.44%
2	0.9348534	93.49%
3	0.9641694	96.42%
4	0.9641694	96.42%
5	0.9478827	94.79%
6	0.9609121	96.09%
7	0.9804560	98.05%
8	0.9673203	96.73%
9	0.9346405	93.46%
10	0.9411765	94.12%
Average	0.9549978	95.50%

For the Random Forest configurations, both the exhaustive method and the partial method reached the same conclusion. The permutation with the highest accuracy found was 1 feature and 900 trees. This configuration achieved an accuracy of 0.9504450. As such, it was

selected to be compared against the best MLP configuration found. Table 8 describes the accuracy obtained for this configuration across each fold.

Table 8: Accuracy of the best Random Forest configuration across every fold of the dataset

Fold	Accuracy score	Accuracy percentage
1	0.9511401	95.11%
2	0.9315961	93.16%
3	0.9446254	94.46%
4	0.9413681	94.14%
5	0.9413681	94.14%
6	0.9543974	95.44%
7	0.9739414	97.39%
8	0.9673203	96.73%
9	0.9542484	95.42%
10	0.9444444	94.44%
Average	0.9504450	95.04%

Multi-class metrics

Table 9 details the multi-class metrics calculated on the global classification problem for the best MLP classifier configuration found, the optimized Random Forest, and the unoptimized Random Forest.

Table 9: Multi-class metrics for best MLP and Random forest configuration, and unoptimized Random Forest.

Model	Global accuracy	Average accuracy across classes	Average precision	Average recall
Multilayer Perceptron	0.9550048	0.8693319	0.9350511	0.8693318
Random Forest	0.9504401	0.8311933	0.9525095	0.8311933

Unoptimized Random Forest	0.9396805	0.8231180	0.9239716	0.8231180

Noticeably, the experimental Random Forest outperformed the unoptimized Random Forest at every metric, on the second negative order of magnitude. This indicates that the optimization process carried out in the Random Forest model does have a marked improvement in performance against classifiers of the same type that use the standard values. It is important to remember that said optimization process, incurs in a time cost, while using the default hyperparameters does not. Even though the second order of magnitude is considered of enough significance to justify the time cost that the optimization process requires.

The MLP outperforms the optimized Random Forest in global accuracy, average accuracy, and average recall, but noticeably the Random Forest outperforms the MLP on average precision. Even if precision shows an advantage on the second negative order of magnitude, which may be considered significance, the fact that every other metric was greater in the MLP model indicates that said advantage may be a case-dependent fluke. More interestingly, global accuracy only differs starting on the third negative order of magnitude, which can be considered small enough to assume equal performance. Backing this, the Welch's t-test value obtained in the comparison between the models with the values in tables 7 and 8 gave a value $t = 0.727708624$, and T value at $\alpha = 0.05$ and Welch's 17 degrees of freedom is $T = 2.109815578$. Since $t < T$, the null hypothesis (which was that the means differed) cannot be rejected. As such, it can be affirmed with 95% certainty that the models are equivalent in global accuracy.

Model runtime

As the models were equivalent in accuracy, the one with the shortest runtime will be chosen instead. Table 10 details the runtime of each model's total experimentation time, and their chosen model configuration classification time. As both the exhaustive and partial experimentation methods reached the same configuration, the runtime for their classification stage is the same. The scoring stage consists of the calculation of the model's multi-class metrics and the plotting of the appropriate curves.

Table 10: Runtimes for MLP, exhaustive method Random Forest, and partial method Random Forest

Metric	Best MLP configuration	Random Forest (Exhaustive method)	Random Forest (Partial method)
Experimentation time	5938.79 seconds	3454.59 seconds	231.96 seconds
Classification time	69.20 seconds	28.36 seconds	28.36 seconds
Scoring stage	0.87 seconds	0.81 seconds	0.81 seconds

On total experimentation runtime, the lowest runtime was achieved by the Random Forest using the partial optimization method, which is an order of magnitude smaller than the other two models. In comparison to the Random Forest that used the exhaustive experimentation method, runtime for the partial method is 14.9 times smaller than exhaustive method. It should be noted, though, that the partial optimization method does not guarantee the best possible tuning for the Random Forest model, while the exhaustive method does. Comparing the Random Forest using partial optimization method and the MLP, the experimentation time is 25.6 times smaller than the latter. Comparing Random Forest using exhaustive optimization and MLP, the former is 1.72 times smaller than the latter. Still, the random forest is definitely superior. On the classification stage, the runtime for both Random Forests

is 2.44 times smaller than that of the MLP. The scoring stage runtime for all models was so similar, it may be considered the same. Nonetheless, there was no statistical test to determine significance, so no superiority beyond this numbers can be affirmed. As such, it is simply concluded that the partial experimentation method's runtime is 14.9 times smaller than the exhaustive experimentation method in the Random Forest classifier, and 25.6 times smaller than the MLP's experimentation method.

Precision-Recall and ROC curves

Figure 3 to Figure 6 display the Precision-Recall and ROC curves plotted from the predictions for the chosen MLP and Random forest configurations, respectively, over the dataset. Additionally, the AUC values are also displayed. The dotted line represents the average of the plot of each class, which is shown with different colors.

Figure 3: Multi-class Precision-Recall curve for the MLP classifier

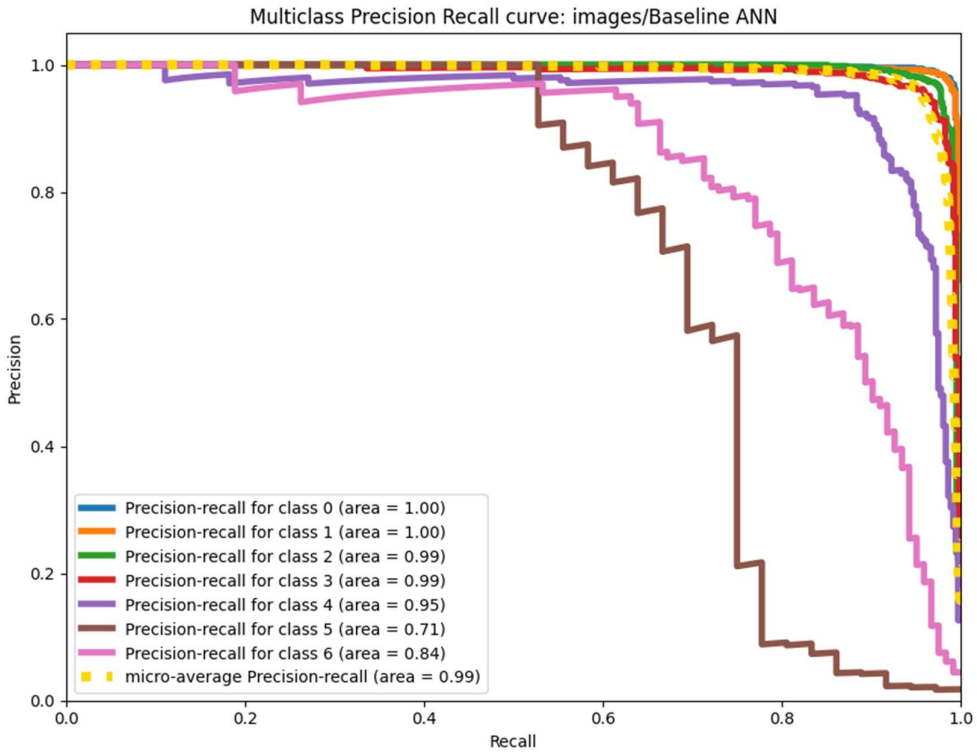


Figure 4: Multi-class Receiver Operating Characteristics curve for the MLP classifier

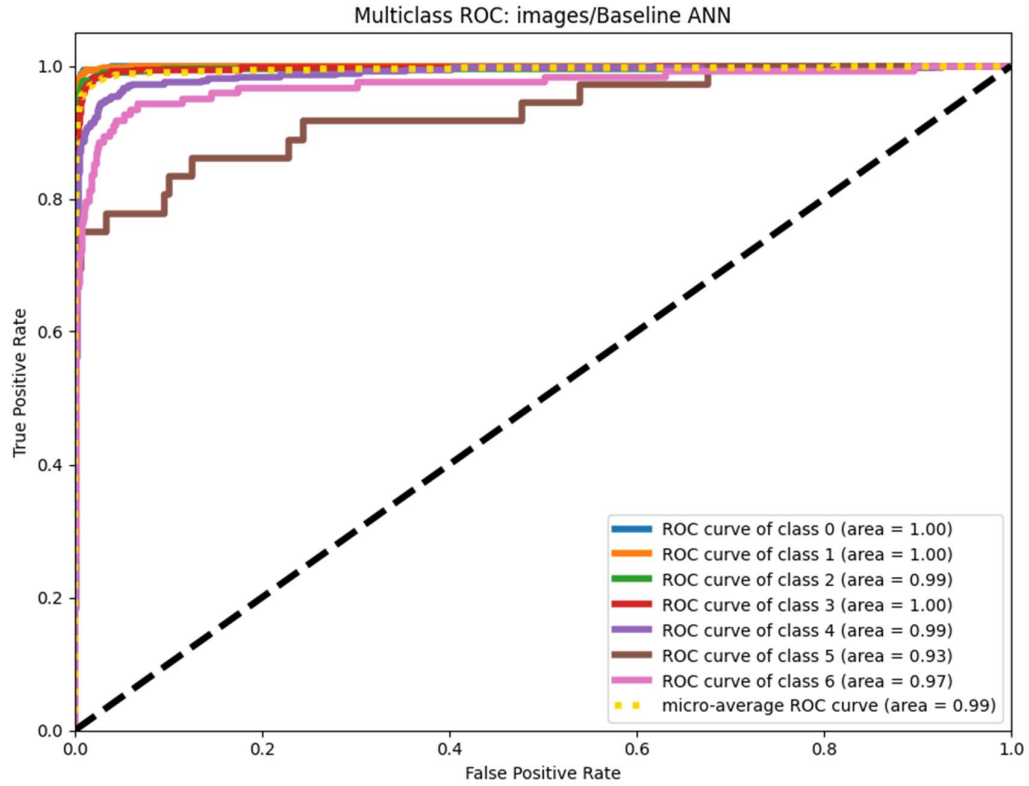


Figure 5: Multi-class Precision-Recall curve for the Random Forest classifier

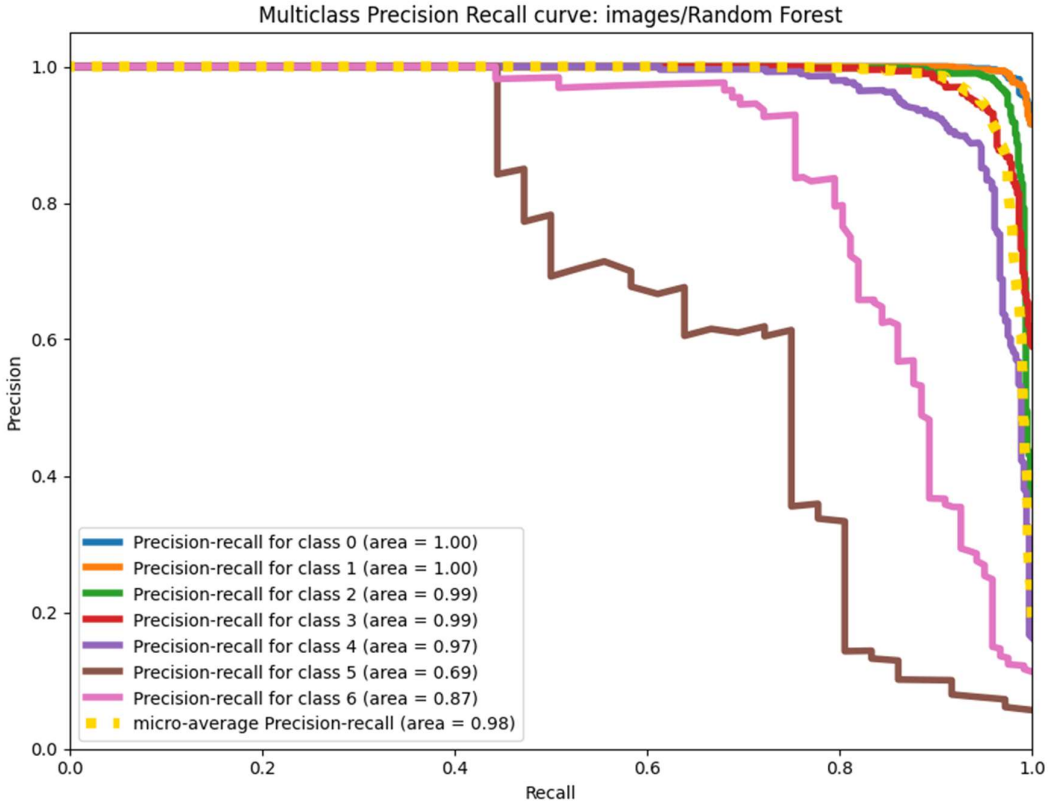
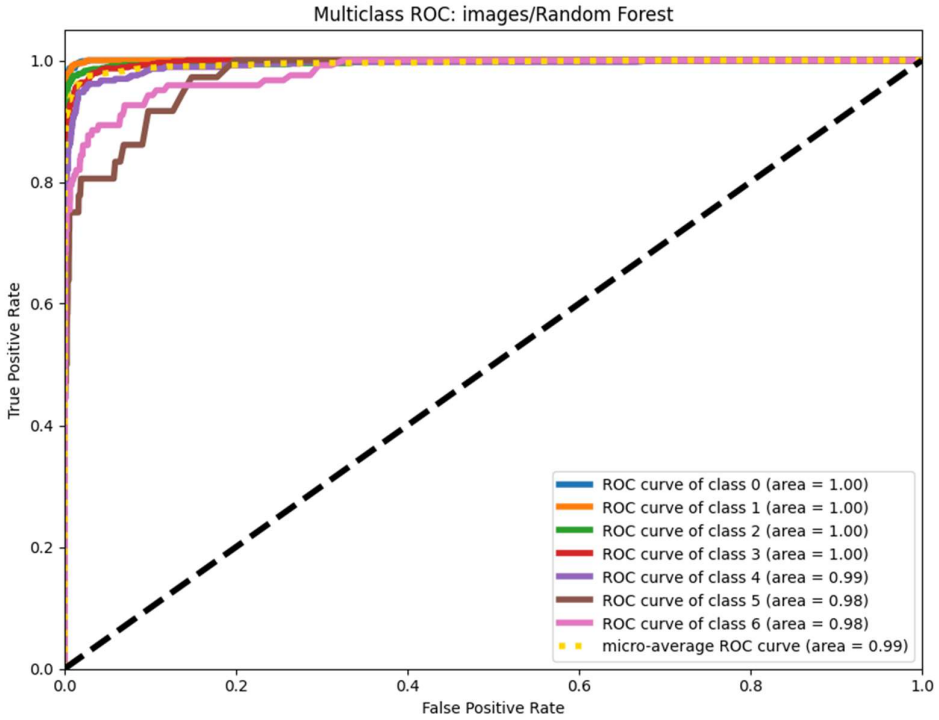


Figure 6: Multi-class Receiver Operating Characteristics for the Random Forest Classifier



In all scenarios, the average area under the curve value for the problem was over 0.98. From this, it can be concluded that the classification problem at hand was solved successfully with great results. Additionally, it should be noted that the curves for CL5 and CL6 classification across all diagrams are significantly lower than the rest. This may be due that, after resampling, there were significantly less instances of these classes in the dataset, which may have biased the models against them. Even though, all other classes have extremely positive curves and high AUC values, reaching an AUC of 1 across many instances. From this, it can be concluded that both of the chosen MLP and Random Forest configurations were successful in classifying frequency of alcohol consumption based on metadata and personality indicators.

CONCLUSIONS

Through the analysis of the multi-class metrics calculated and the curves plotted, it can be concluded that both the MLP classifier and the Random Forest classifier were highly successful in predicting frequency of alcohol use using psychological indicators and metadata in this particular problem, reaching values of average AUC of either 0.99 or 0.98 on all graphs. In terms of metrics performance, it was determined that the models had an equal average performance, with a 95% certainty. In terms of runtime, it was numerically determined that the partial method of experimentation was 14.9 times faster than the exhaustive experimentation method in Random Forest, and 25.6 times faster than the experimentation method for the MLP. Even if said time advantage is numeric and not statistical, this leads to the conclusion that Random Forests classifiers may be able to be used in the Ecuadorian University context to predict substance abuse with similar results to MLP and shorter training times. Even so, due to the numerical nature of this result, further investigation is required to ascertain this.

The following limitations of the study must be considered. The optimization process for the MLP is done through random generation and testing and, as such, offers no certainty in the maximum capabilities of MLP classification in the problem. Additionally, the possible configurations exploration spaces chosen for the optimization of the model were determined arbitrarily, and differently sized exploration spaces may yield different results both in metric scoring and in optimization time. Finally, it must be considered that these results are highly dependent on the dataset used, and different datasets and classification problems may yield

different results. On the same topic, the dataset itself was not built in Ecuador, so its results may not be applicable to the Ecuadorian social and psychological context.

There are possible future research options that could be used for the expansion of this investigation. First, a genetic optimization method may be proposed for the optimization of the MLP. This method would use genetic algorithms to sequentially improve the accuracy of the configurations generated and may yield a configuration with considerably better performance at the cost of a larger optimization time. Second, different feature selection methods may be used in the dataset. This could vary the overall performance of the models or favor one model over the other according to what the feature selection process prioritizes. Finally, the research may be replicated with a dataset built in either the Ecuadorian context or the Latin American context. This would validate that the results found were not specific to a certain social or cultural context. Unfortunately, as building such dataset would imply dealing with human subjects, doing would have ethical implications and, as such, a legal and ethical process would have to be started to obtain an authorization to gather the data.

BIBLIOGRAPHICAL REFERENCES

Ávila-Toscano, J.H., Hoyos Pacheco, S. L., González, D.P., & Cabrales Polo, A. (2011). Relación Entre Ansiedad Ante Los Exámenes, Tipos De Pruebas Y Rendimiento Académico En Estudiantes Universitarios. *Psicogente, 14*(26),255-268. ISSN: 0124-0137. Available at: https://www.redalyc.org/articulo.oa?id=497552359004

Batista, G. E., Bazzan, A. L. C, & Monard, M. C. (2003, December 3-5). *Balancing Training Data for Automated Annotation of Keywords: A Case Study.* [Paper Presentation]. II Brazilian Workshop on Bioinformatics, Macae, RJ, Brazil. Available at https://www.researchgate.net/publication/221322870_ Balancing_Training_Data_for_Automated_Annotation_of_Keywords_a_Case_Study

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6(*1), 20–29. https://doi.org/10.1145/1007730.1007735

Blanco, C., Okuda, M., Wright, C., Hasin, D. S., Grant, B. F., Liu, S. M., & Olfson, M. (2008). Mental health of college students and their non-college-attending peers: results from the National Epidemiologic Study on Alcohol and Related Conditions. *Archives of general psychiatry, 65*(12), 1429–1437. https://doi.org/10.1001/archpsyc.65.12.1429

Borges, L., & Angeli dos Santos, A.A. (2016). Sintomatología depresiva y desempeño escolar: un estudio con niños brasileños. *Ciencias Psicológicas, 10*(2), 189-197. Recuperado en 30 de abril de 2021, de http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-42212016000200008&lng=es&tlng=es.

Burman, I., Som, S., Hossain, S. A., & Sharma, M. (2019). Mining to Discover Association of Psychological Factors with Student Academic Performance. *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 432–435. https://doi.org/10.1109/iscon47742.2019.9036194

Correia, J., Trancoso, I., & Raj, B. (2016). Detecting Psychological Distress in Adults Through Transcriptions of Clinical Interviews. *Advances in Speech and Language Technologies for Iberian Languages,* 162–171. https://doi.org/10.1007/978-3-319-49169-1_16

Cui Yuan. (2014). Data mining techniques with its application to the dataset of mental health of college students. *2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)*, 391–393. https://doi.org/10.1109/wartia.2014.6976277

Great Learning Team. (2020, March 11). *Random Forest Algorithm- An Overview.* GreatLearning. https://www.mygreatlearning.com/blog/random-forest-algorithm/

Hargreaves, D. (2021, January 11). Stratified K-Fold: What It Is & How to Use It. Towards Data Science. https://towardsdatascience.com/stratified-k-fold-what-it-is-how-to-use-it-cf3d107d3ea2

Kashani, J. H., & Orvaschel, H. (1990). A community study of anxiety in children and adolescents. *The American journal of psychiatry, 147*(3), 313–318. https://doi.org/10.1176/ajp.147.3.313

Li, F., & Ding, Y. (2019). Data Mining in Cognitive Function Training of Depression Patients Applications. *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, 98–101. https://doi.org/10.1109/itme.2019.00033

Liu, Y.Z., Wang, Y. X., & Jiang, C. L. (2017). Inflammation: The Common Pathway of Stress-Related Diseases. *Frontiers in human neuroscience, 11,* 316. https://doi.org/10.3389/fnhum.2017.00316

Ortiz, R., et al. (2008). Prevalencia de la ansiedad en universitarios. Intraforo UV (2008). Memorias II Foro Intrauniversitario de Investigación en Salud. Universidad Veracruzana. Xalapa-Veracruz: México

Qinghua, J. (2016). Data Mining and Management System Design and Application for College Student Mental Health. *2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, 410–413. https://doi.org/10.1109/icitbs.2016.96

Ratnaparkhi, B., Katore, L., & Umale, J. S. (2015). Improved student psychology prediction & recommendation strategy using 2 state data analysis. *2015 Global Conference on Communication Technologies (GCCT),* 869–873. https://doi.org/10.1109/gcct.2015.7342786

Reivan-Ortiz, G. G., Pineda-García, G., & León Parias, B. D. (2019). Psychometric Properties of The Goldberg Anxiety and Depression Scale (GADS) In Ecuadorian Population. *International Journal of Psychological Research, 12*(1), 41–48. https://doi.org/10.21500/20112084.3745

Ruisoto, P., López-Guerra, V. M., Paladines, M. B., Vaca, S. L., & Cacho, R. (2020). Psychometric properties of the three versions of the Perceived Stress Scale in Ecuador. *Physiology & Behavior, 224*, 113045. https://doi.org/10.1016/j.physbeh.2020.113045

Saleh, D., Camart, N., & Romo, L. (2017). Predictors of Stress in College Students. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00019

Secretaría Nacional de Planificación y Desarrollo. (2016). *Plan Nacional para el Buen Vivir 2013–2017*. Accessed on April 30, 2021, from http://www.buenvivir.gob.ec/versiones-plan-nacional

Torres, C., Otero, P., Bustamante, B., Blanco, V., Díaz, O., & Vázquez, F. L. (2017). Mental Health Problems and Related Factors in Ecuadorian College Students. *International*

journal of environmental research and public health, 14(5), 530. https://doi.org/10.3390/ijerph14050530

Tyulyupo, S., Andrakhanov, A., Dashieva, B., & Tyryshkin, A. (2018). Adolescents Psychological Well-Being Estimation Based on a Data Mining Algorithm. *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT),* 475–478. https://doi.org/10.1109/stc-csit.2018.8526628

Vicent, M., Inglés, C. J., Sanmartín, R., Gonzálvez, C., Jiménez-Ayala, C. E., & García-Fernández, J. M. (2020). Psychometric properties of the child and adolescent perfectionism scale in Ecuadorian adolescents. *Journal of Affective Disorders, 272*, 176–182. https://doi.org/10.1016/j.jad.2020.04.036

Zainal-Mokhtar, K., & Mohamad-Saleh, J. (2013). An Oil Fraction Neural Sensor Developed Using Electrical Capacitance *Tomography Sensor Data. Sensors, 13*(9), 11385–11406. https://doi.org/10.3390/s130911385

APPENDIX A: AVERAGE ACCURACY OBTAINED ACROSS 100 RANDOM MLP CONFIGURATIONS

Iteration	Configuration				Average accuracy
	Number of hidden layers	Number of neurons in each hidden layer	Activation function	Learning rate	
1	2	50	identity	0.3	0.4029997
2	3	60	relu	0.5	0.3776500
3	3	70	relu	0.1	0.8288689
4	2	50	Identity	0.1	0.4000724
5	3	80	Tanh	0.3	0.5060037
6	1	30	Relu	0.5	0.5696398
7	2	50	Relu	0.3	0.7589523
8	1	60	Logistic	0.5	0.8239137
9	2	50	Tanh	0.1	0.9465255
10	1	100	Logistic	0.1	0.8167731
11	1	30	Tanh	0.1	0.8020736
12	1	80	Relu	0.1	0.7222254
13	3	20	Logistic	0.5	0.8780663
14	1	70	Logistic	0.5	0.8252368
15	1	60	Tanh	0.3	0.7864235
16	1	40	Identity	0.5	0.3840902
17	1	40	Tanh	0.1	0.7691895
18	2	90	Tanh	0.1	0.9484810
19	3	40	Identity	0.1	0.3919248
20	2	70	Logistic	0.1	0.9386994
21	2	80	Logistic	0.1	0.9449054
22	1	70	Relu	0.3	0.7222063
23	1	60	Logistic	0.1	0.8040653
24	1	20	Identity	0.1	0.3850855
25	2	90	Relu	0.5	0.6549233
26	3	20	Tanh	0.5	0.7176811
27	3	60	Relu	0.3	0.6625982
28	3	40	Logistic	0.3	0.2305135
29	1	30	Tanh	0.1	0.8020736
30	1	40	Logistic	0.3	0.8210140
31	3	50	Tanh	0.1	0.9533680
32	1	50	Identity	0.5	0.3847289
33	2	70	Relu	0.1	0.9308839
34	3	80	Logistic	0.5	0.2223542
35	2	30	Relu	0.1	0.8516468
36	1	30	Identity	0.3	0.4036448
37	1	70	Relu	0.5	0.7127068
38	1	90	Logistic	0.3	0.8128462
39	2	100	Relu	0.1	0.9197281

40	2	30	Identity	0.1	0.4072449
41	2	10	Identity	0.1	0.3941890
42	2	10	Identity	0.3	0.3997530
43	2	40	Tanh	0.1	0.9367493
44	2	50	Logistic	0.3	0.9452173
45	1	90	Relu	0.5	0.7545049
46	3	50	Tanh	0.5	0.2041291
47	2	50	Relu	0.1	0.8774137
48	2	60	Identity	0.3	0.3974548
49	1	60	Logistic	0.5	0.8239137
50	1	50	Logistic	0.5	0.8193449
51	1	80	Tanh	0.3	0.7981638
52	3	30	Relu	0.3	0.6227619
53	3	60	Relu	0.3	0.6625982
54	3	70	Relu	0.5	0.4209576
55	1	50	Logistic	0.1	0.8229631
56	3	50	Identity	0.1	0.3925805
57	3	80	Logistic	0.3	0.2272562
58	3	40	Logistic	0.3	0.2305135
59	2	10	Tanh	0.1	0.7065487
60	3	40	Logistic	0.3	0.2305135
61	1	30	Logistic	0.5	0.7574408
62	1	40	Identity	0.3	0.3915895
63	1	80	Relu	0.5	0.9449011
64	2	50	Tanh	0.3	0.7896936
65	1	40	Tanh	0.3	0.7896936
66	1	40	Relu	0.1	0.6853633
67	2	10	Logistic	0.5	0.7039461
68	2	10	Tanh	0.5	0.6306189
69	2	100	Tanh	0.1	0.9510858
70	2	70	Logistic	0.5	0.9373954
71	1	90	Tanh	0.1	0.8092951
72	2	50	Tanh	0.1	0.9465255
73	2	90	Relu	0.5	0.6549233
74	1	100	Relu	0.1	0.7469928
75	1	10	Relu	0.1	0.4717996
76	1	10	Tanh	0.3	0.5634083
77	2	10	Tanh	0.5	0.6306189
78	1	20	Tanh	0.3	0.7124343
79	3	60	Tanh	0.3	0.9549978
80	1	10	Logistic	0.1	0.5862479
81	3	60	Tanh	0.5	0.1809499
82	2	30	Tanh	0.3	0.9109898
83	3	40	Tanh	0.3	0.9367354
84	2	60	Tanh	0.1	0.9465276
85	1	50	Relu	0.1	0.6645292
86	3	60	Tanh	0.1	0.9533787

87	1	70	Tanh	0.3	0.8030540
88	2	90	Tanh	0.3	0.9481606
89	3	80	Relu	0.5	0.4889581
90	1	40	Tanh	0.1	0.8340572
91	2	10	Relu	0.5	0.3472717
92	1	90	Relu	0.5	0.7545049
93	3	30	Tanh	0.1	0.9364203
94	3	20	Relu	0.5	0.3041291
95	1	30	Relu	0.1	0.6142726
96	1	100	Tanh	0.5	0.7333152
97	1	50	Identity	0.3	0.3902929
98	1	60	Tanh	0.1	0.8457655
99	1	40	Tanh	0.3	0.7896936
100	1	90	Relu	0.5	0.7545049

APPENDIX B: AVERAGE ACCURACY OBTAINED ACROSS EVERY PERMUTATION OF THE HYPERPARAMETERS FOR THE RANDOM FOREST CLASSIFIERS

Number of random features selected	Number of decision trees in the Random Forest	Average Accuracy
1	100	0.9445712
1	200	0.9475057
1	300	0.9494620
1	400	0.9484839
1	500	0.9491360
1	600	0.9491360
1	700	0.9488099
1	800	0.9488099
1	900	0.9504402
1	1000	0.9497881
2	100	0.9426149
2	200	0.9448973
2	300	0.9471797
2	400	0.9468536
2	500	0.9462015
2	600	0.9462015
2	700	0.9471797
2	800	0.9484839
2	900	0.9475057
2	1000	0.9478318
3	100	0.9432670
3	200	0.9416368
3	300	0.9445712

3	400	0.9458754
3	500	0.9452233
3	600	0.9455494
3	700	0.9458754
3	800	0.9452233
3	900	0.9455494
3	1000	0.9448973
4	100	0.9377242
4	200	0.9400065
4	300	0.9409847
4	400	0.9429410
4	500	0.9442452
4	600	0.9432670
4	700	0.9426149
4	800	0.9426149
4	900	0.9419628
4	1000	0.9413107
5	100	0.9347897
5	200	0.9383763
5	300	0.9373981
5	400	0.9377242
5	500	0.9380502
5	600	0.9367460
5	700	0.9367460
5	800	0.9380502
5	900	0.9377242
5	1000	0.9380502
6	100	0.9325073
6	200	0.9338115
6	300	0.9354418
6	400	0.9357679
6	500	0.9347897
6	600	0.9354418
6	700	0.9357679
6	800	0.9357679
6	900	0.9360939
6	1000	0.9360939
7	100	0.9305510
7	200	0.9308771
7	300	0.9312031
7	400	0.9328334
7	500	0.9325073
7	600	0.9315292
7	700	0.9321813
7	800	0.9308771
7	900	0.9315292
7	1000	0.9321813

8	100	0.9282687
8	200	0.9289208
8	300	0.9285947
8	400	0.9279426
8	500	0.9282687
8	600	0.9282687
8	700	0.9279426
8	800	0.9289208
8	900	0.9282687
8	1000	0.9276166
9	100	0.9253342
9	200	0.9263124
9	300	0.9266384
9	400	0.9256603
9	500	0.9282687
9	600	0.9272905
9	700	0.9276166
9	800	0.9272905
9	900	0.9259863
9	1000	0.9259863