Contents

[**1 Part 1** 2](#_Toc97502864)

[**1.1** **Dataset Summary** 2](#_Toc97502865)

[**1.2** **Principal Component Analysis (PCA)** 2](#_Toc97502866)

[**2** **Part 2** 6](#_Toc97502867)

[**2.1** **Dataset Summary and Pre-processing** 6](#_Toc97502868)

[**2.2** **Model Development** 6](#_Toc97502869)

[**2.2.1 Multiple Linear Regression with all Input Features** 6](#_Toc97502870)

[**2.2.2 Multiple Linear Regression with Significant Features** 6](#_Toc97502871)

[**2.2.3 Decision Tree** 7](#_Toc97502872)

[**2.2.4 XGBoost** 7](#_Toc97502873)

[**2.3** **Model Evaluation** 7](#_Toc97502874)

[**3** **Part 3** 8](#_Toc97502875)

[**3.1** **Dataset Summary and Pre-processing** 8](#_Toc97502876)

[**3.2** **Model Development** 8](#_Toc97502877)

[**3.2.1 Logistic Regression with all Input Features** 8](#_Toc97502878)

[**3.2.2 Logistic Regression with Significant Features** 8](#_Toc97502879)

[**3.2.3 Decision Tree** 9](#_Toc97502880)

[**3.2.4 XGBoost** 9](#_Toc97502881)

[**3.3** **Model Evaluation** 9](#_Toc97502882)

[**References** 11](#_Toc97502883)

# **1 Part 1**

## **Dataset Summary**

The EWCS (European Working Conditions Survey 2016) dataset contains 7647 rows and 11 columns. The data types of all the columns are loaded as integers. Q2a (Gender) can be converted to factor data type later as it has only 2 unique values (1 and 2) representing male and female respectively. Q2b represents the age. Q87 series (Q87a, Q87b, Q87c, Q87d, Q87e) contain the values to the personal feelings of the individuals over the last two weeks. Q90 series (Q90a, Q90b, Q90c, Q90f) contain the values to the feelings of the individuals regarding their jobs. Refer to Table 1 for some examples of the values for each column.

Table 1 Internal structure of EWCS dataset

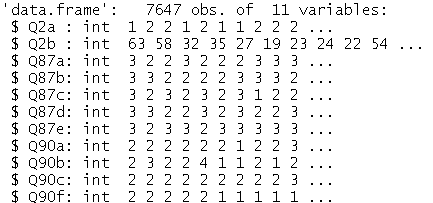
****

Figure 1 shows a summary of the distribution of the variables. The values for Q2a are not very useful for now as it is not converted to factor data type yet. However, the mean of 1.49 indicates the gender distribution is quite balanced as it roughly sits in the middle of 1 (male) and 2 (female). All survey questions have a median of 2 (Most of the time) except for the last question (Q90f) with a median of 1 (Always).

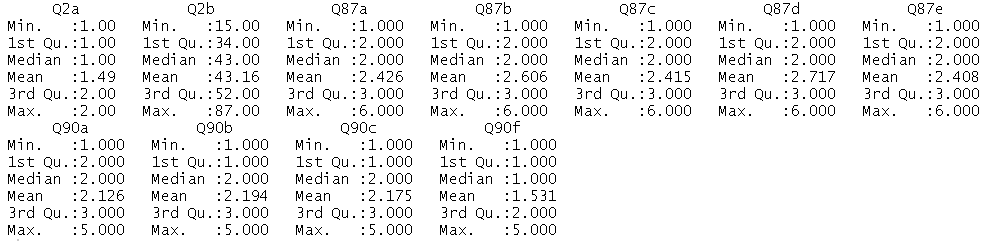
****

Figure 1 Summary of Variables

## **Principal Component Analysis (PCA)**

Principal component analysis (PCA) will be used to summarize the information in the EWCS data. PCA is an unsupervised technique used mainly to reduce the dimensionality of a dataset while preserving the maximum amount of variance [1]. It produces linear combinations of the initial variables to form the principal components [2]. Normality assumptions of the variables are not required here as we are using PCA for descriptive purposes as opposed to inferential purposes. Columns 3-11 representing the survey values will be used for the PCA. The columns will be centered and scaled before calculating the principal components. This simplifies the computation and reduces any bias due to differences in scale. If one component varies less than another because of their respective scales (1000s vs. 10s range), PCA might determine that the direction of maximal variance more closely corresponds with the larger range (1000s) axis.

Table 2 shows the importance of each principal component. A total of 9 principal components (PC1-PC9) are calculated by the prcomp function in R. The standard deviation is equivalent to the eigenvalue after centering and scaling. The eigenvalues are the coefficients of the respective eigenvectors. Ranking the eigenvectors from highest to lowest eigenvalues produces the principal components in order of significance. PC1 is the direction representing the largest variation, followed by PC2, and so on.

Table 2 Importance of each principal component

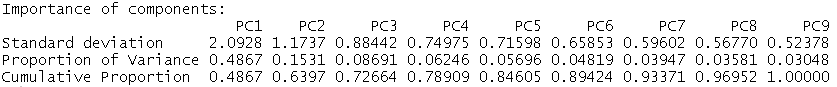


Figure 2 shows a scree plot that allows us to visually inspect the number of principal components to retain to reduce the dimensionality of the dataset. This plot corresponds to the values given in the second row of Table 2. The proportion of variance refers to the amount of variance each component accounts for in the data (e.g. PC1 accounts for 48.67% of the total variance in the data, PC2 accounts for 15.31% of the total variance in the data, and so on). The scree plot shows us where the variance explained value starts to level off. From the plot, we can determine that the first 3 principal components explain more than 70% of the variance in the data and it can be a good starting point to use these for further analysis or modeling.

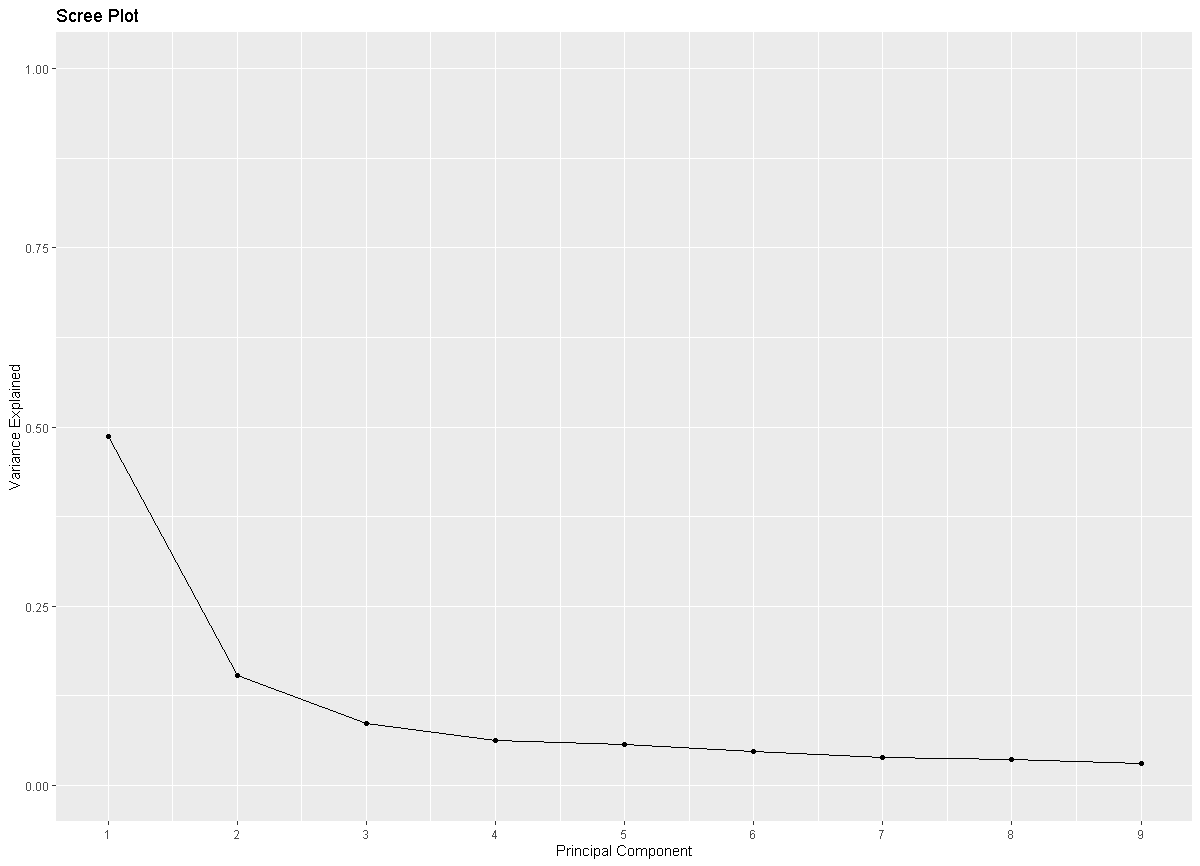


Figure 2 Scree plot showing the variance explained by each principal component

Figure 3 shows the cumulative variance plot where we can visualize the total percentage of variance explained for different cut-offs of principal component numbers. For example, if we are interested in explaining at least 80% of the variance using the minimum number of principal components, we can choose the first 5 principal components as shown in the plot. These PCs explain 84.605% of the total variance in the dataset.

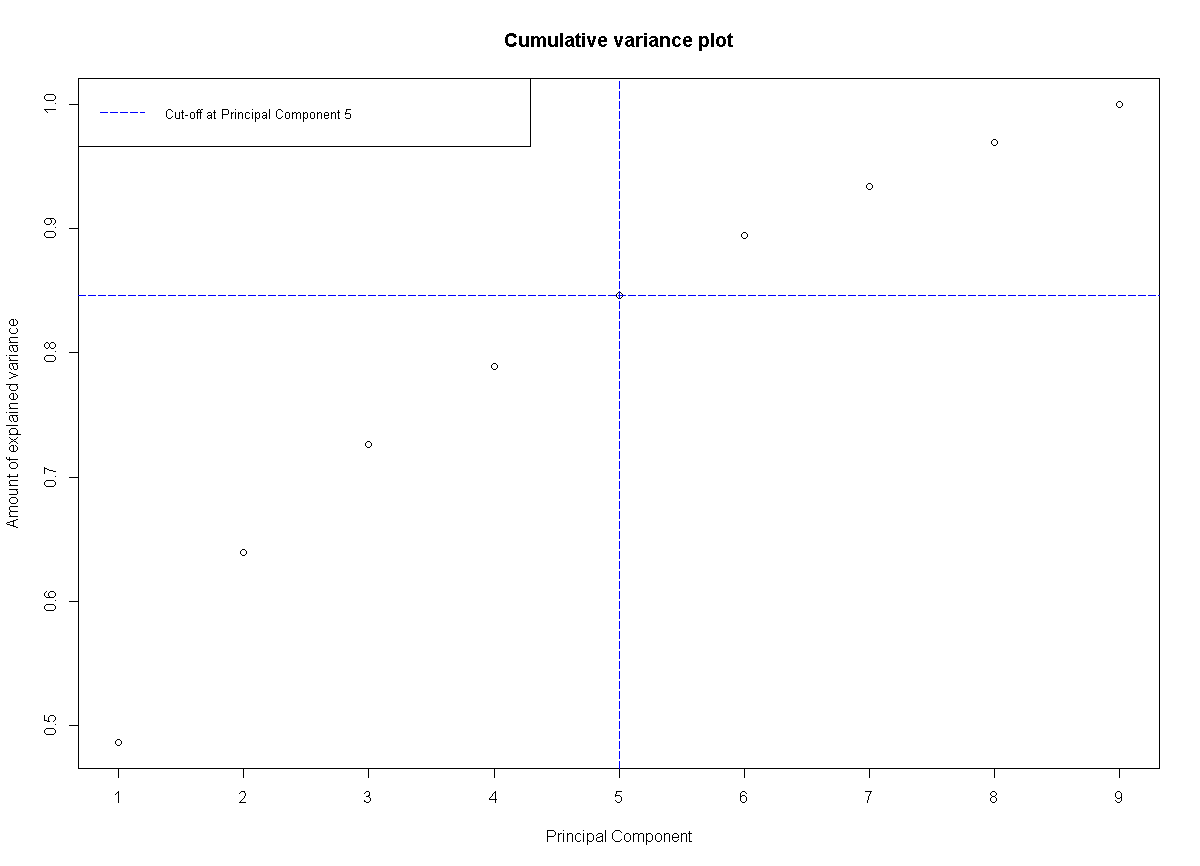


Figure 3 Cumulative variance plot showing the total percentage of variance explained vs. the number of principal components

Q2a (Gender) variable is changed to factor type for categorical analysis. From Figure 4, it can be seen that other than the female survey answers being slightly more spread out, there is no visible difference between males and females.

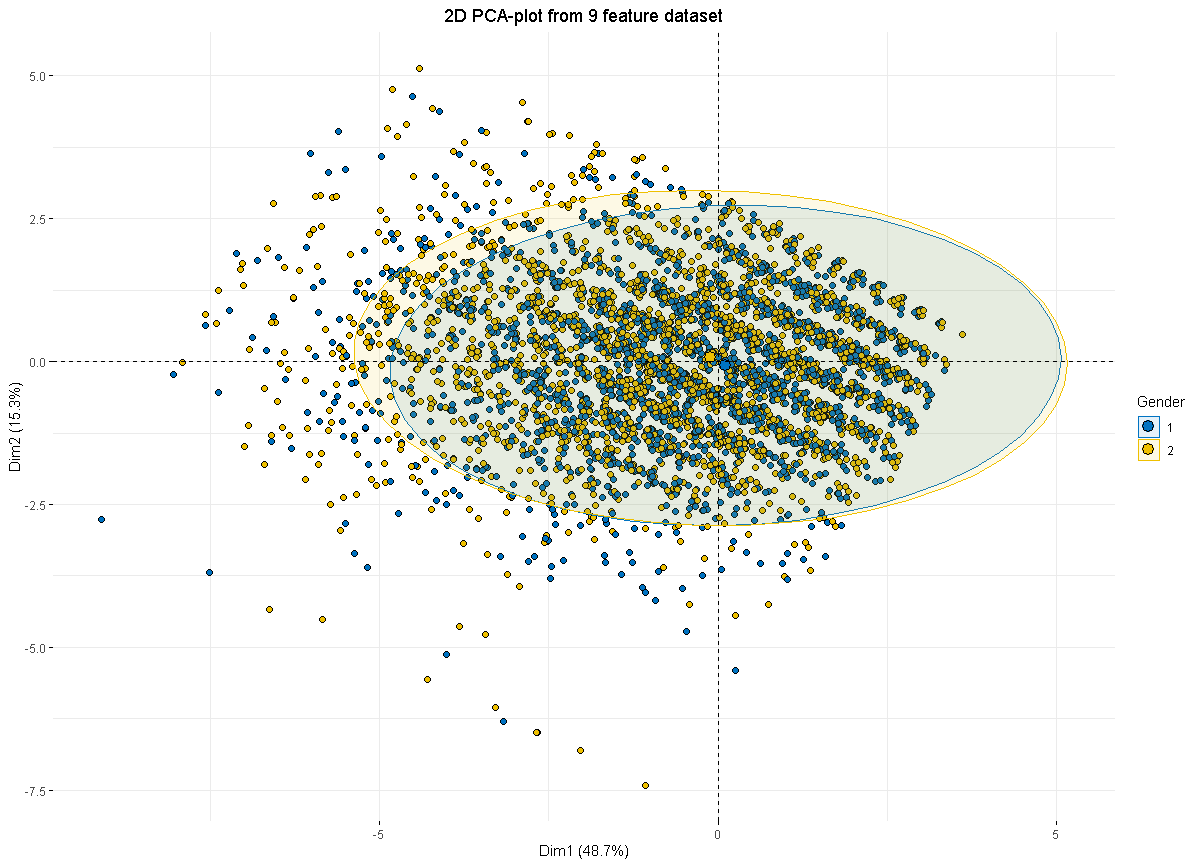


Figure 4 2D PCA-plot from 9 feature dataset

Figure 5 shows the PCA biplot for PC1 and PC2 with the loadings of the standardized original variables. It can be observed that the Q87 series have higher loadings in PC1 while the Q90 series have higher loadings in PC2.

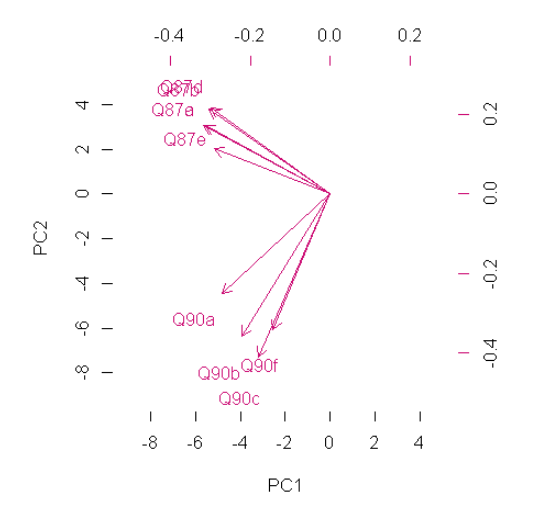


Figure 5 PCA biplot

# **Part 2**

## **Dataset Summary and Pre-processing**

The student performance datasets are available in two distinct subjects: Mathematics and Portuguese language. The first period and second period grades (G1/G2) are removed from the datasets as the ability to predict the final grade (G3) using alternative data can be much more useful [3].

## **Model Development**

### **2.2.1 Multiple Linear Regression with all Input Features**

The baseline model used for both datasets is a multiple linear regression model with all input features fed to predict the continuous output value (G3) with a possible range from 0-20. A random seed is set for the train/test split to ensure that the randomization works the same on different machines and the results are reproducible [4]. The default 70/30 train/test split is used where 70% of the data are used for training and 30% of the data are used for testing. The root mean square error (RMSE) [5] would be used to assess the predictive performance of the models.

### **2.2.2 Multiple Linear Regression with Significant Features**

Based on the p-values [6] of the input features from the previous models, a subset of the features are chosen to be fed to the new multiple linear regression models. The selection criteria were based on a 5% significance threshold of the p-values. Table 3 shows the significant features of both datasets. It can be observed that the “failures” feature is present in both datasets and is probably an important feature for prediction purposes.

Table Significant features of both datasets

|  |  |  |
| --- | --- | --- |
|  | **Mathematics Dataset** | **Portuguese Dataset** |
| **Significant Features** | failures, romanticyes | schoolMS, sexM, failures, schoolsupyes, higheryes, famrel |

### **2.2.3 Decision Tree**

A decision tree can be more flexible than linear regression models due to its ability to model non-linear information [7]. The structure of the tree closely resembles a bunch of if-else statements and is thus highly interpretable. However, it tends to overfit if unpruned as it can keep splitting the data till every node contains only 1 sample. All the features were fed into the model as it has an internal feature selection mechanism through the split structure. If a feature is not useful in increasing the purity of the data, it would not be chosen as a split point. This naturally eliminates bad features without the additional need to perform feature selection beforehand [8].

Pruned trees based on the optimal complexity parameters (CP) are also tried out. The CP is used to control and select the optimal decision tree size. The core idea is that the cost of the tree will increase if the number of terminal nodes increases, aka additional splits. If the next split does not increase the purity of the child nodes enough to offset the complexity cost, the tree will terminate the growing process at the parent node [9].

### **2.2.4 XGBoost**

XGBoost stands for eXtreme Gradient Boosting which is an ensemble of decision trees that are built sequentially. A tree is built to reduce the errors of the previous tree [10]. Each individual tree is similar to a decision tree. However, XGBoost offers higher scalability and performance with optimized computation algorithms. It also has regularization hyperparameters to prevent overfitting with so many trees.

## **Model Evaluation**

Table 4 shows the model results based on the root mean square error for both datasets. It can be observed that the mathematics results are much harder to predict accurately relative to the Portuguese results. This is consistent across all 5 models. The XGBoost model has the lowest RMSE on both datasets. An interesting observation is that the linear regression models might not necessarily perform better with using just the significant features. Predictive performance is tied more closely to the association power between the input and output features rather than the significance level of the input features. Given that the size of the datasets are rather small (less than 1000 rows), the predictive performance has room to improve through the collection of more data and ensuring the quality of the data is robust.

Table Model results based on RMSE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Multiple Linear Regression with all Input Features** | **Multiple Linear Regression with Significant Features** | **Decision Tree** | **Pruned Decision Tree** | **XGBoost** |
| **Mathematics Dataset** | 3.9742 | 4.2926 | 3.8887 | 3.7850 | 3.6388 |
| **Portuguese Dataset** | 2.7717 | 2.8148 | 3.1503 | 2.8862 | 2.7432 |

# **Part 3**

## **Dataset Summary and Pre-processing**

The bank marketing dataset is used for this part. The “y” label is converted from factor to numeric data type with the following mapping: "yes": 1 and "no": 0. This helps to facilitate the prediction process later on as the machine learning models can only work with numbers.

## **Model Development**

### **3.2.1 Logistic Regression with all Input Features**

The baseline model used for the dataset is a logistic regression model with all input features fed to predict the binary output (y) with a possible value of either 1 or 0. Similar procedures to the regression task such as random seed and 70/30 train/test split are deployed here as well. The accuracy metric [11] would be used to assess the predictive performance of the models. Another useful metric from the model summary is the Akaike information criterion (AIC). The AIC values are used in binary classification tasks as the R-square is only available for regression tasks. The AIC value is not meaningful by itself. It has to be compared with different logistic regression models to make sense. The AIC value is 1629.1 for this model with all features used.

### **3.2.2 Logistic Regression with Significant Features**

Based on the p-values of the input features from the previous model, a subset of the features are chosen to be fed to the new logistic regression model. The selection criteria were based on a 5% significance threshold of the p-values. Table 5 shows the significant features. Figure 6 shows the summary of this model. The AIC value is 1625.3 for this model which is slightly worse when compared to the model with all features used. The “month” feature appears a lot of times in the summary table. It records the last contact month of the year for the client with the bank. It makes sense as a non-active client has less probability of subscribing to a term deposit.

Table 5 Significant features of both datasets

|  |  |
| --- | --- |
|  | **Bank Marketing Dataset** |
| **Significant Features** | loan, contact, month, duration, campaign, poutcome |

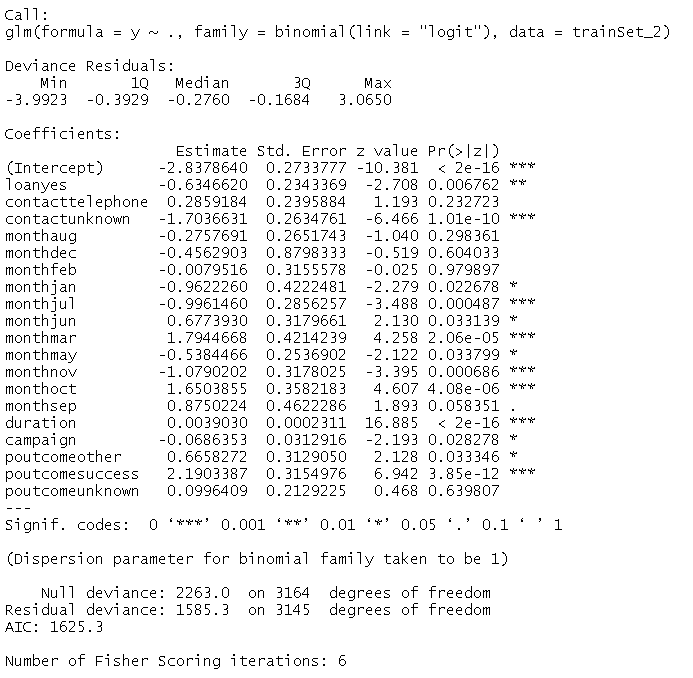
****

Figure Summary of logistic regression model

### **3.2.3 Decision Tree**

Similar steps are done here as with part 2. The only difference is the method used in rpart is “class” instead of “anova” to accommodate for the binary classification task.

### **3.2.4 XGBoost**

Similar steps are done here as with part 2. The only difference is the method used in xgboost is “binary:logistic” instead of the default value to accommodate for the binary classification task.

## **Model Evaluation**

Table 6 shows the model results based on accuracy. It can be observed that the results are rather similar despite the different models used or processing techniques such as feature selection and pruning. The decision tree model has the highest accuracy, albeit only by a slight margin. An interesting observation is that the logistic regression models might not necessarily perform better with using just the significant features. Predictive performance is tied more closely to the association power between the input and output features rather than the significance level of the input features.

Table 6 Model results based on RMSE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Multiple Linear Regression with all Input Features** | **Multiple Linear Regression with Significant Features** | **Decision Tree** | **Pruned Decision Tree** | **XGBoost** |
| **Accuracy** | 0.9004 | 0.8909 | 0.9041 | 0.9034 | 0.9019 |

# **References**

[1] <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>

[2] <https://strata.uga.edu/software/pdf/pcaTutorial.pdf>

[3] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[4] <http://rfunction.com/archives/62>

[5] <https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/>

[6] <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/>

[7] <https://pdf.co/blog/decision-trees-in-machine-learning>

[8] <https://www.displayr.com/how-is-splitting-decided-for-decision-trees/>

[9] https://quick-adviser.com/what-is-the-optimal-value-of-the-complexity-parameter-cp/

[10] <https://xgboost.readthedocs.io/en/stable/R-package/xgboostPresentation.html>

[11] https://www.cuemath.com/accuracy-formula/