Practical Work II - Article

This article has been carried out for the ANADI Curricular Unit

1st Mariana Lages
Instituto Politécnico do Porto
ISEP - DEI
Porto, Portugal
1200902@isep.ipp.pt

2nd Francisco Redol
Instituto Politécnico do Porto
ISEP - DEI
Porto, Portugal
1201239@isep.ipp.pt

3rd Miguel Jordão
Instituto Politécnico do Porto
ISEP - DEI
Porto, Portugal
1201487@isep.ipp.pt

***Abstract*—The main objective of this report is to compare different statistical tools through the analysis of the performance of a cyclist group. Its conclusions derive from a data set consisting of 11 columns over 1000 rows. These columns contain intel about the athletes' ID, Gender, Team, Background, Pro level, Winter Training Camp completion, Date of Birth and Altitude, VO2 and Heart rate results. As means to improve comprehension and pattern discovery, R language features such as the ability to plot linear regression graphics or build complex Decisions Trees, Neural Networks, and KNN Algorithms allowed for a comparison of the different techniques using several crucial metrics, such as Accuracy, Precision, Sensitivity, Specificity and F1 Score.**

***Index Terms*—linear regression, multi-linear regression, decision trees, neural networks, k-nearest neighbors, hypothesis tests, data analysis, data modeling, machine learning, R language**

I. Introduction

Exploratory analysis is elected as the first step in a data analysis process, providing insights into a data set's characteristics. By exploring and summarizing data through numerous statistical and graphical procedures, researchers can identify patterns, anomalies, and potential outliers, which can guide further analysis and hypothesis testing, along with formulating theories around a certain subject.

In this research, we will discuss key statistical aspects of that character, such as exploratory analysis, machine learning models, and data modeling based on the provided data sets. By giving an insight into these techniques and their applications, we aim to emphasize the importance of these backbone concepts while doing scientific research.

The main motivation of this practical work is to apply machine learning algorithms in data exploration and comparison using appropriate statistical tests. The goal is to analyze a data set collected from professional cyclists during their preseason training period. The objective is to validate potential relationships between the collected data and measurements related to various performance indicators of the athletes.

The specific objectives of this work are as follows:

- Conduct a state-of-the-art review of different machine learning algorithms applicable to the analysis of the provided data set.
- Develop models using classification/regression techniques to analyze the data on cyclists' preparation for the upcoming competition season.
- Obtain results from the developed models and perform an in-depth analysis of the findings.
- Compare the performance of different machine learning algorithms used in the study.
- Conclude and provide a synthesis of the overall findings.

The methodology followed in this work involves the following steps:

- **Data Collection:** A data set consisting of structured data related to the preparation of cyclists for a new competition season after their winter training has been obtained. The data set comprises 1000 observations and includes 11 variables such as ID, Gender, Team, Background, Pro_Level, Winter Training Camp, Altitude_results, Vo2_results, Hr_results, Date of birth, and Continent.
- **Literature Review:** A comprehensive review of the state-of-the-art machine learning algorithms relevant to the analysis of the data set is conducted. This step helps in identifying the most suitable algorithms for the given task.
- **Model Development:** The selected machine learning algorithms, including linear regression, multi-linear regression, decision trees, k-nearest neighbors, and neural networks, are applied to develop classification/regression models based on the data set.
- **Model Evaluation:** The developed models are evaluated using appropriate performance metrics and statistical tests. The results are analyzed to gain insights into the relationships between the collected data and performance metrics.
- **Comparison and Discussion:** A comparative analysis is performed to assess the performance of different machine learning algorithms. The strengths and limitations of each algorithm are discussed based on the obtained results.
- **Conclusion:** The findings from the analysis are summarized, and conclusions are drawn regarding the identified relationships between the collected data and the performance indicators of the cyclists.

By following this methodology, the objective is to gain a deeper understanding of the data set and provide insights into the factors influencing the performance of professional cyclists

during their training and preparation for competitions.

II. Literature Review

In order to comprehensively comprehend the application of machine learning models in the context of cycling, specifically in the pre-season phase, a literature review was conducted to analyze performance analysis and the specific models utilized in this study. This section presents a concise overview of the significant findings from prior research and studies pertaining to the analysis of data collected during the pre-season period of cyclists. By examining the current state of the field, we can identify any existing gaps, challenges, and opportunities for harnessing machine learning techniques to analyze the data set encompassing cyclist preparation.

A. Machine Learning

Machine learning has emerged as a transformative tool across various domains, revolutionizing the analysis and processing of data. It encompasses a wide range of algorithms and techniques that enable computers to learn from data without explicit programming. In recent years, machine learning has experienced remarkable advancements, driven by the availability of large data sets and increased computational power. These advancements have facilitated the development of sophisticated models capable of extracting meaningful patterns and making accurate predictions. From conventional approaches like regression and decision trees to more advanced methods like neural networks, machine learning techniques offer immense potential for tackling complex problems and uncovering valuable insights from data that would otherwise be challenging to analyze.

B. Linear Regression and Multi-linear Regression

Linear regression is a widely-used statistical model for predicting a continuous outcome variable based on one or more independent variables. It assumes a linear relationship between the predictors and the response variable. In the context of cyclist preparation, linear regression can be used to estimate the impact of individual variables, such as altitude results, VO2 results, and heart rate results, on performance metrics. Multi-linear regression extends this concept by allowing for the inclusion of multiple predictors simultaneously, enabling a more comprehensive analysis of the factors influencing performance.

C. Decision Trees

Decision trees are tree-based models that partition the data based on a series of hierarchical decisions. Each internal node represents a test on a specific attribute, while each leaf node represents a class label or a prediction. Decision trees are interpretable and easy to understand, making them valuable for gaining insights into the most influential variables in the cyclist preparation data set. Additionally, decision trees can handle both categorical and numerical data, making them suitable for analyzing the diverse range of variables in the data set.

D. Neural Networks

Neural networks are a class of machine learning algorithms inspired by the structure and function of biological neural networks. These models consist of interconnected nodes, or artificial neurons, organized into layers. Neural networks can capture complex nonlinear relationships and are particularly effective in tasks such as classification and regression. By training a neural network on the cyclist preparation data set, it is possible to uncover intricate patterns and dependencies between the input variables and performance metrics.

E. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric algorithm that classifies new data points based on the majority vote of their k-nearest neighbors in the training data set. KNN is particularly suitable for classification tasks and has been successfully applied in various domains. By identifying the most similar instances to a given data point, KNN can provide valuable insights into the relationships between different variables in the cyclist preparation data set.

F. Cross Validation

Machine learning models require rigorous evaluation to ensure their reliability and effectiveness in making accurate predictions. One widely used technique for model evaluation is cross-validation. Cross-validation allows us to estimate a model's performance by assessing its predictive capabilities on different subsets of the data. For this test study it will only be analysed the K-fold Cross-Validation.

K-fold Cross-Validation is a robust method for estimating model accuracy. It divides the data set into k subsets or folds. The model is trained on k-1 folds and tested on the remaining fold, iteratively repeating this process until each fold has served as the test set. The prediction errors from each iteration are averaged to obtain the cross-validation error, which serves as a performance metric for the model. By evaluating the model on different subsets of the data, k-fold cross-validation provides a more accurate estimate of the test error rate compared to other methods. Choosing the appropriate value of k is crucial in k-fold cross-validation. While there is no definitive rule, commonly used values include 5 and 10. The selection of k depends on factors such as data set size, computational resources, and the desired level of precision in estimating model performance.

III. Pre-Processing of the data set

For our statistical methods to be applied, we needed to make our data set ready to be explored. Reuniting the requirements of our different models and analysis methods, we could verify that the ID did not have emphasis for any examination, which made us import the data set without it. Continuing the analysis, we could see that attributes such as Gender and Continent were suitable candidates for label and one-hot encoding, respectively. Not only that, but the date of birth did not pose immediate information for our models to dissect, which made us calculate the age and append it to the data set

(after removing that date from the data set). However, before any following step, it was crucial to verify beforehand if any row had missing information. Fortunately, the data set had no information absent, which allowed us to move to the next phase.

Post-Processing left us with a data set in which binary attributes such as male or female (on the Gender column) are represented by ones and zeros and, for variables with more than two feasible options, one-hot encoding created columns to allow binary representation, making the data ready to be utilized on the next steps. Not only that, a Min-Max Normalization technique was applied, allowing for data scaling, increased interpretability and model performance for the following operations.

Ultimately, these columns were the ones that were accessible: Gender, Pro.level, Winter.Training.Camp, TeamGroupA, TeamGroupB, TeamGroupC, TeamGroupD, TeamGroupE, BackgroundCobblestones, BackgroundHill, BackgroundMountain, BackgroundNone, BackgroundSprinter, BackgroundTimeTrial, ContinentAfrica, ContinentAsia, ContinentAustralia, ContinentEurope, ContinentNorthAmerica, ContinentSouthAmerica, altitude_results, vo2_results, hr_results and age.

IV. Analysis of the data set

Ensuring a relevant analysis is a process that depends on a coherent selection of data fields. That leads to an exquisite evaluation of the available information. A set of plots created to aid in that process allowed us to extract pointer information about our data, making it easier to get conclusions later.

First, numeric data fields (altitude, VO2, heart rate, and age) were elected to generate a box plot. Visual clues allow us to infer that six outliers are present in our data set. However, these represent 0.6% (6 out of 1000 rows) of our data, meaning they are negligible on the scope of the data set.

Afterward, a set of scatter plots illustrating relations between numeric attributes originated, allowing for comprehension of how these values evolve. The altitude vs VO2 plot (Figure 9) demonstrated how an increase in elevation proportionally augmented VO2 values in athletes, explained by reduced oxygen availability and a rise in red blood cell production. The findings from the previous study receive confirmation from the altitude vs heart rate plot (Figure 10), which represents an increase in heart rate according to a height increase, explained by the reduced oxygen concentration. Lastly, an increased VO2 leads to a higher heart rate since there is a greater oxygen demand, and improved efficiency of its usage, leading to increased cardiac output. All are demonstrated by the VO2 vs heart rate plot (Figure 11).

However, as the focus point of the article is the comparison between regression and classification algorithms for different attributes, it is crucial to select the most influential data fields since an overflow of data can lead to an incorrectly taught model or a lack of performance, for instance. Beforehand, it's important to note that correlation does not imply causation. The following findings are based solely on the correlation matrix, and further analysis or experimentation would be needed to draw more definitive conclusions. Upon inspecting the graphical depiction, it's possible to verify a weak negative correlation (-0.0907) between the professional level and gender, suggesting a slight difference in professional levels between genders, with a tendency for higher professional levels in males. Winter Training Camp does not reveal strong correlations with other variables in the data set. Its correlation coefficients with other variables are relatively low and close to zero. None of the Team Group variables shows significant correlations with other variables in the data set. The correlation coefficients are generally close to zero, indicating a weak relationship. Background variables have relatively low correlations with other variables. These variables do not strongly correlate with performance-related variables, suggesting that the background does not significantly influence performance. The Continent variables show weak correlations with other variables. These variables do not appear to have a solid bond with performance or other measured attributes. Altitude Results, VO2 Results, and HR Results show moderate correlations, indicating some interdependence. These variables also show weak correlations with age, suggesting it may have a limited influence on these performance-related measures (Figure 12).

V. Performance Analysis Of Learning Techniques

A. Regression

In this study, we adopted a statistical approach to develop and evaluate simple and multiple linear regression models. The objective was to predict the dependent variable "altitude_results" based on the independent variable "hr_results".

To start this analysis, we adopted the good practice of creating a data sample that divides the normalized data set into two sets: one for training and one for testing. For this, the data was divided in a **70/30 ratio**, in which 70% of the records were inserted in the training set and the remaining 30% in the test set.

Subsequently, we performed a simple linear regression model, utilizing "hr_results" as the dependent variable and "altitude_results" as the independent variable.

The resulting equation for the model was as follows:

$$altitude\_results = 0.02919 + 0.86157 * hr\_results \quad (1)$$

In addition, we created a scatter plot to visualize the regression line and the relationship between the variables in question.

The graph obtained is present in the Figure 5.

By analyzing the regression line and the respective scatter plot, it is possible to identify the direction of the relationship by the positive growth trend of the regression line. We can also assess the strength of the relationship by checking the proximity of the points to the line, which are mostly close.

Another factor to consider is the positive slope of the regression line, which indicates that there is a general increase

in the variables as one increases about the other. We also observed that the dispersion of the points is mostly balanced on both sides of the line, so we can say a good fit of the model.

This analysis allows us to infer a positive relationship between the variables, reinforcing the reliability of the results obtained and the validity of the regression model used. Through the regression line and the scatter plot, it is possible to clearly understand the direction, strength, and fit of the model, contributing to an accurate interpretation of the collected data.

The next step consisted of calculating the **Mean Absolute Error (MAE)** and the **Root Mean Squared Error (RMSE)** of the model over the 30% of the test cases. For this, the *predict()* function was used to predict the values based on the test data. After obtaining the predictions, the difference between the actual values and the predicted values was calculated, which was stored in the "dif" variable. It is through this variable that MAE and RMSE values are obtained, which are used to assess the quality of forecasts.

The MAE is calculated as the mean of the absolute differences between the actual and predicted values, while the RMSE is calculated as the square root of the mean squared differences between the actual and predicted values. The values obtained were **0.0905** and **0.1089**, respectively, which can be considered as reduced values, indicating a better quality of forecasts. These values represent a smaller discrepancy between the actual and predicted values.

These results demonstrate the good quality of the regression model used in predicting the test values. Obtaining a low MAE and RMSE reinforces the reliability of the predictions made, contributing to the validation and usefulness of the model in the context of the study in question.

However, it was decided to test the possibility of obtaining more accurate results using a more complex model, such as a **Multiple Linear Regression** model. In this sense, it was decided to include the variable "vo2_results" as a new independent variable in the model.

By considering multiple independent variables, the **Multiple Linear Regression** model is able to more fully capture and quantify the interactions and joint effects of these variables on the dependent variable.

Thus, when performing this comparative analysis, it will be possible to assess whether the multiple linear regression model represents a more appropriate and accurate approach to the problem in question, contributing to the advancement of scientific knowledge in the study area.

We used the training data set to build the model, resulting in the following equation:

$$altitude\_results = 0.01565 + 0.28762 * hr\_results + 0.59886 * vo2\_results \quad (2)$$

To assess whether the more complex model presents better results, we calculated the MAE and RMSE and compared the values obtained with those from the simple linear regression. For this, we follow the same reasoning used previously in simple linear regression.

The MAE and RMSE results obtained for the multiple linear regression model were **0.0867** and **0.1075**, respectively.

These values indicate a reduction in MAE and RMSE compared to the simple linear regression model. Therefore, we can conclude that the more complex model, which includes the "vo2_results" variable, presents a better prediction quality compared to the simple linear regression model.

The next objective was to predict the attribute "vo2_results" through the application of three different models: multiple linear regression, decision tree, and neural network.

For all models, we consider "vo2_results" as the dependent variable and "altitude_results" and "hr_results" as independent variables. We calculated the MAE and RMSE to allow future comparisons of results between the different models.

In the case of the **multiple linear regression model**, we followed the same approach used previously and obtained the following equation:

$$vo2\_results = 0.01845 + 0.14254 * altitude\_results + 0.83559 * hr\_results \quad (3)$$

As for the **decision tree model**, we used the *rpart()* function with the training data set and generated a visual graph using the *rpart.plot()* function. The result is seen in the Figure 6.

When analyzing the obtained decision tree from Figure 6, we can observe that its depth is relatively low, which indicates that the applied model is not very complex.

It is evident that the **"hr_results"** variable plays a crucial role in the decisions taken by the tree. This suggests that heart rate is a determining factor in predicting the attribute under study.

When examining the leaves of the tree, it is noticeable that most of the values are concentrated between **0.6325 and 0.7771**, representing **29.70%** of the total observations (n=210). On the other hand, there is a smaller concentration of values below 0.3313, corresponding to only 5.94% of the observations (n=42).

With regard to the **neural network**, three models were created with different parameters, which were based on those used in the theoretical-practical classes. All models were performed using the *neuralnet()* function and variations were made in the parameters, having created a neural network with **1 internal node**, with **3 internal nodes** and also with **2 internal levels**, where the first level had **6 internal nodes** and the second had **2 internal nodes**.

The *plot()* function was used to generate the graphs of each network and these were the results obtained:

Next, the MAE and RMSE values obtained for each of the developed models were compared. In the case of the neural network, only the model with an internal node was considered,

since this is the most simplified model among those tested. The following Table I presents the results obtained:

TABLE I
MAE AND RMSE VALUES OF EACH MODEL PERFORMED

Model	MAE	RMSE
Multiple Linear Regression	0.0431	0.0536
Decision Tree	0.0505	0.0643
Neural Network (1 internal node)	0.0438	0.0544

When analyzing the Table I, we observe that the **Multiple Linear Regression** model presents the lowest value of MAE and RMSE, indicating better performance in predicting the results. On the other hand, the decision tree model demonstrates the highest value of MAE and RMSE, suggesting a lower accuracy in making predictions.

Finally, a hypothesis test was carried out to verify whether the results obtained in the two best models are statistically significant.

Based on the results presented in the previous Table I, it was concluded that the two best models are the **Multiple Linear Regression and the Neural Network with 1 internal node**.

Then, a **Student's t-test** was performed with the following hypotheses:

- H0: The results obtained in both models are statistically significant
- H1: The results obtained in the two models are not statistically significant

After performing the test, a p-value of **0.9324** was obtained. Considering a significance level of 5%, it is possible to verify that the p-value is **greater than alpha (0.05)**, so there is **not enough statistical evidence to reject H0**.

Thus, it is concluded that the results obtained for the two models are **statistically significant**.

When considering this aspect, the evaluation of the performance of the models is based on the comparison of the MAE and RMSE values, being observed that the **multiple linear regression model** presents smaller values for both metrics, indicating better performance concerning the other model.

B. Classification

*1) **Pro Level**:* Using the **Pro Level attribute** for the classification of the athletes, it was possible to investigate and compare the precision of three different models. Based on the gender, altitude_results, vo2_results and hr_results, which were the attributes that, according to the correlation plot, demonstrated a higher coefficient, managed to set up and train the respective models, comparing and evaluating their results post-test. These are the Decision Tree, Neural Network and K-Nearest neighbors.

The **Decision tree** (with its design represented in Figure 10) proved to be the least accurate of the group. The **Neural Network** (demonstrated in Figure 11) provided **optimal accuracy** when created with one node after a trial-and-error procedure, taking the crown in terms of accuracy, but by a very slim margin. Third, the **K-nearest neighbors**, after research (Figure 12), proved to have an **optimal number of neighbors of forty-one**. It's important to note that this type of algorithm is known for not performing an explicit training process (it does not store the training data and calculates on-demand for each prediction). It can elaborate faster training times but take longer to predict because it needs to calculate the distance between the training points and the new point for each prediction.

With their architecture defined, the models had their algorithms ready to be trained and tested. The results of this iteration are presented in Table III, allowing for a more comprehensive overview of the performance of these techniques.

The **Neural network had the highest accuracy**, also being **more precise than the K-nearest neighbors and the Decision tree**. Surprisingly, the **K-nearest neighbors had the highest sensitivity/recall**, which means it identified the **most cyclists who could be distinguished with a Pro Level**. The **Neural Network also has superior specifity**, followed by the **Decision tree and K-nearest neighbors**, though the **K-nearest neighbors has the highest F1-Score**.

After testing the previous algorithms, **K-nearest neighbors** and **Neural Network** were elected for this training technique. A cvf of ten (10) was decided for these models, permitting the algorithms to have improved training, since the data set was sub-partitioned ten times.

For the **Neural Network** and the **K-Nearest Neighbours** models, a **mean accuracy of 68.44% and 59.81%, respectively, representing a clear advantage by the neural network** (key to note a **standard deviation of 5% and 5.47%**). Although it can indirectly increase model accuracy, which was not the case, it helps the models deliver more reliable answers based on the provided data.

As a way of verifying if there are significant differences performance-wise between the top 2 models, a **Student's t-test was applied**. This particular test is frequently employed when it's intended to compare two groups/levels of an independent variable. Reutilizing the previously calculated data (contains the average accuracy post-k-fold cross-validation) a hypothesis was formed:

- **H0**: There is a significant difference in the performance between the top two models;
- **H1**: There is not a significant difference in the performance between the top two models;
- **Significance Level (alpha)**: 0.05 (5%)

After conducting the test, a p-value: of **3.17*10÷3** at a significance level of 5% allowed us to dismiss the null hypothesis. So we can safely conclude that there are **no significant differences in performance between these models**, which is a trustworthy result, based on the obtained values.

*2) **Winter Training Camp**:* Within the scope of this study, a second predictive study was carried out for the attribute

”Winter.Training.Camp”. To analyze its predictive capacity, two different models were used: a **Decision tree and a Neural network**.

Initially, we will proceed with the creation of the decision tree model, as well as the construction of several models of neural networks, in which the parameters will be varied. We will then select the two best models, based on the lowest MAE and RMSE criteria.

But even before the creation of the models, we adopted once again the good practice of creating a data sample that divides the normalized data set into two sets: one for training and one for testing. For this, the data was divided in a **70/30 ratio**, in which 70% of the records were inserted in the training set and the remaining 30% in the test set.

Then, we proceeded to the creation of the decision tree model, utilizing ”Winter.Training.Camp” as the dependent variable and all the others as independent variables. We used the *rpart()* function with the training data set and generated a visual graph using the *rpart.plot()* function.

The result is present in the Figure 13.

When analyzing the obtained **Decision tree**, we can observe that its depth is relatively high, which indicates that the applied model is a little bit complex.

It is evident that the **”altitude_results”** and the **”hr_results”** variable play a crucial role in the decisions taken by the tree. This suggests that altitude and heart rate are determining factors in predicting the attribute under study.

When examining the leaves of the tree, it is noticeable that most of the values are concentrated between **0.4934 and 0.7289 (both heart rate)**, representing **34.72%** of the total observations (n=250).

On the other hand, the lowest values are observed in men with a **heart rate equal to or above 0.3554** and **altitude below 0.3882**, which corresponds to only **1.81%** of the total data (n=13).

For the **Neural Network** model, four distinct models were created, each with a specific number of internal nodes. These models were developed with 1, 2, 3 and 4 internal nodes, respectively.

For each model, MAE and RMSE calculations were performed by performing the prediction of values and the difference between actual and predicted values. The results of these calculations are presented in the Table VI-D.

From the Table VI-D, it is possible to identify the two best models. The neural network with 1 internal node was selected, along with the decision tree. Although the decision tree did not present the best performance in relation to the other developed networks, it was selected to allow the evaluation of two different models.

To evaluate the prediction accuracy of the attribute under study in the two best-selected models, we used the k-fold cross-validation method. We decided to divide the data set into **11 folds**, as using 12 folds resulted in errors during the calculation.

We assigned a fold to each observation in the data set by generating random samples of numbers from 1 to 11. Then, we applied the **k-fold cross-validation** method to calculate the mean and standard deviation of the prediction success rate of the attribute under study in the selected models.

For the decision tree, we obtained a **mean of 0.7047** and a **standard deviation of 0.0511** in the prediction hit rate. As for the neural network with only 1 internal node, the values were **0.6710 for the mean** and **0.0508 for the standard deviation** of the prediction success rate.

Subsequently, a hypothesis test was carried out to verify if there is a significant difference in the performance of the two best previously obtained models.

For such, a **Student’s t-test** was performed with the following hypotheses:

- **H0**: There is a significant difference in the performance of the two models
- **H1**: There is no significant difference in the performance of the two models

After performing the test, a p-value of **0.9787** was obtained. Considering a significance level of 5%, it is possible to verify that the p-value is **greater than alpha (0.05)**, so there is **not enough statistical evidence to reject H0**.

Thus, it is concluded that there is a significant difference in the performance of the two models.

Finally, the objective was to identify which of the two models presents the best performance, according to the criteria: Accuracy, Sensitivity, Specificity and F1-Score.

To carry out this identification, it was necessary to apply the **k-fold cross-validation** method again, taking into account the aforementioned criteria. To calculate each of these criteria, the **confusion matrix** obtained for each of the models was used and the values were stored in their respective matrices.

Finally, the averages of the performance metrics were calculated, resulting in the following values present in the Table II:

TABLE II
PERFORMANCE METRICS RESULTS FOR EACH OF THE MODELS

Model	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Decision Tree	0.7111	0.6370	0.3398	0.9073	0.4380
Neural Network	0.6617	0.5151	0.5936	0.6997	0.5463

Accuracy represents the proportion of correct predictions about the total samples. In the case of the **decision tree**, we obtain an accuracy of 0.7111292, indicating that, on average, **71.11%** of the predictions were correct. For the **neural network with 1 internal node**, the accuracy is a little lower, at **66.17%**.

Precision refers to the ratio of correctly predicted positive outcomes to the total predicted positive outcomes. In the **decision tree**, the precision is **0.6370409**, while in the **neural network with 1 internal node**, the precision is **0.5151077**.

Sensitivity, also known as the True Positive Rate, represents the proportion of correctly identified positive results to the

total number of actual positive results. In the **decision tree**, the sensitivity is **0.3397574**, while in the **neural network with 1 internal node**, the sensitivity is **0.5936268**.

Specificity, also known as the True Negative Rate, represents the proportion of correctly identified negative results in relation to the total number of actual negative results. In the **decision tree**, the specificity is **0.9072962**, while in the **neural network with 1 internal node**, the specificity is **0.6996822**.

The **F1-Score** is a measure that combines accuracy and sensitivity into a single value, providing an overall measure of performance. In the **decision tree**, the F1-Score is **0.4380479**, while in the **neural network with 1 internal node**, the F1-Score is **0.5462835**.

When analyzing these results, we can observe that the **decision tree** presents a **slightly higher accuracy**, as well as a **greater specificity** compared to the neural network with 1 internal node. However, the **neural network** shows a **higher sensitivity** and a **higher F1-Score**.

*3) **Gender**:* Finally, as the culmination of the **Classification** study, an analysis was conducted to assess the predictive capability of the "gender" attribute. Two distinct machine learning models, namely the **Neural Network** and the **K-Nearest Neighbors**, were employed for this purpose.

Initially, multiple **Neural Network** models were created, varying the internal nodes utilized, and several **K-Nearest Neighbors** models were constructed. The selection process involved identifying the two best models based on criteria such as the lowest **Mean Absolute Error (MAE)** and **Root Mean Square Error (RMSE)** for the **Neural Network**, and the highest accuracy for the **K-Nearest Neighbors**.

Prior to model creation, a standard practice was followed, involving the division of the normalized data set into training and testing sets. A **70/30 ratio** was adopted, allocating **70%** of the records to the training set and the remaining **30%** to the test set.

To conduct a predictive study on this attribute, it was necessary to determine the optimal model for each **Machine-learning** method. This entailed identifying the ideal number of internal nodes for the **Neural Network** and the optimal K value for the **K-Nearest Neighbors**.

Regarding the **Neural Network**, multiple models were trained with varying numbers of internal nodes to find the best-performing configuration. The various trained models are presented in Table VI-D:

Following evaluation, the model with **16 nodes in the first layer, 10 nodes in the second layer, 6 nodes in the third layer, and 2 nodes in the fourth layer demonstrated superior performance**. There were other models with a similar number of nodes but this demonstrates This model yielded an **MAE** of **0.0727091** and an **RMSE** of **0.2588458**, establishing it as the most effective **Neural Network** model. In Figure 14 is possible to see the full network.

Subsequently, a search for the optimal **K-Nearest Neighbors** model was conducted. This involved testing different values of K to identify the model with the highest accuracy. Through fifty iterations, the model with K=1 emerged as the best performer, achieving an accuracy of **0,85086**, as depicted in Figure 15.

On the Y-axis, the model accuracy values are depicted, while the K values are represented on the X-axis. As evident from the plot, the red circle corresponds to the highest accuracy value, reaching **85.086%**, and is associated with a K value of **one**. This indicates that the model achieved optimal performance with this specific K value, as demonstrated graphically.

In order to leverage the best models from two distinct machine learning approaches, the **K-Cross Validation** technique was employed. This method involves dividing the normalized data set into k subsets or folds of equal size. The models are then trained and assessed k times, with each iteration using a different fold as the test set and the remaining folds as the training set. This approach is crucial for obtaining the **average accuracy** of the two models across the K folds. To achieve the **highest accuracy**, various values for K folds were explored, ultimately determining that the optimal value was 11 since the mean value of the accuracy for both models was the highest one achieved and the standard deviation was the lowest one, having the Neural Network with a mean accuracy value of **0,8016012** and standard deviation of **0,04812821** while the K-Nearest Neighbours had an accuracy mean value of **0.5400159** and a standard deviation of **0,025018**. Based on these values, it can be inferred that the best **Neural Network** model demonstrates higher overall accuracy compared to the **K-Nearest Neighbors** model. However, it should be noted that the **Neural Network** model exhibits a higher degree of variability in the values, as evidenced by a comparatively higher standard deviation. Based on these values, it is possible to deduce that the best **Neural Network** model presents a better accuracy overall compared to the **K-Nearest Neighbours**, but has a higher inconsistency in the values, presenting a higher standard deviation.

The utilization of **K-fold cross-validation** yields a more dependable estimation of a model's performance when compared to a single train-test split. It helps alleviate the impact of random fluctuations in the train-test split, ensuring a more robust evaluation of the model's capacity to generalize to unseen data.

Afterward, a hypothesis test was conducted to determine whether the results obtained from the two best models are statistically significant.

A **Student's t-test** was performed with the following hypotheses:

- **H0:** The results obtained in both models are statistically significant.
- **H1:** The results obtained in the two models are not statistically significant.

Upon conducting the test, a p-value of **0.7841** was obtained. Considering a significance level of 5%, it can be observed that the p-value is **greater than the alpha value** (0.05). Therefore, there is insufficient statistical evidence to reject H0. Consequently, it is concluded that the results obtained for both models are **statistically significant**.

To determine the best model, various criteria such as **accuracy, precision, sensitivity, specificity, and F1 score** were utilized. These criteria provide insight into the performance of each model. To obtain more reliable estimations of the models' performance, these values were calculated **for each K-fold iteration, and subsequently, the mean value was determined**. By employing 11 folds, the following values were obtained for the two models, as shown in Table VI:

Based on the values from Table VI the performance of the two models was compared in terms of predictive accuracy, precision, sensitivity, specificity, and F1 score. It is possible to deduce:

- **Accuracy:** a model that represents the overall correctness of its predictions. The **K-Nearest Neighbours** model achieved a mean accuracy of **0.5276064**, while the **Neural Network model with 16, 10, 6, and 2 nodes obtained a higher mean accuracy of 0.7908638**. This indicates that the **Neural Network** model generally performed better in predicting the **"Gender"** attribute;
- **Precision:** measures the proportion of correctly predicted positive instances (True Positives) out of all instances predicted as positive. The **Neural Network model exhibited a slightly higher mean precision of 0.7929163** compared to the **K-Nearest Neighbours model's mean precision of 0.5439251**. This suggests that the **Neural Network model had a better ability to correctly predict the positive instances**;
- **Sensitivity (Recall):** quantifies the ability of a model to correctly identify positive instances (True Positives) out of all actual positive instances. The **Neural Network model demonstrated a higher mean sensitivity of 0.8149046**, indicating its better performance in identifying the actual positive instances. In contrast, the **K-Nearest Neighbours model had a lower mean sensitivity of 0.5644452**;
- **Specificity:** measures the ability of a model to correctly identify negative instances (True Negatives) out of all actual negative instances. The **Neural Network** model achieved a mean specificity of **0.7613693**, indicating its better performance in correctly identifying the negative instances. On the other hand, the **K-Nearest Neighbours** model had a lower mean specificity of **0.4871879**;
- **F1 Score:** combines precision and sensitivity into a single metric, providing a balanced evaluation of a model's performance. The **Neural Network model achieved a higher mean F1 score of 0.8014577**, suggesting its **better overall performance in terms of precision and sensitivity**. The **K-Nearest Neighbours model obtained a lower mean F1 score of 0.5521911**.

Based on these results, the **Neural Network model with 16, 10, 6, and 2 nodes outperformed the K-Nearest Neighbors** model in terms of accuracy, precision, sensitivity, specificity, and F1 score. Therefore, the **Neural Network** model is considered to have **the best overall performance for predicting the "Gender" attribute**.

VI. Conclusion

Through the cyclists' data set, we were allowed to study how the different regression and **Machine learning algorithms** compare. **Decision trees**, **Neural Networks**, and **K-nearest neighbors** (**KNN**) are all **Machine learning** algorithms commonly used in various fields of artificial intelligence and data analysis to make predictions, classify data, or find patterns in data sets.

A. Heart Rate and VO2 Results

After studying the heart rate, we inferred that a more sophisticated regression model was achievable. So, when dissecting the VO2 results, a **Multiple Linear Regression** model was used, allowing for a **strengthened accuracy and ability to detect patterns**. It performed slightly better than the single-node **Neural Network and substantially better than the Decision tree**.

B. Pro Level

After conducting a thorough examination of the given diagrams and table, we could verify that the three models had excellent results. However, based on the required application, even if the **accuracy might be lower, it may be benefic to use a specific model**. For example, if you need a **low rate of false negatives you should choose the model with the highest sensitivity/recall**. In this case, the **KNN** model. But, if you need the highest number of guesses, the **Neural Network** may be more suitable. All goes down to the use case.

C. Winter training camp

On the Winter Training Camp results, the data enlightened that, once again, the single-node **Neural Network** was the best performing. However, for the sake of evaluating two different models, the **Decision tree** was elected. A **higher F1-Score belongs to the Neural Network by a substantial margin**, even though its **accuracy is inferior**. This is due to an **inferior false-negatives rate (high sensitivity)**. It may be advisable to go with the **Neural Network**, however, as previously said, specific use cases may require the other model.

D. Gender

When studying the gender field, contrary to previous results, which also may be influenced by the machines the algorithms are executed on, the **Neural Network performed best on a 4 layer with 16,10,6 and 2 nodes, respectively**. The closest model was the **KNN** with a k value of **one**, however, all of its **stats were significantly lower, which leads to an incredibly high F1-Score, electing it the best model**.

APPENDIX

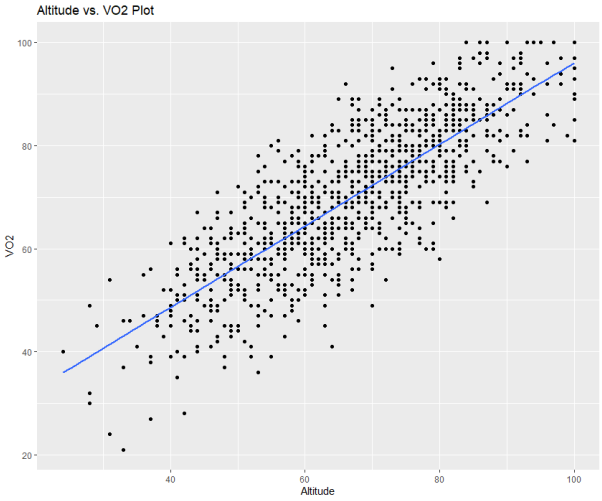

Fig. 1. Altitude vs VO2 Scatter Plot

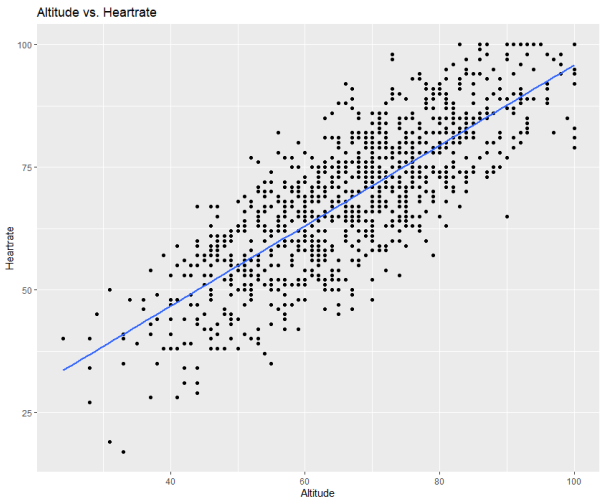

Fig. 2. Altitude vs Heart Rate Scatter Plot

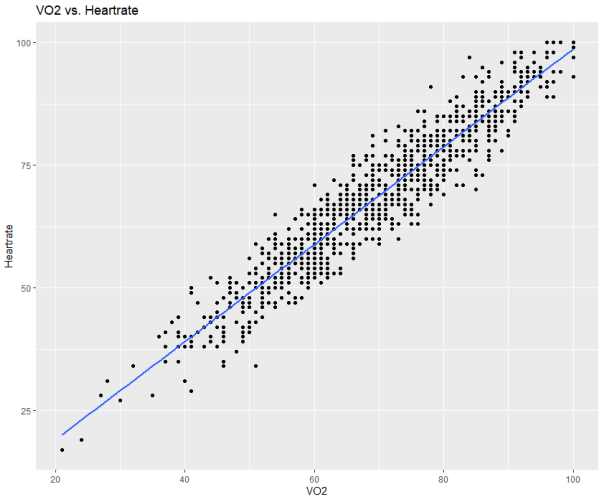

Fig. 3. VO2 vs Heart Rate Scatter Plot

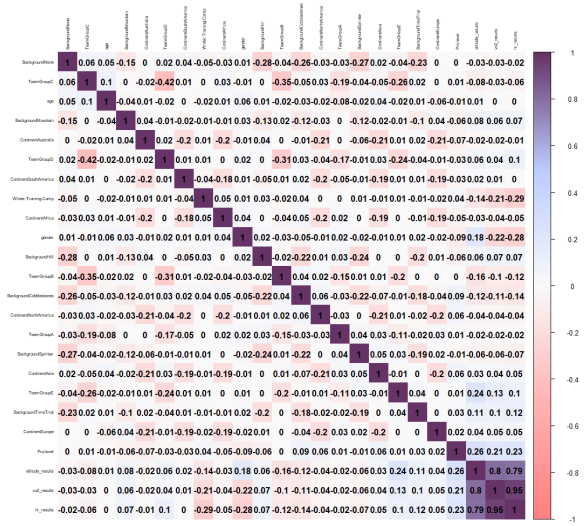

Fig. 4. Correlation Plot

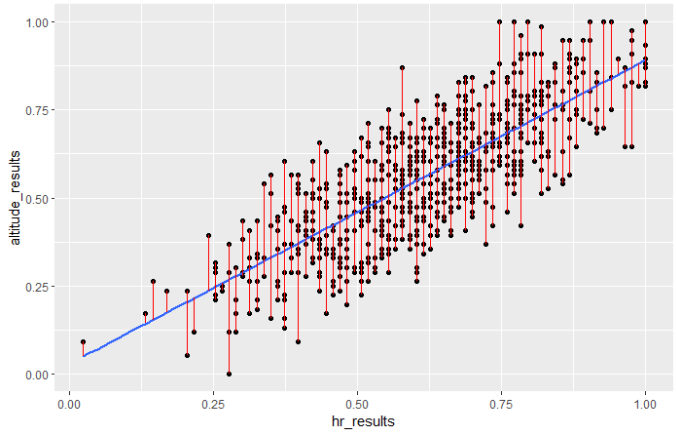

Fig. 5. Scatter plot and regression line showing the correlation between ”altitude_results” and ”hr_results”.

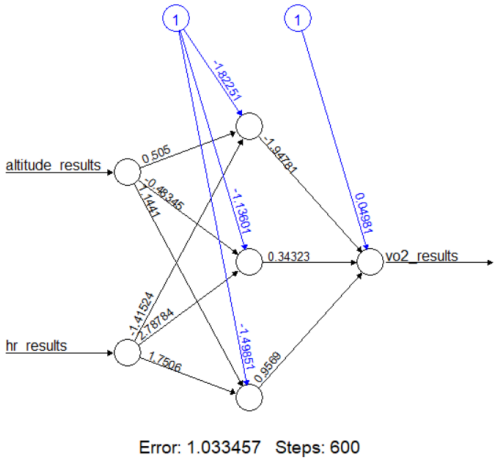

Fig. 8. Neural network plot with 3 internal nodes for vo2_results.

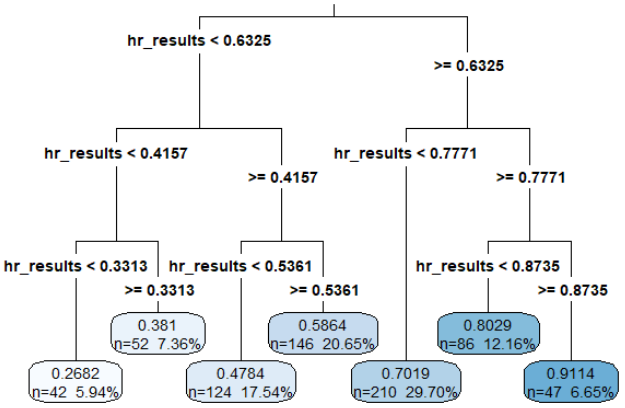

Fig. 6. Decision tree visual graph for the ”vo2_results” attribute.

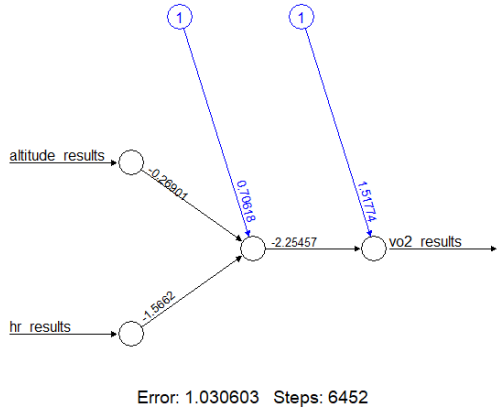

Fig. 7. Neural network plot with 1 internal node for vo2_results.

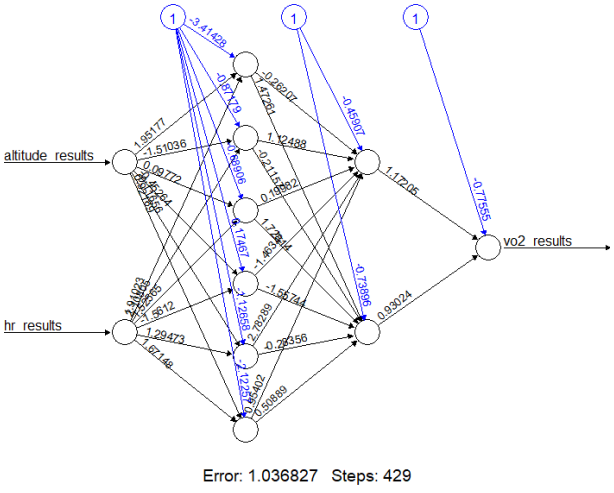

Fig. 9. Neural network plot with 2 internal levels: 6 and 2 internal nodes for vo2_results.

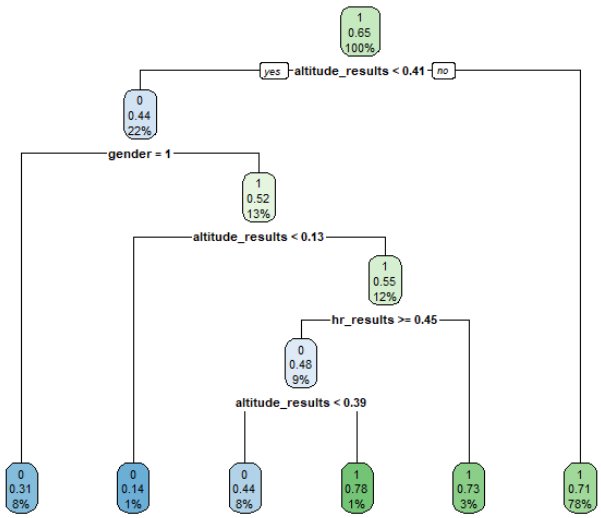

Fig. 10. Decision Tree Representation for the "Pro.Level" attribute

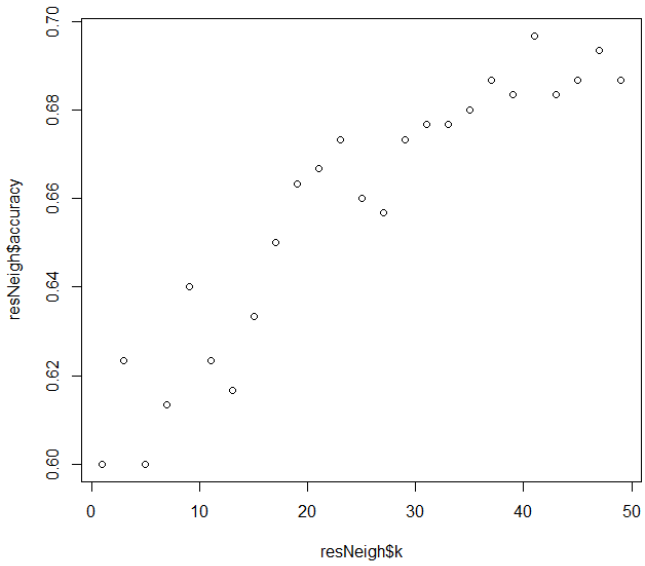

Fig. 12. Plot representing the best K-value for the "Pro.level" KNN Model

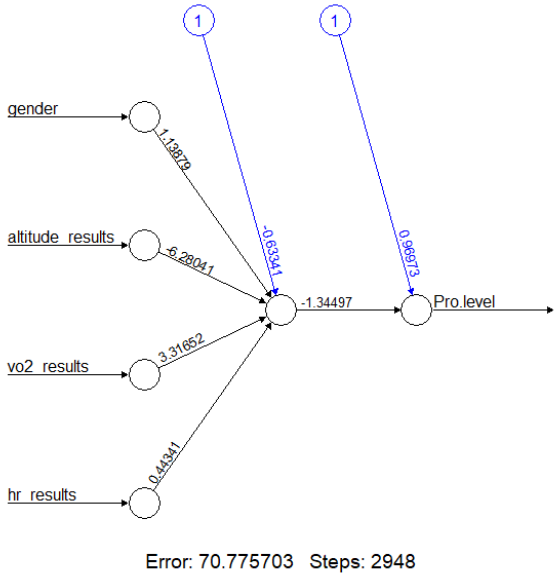

Fig. 11. Neural Network Representation for the "Pro.Level" attribute

TABLE III
PERFORMANCE METRICS RESULTS FOR EACH OF THE MODELS - PRO.LEVEL

Model	Accuracy	Precision	Sensitivity	Specifity	F1-Score
Decision Tree	0.6733	0.7007	0.8900	0.2400	0.7841
Neural Network	0.7000	0.7182	0.9050	0.2900	0.8008
KNN	0.6977	0.7041	0.9400	0.2100	0.8051

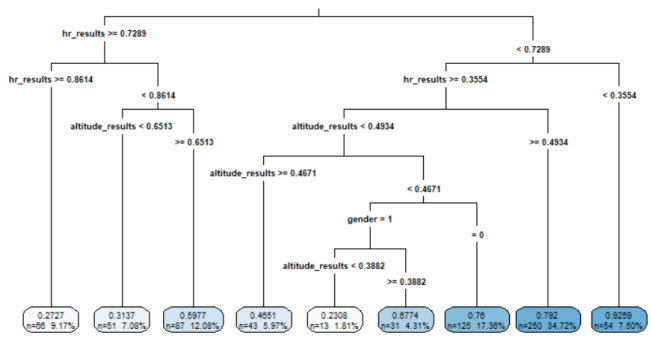

Fig. 13. Decision tree visual graph for the "Winter.Training.Camp" attribute.

TABLE IV
MAE AND RMSE VALUES OF EACH MODEL PERFORMED FOR THE WINTER.TRAINING.CAMP ATTRIBUTE

Model	MAE	RMSE
Decision Tree	0.4232	0.4951
Neural Network (1 internal node)	0.3664	0.4456
Neural Network (2 internal nodes)	0.3889	0.4705
Neural Network (3 internal nodes)	0.4012	0.5008
Neural Network (4 internal nodes)	0.4214	0.5098

TABLE V
NEURAL NETWORK TRAINED MODELS FOR THE GENDER ATTRIBUTE

Model	MAE	RMSE
Neural Network (1 internal node)	0.1583	0.2870
Neural Network (3,1 internal nodes)	0.1017	0.2591
Neural Network (6,2 internal nodes)	0.1109	0.2822
Neural Network (9,3 internal nodes)	0.1209	0.2822
Neural Network (10,6,2 internal nodes)	0.0902	0.2538
Neural Network (16,10,6,2 internal nodes)	0.0727	0.2588

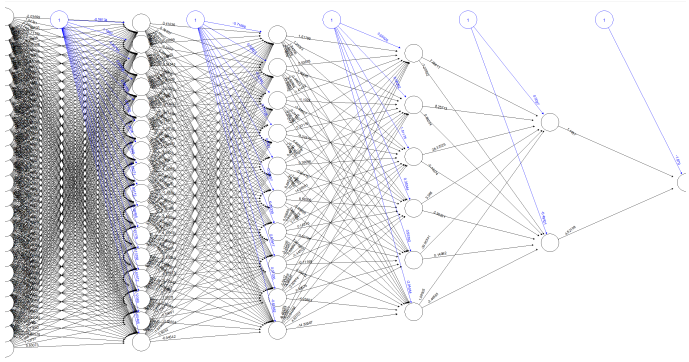

Fig. 14. Neural Network with 16,10,6,2 internal nodes for the "Gender" attribute

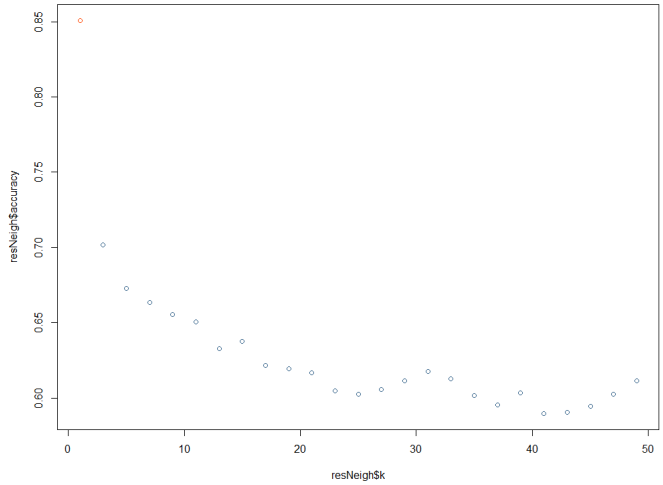

Fig. 15. Plot representing the best K-value for the "Gender" attribute

TABLE VI
DIFFERENT CRITERIA TO EVALUATE THE MODELS FROM THE GENDER ATTRIBUTE

Model	Accuracy	Precision	Sensitivity	Specificity	F1
K-Nearest Neighbours	0.5276	0.5439	0.5644	0.4872	0.5522
Neural Network	0.7908	0.7929	0.8149	0.7614	0.8015

REFERENCES

[1] GeeksforGeeks. (n.d.). *K-NN Classifier in R Programming*. Retrieved from https://www.geeksforgeeks.org/k-nn-classifier-in-r-programming/

[2] DataCamp. (n.d.). *Building Neural Network (NN) Models in R*. Retrieved from https://www.datacamp.com/tutorial/neural-network-models-r

[3] Displayr. (n.d.). *What is a Correlation Matrix?*. Retrieved from https://www.displayr.com/what-is-a-correlation-matrix/

[4] Stats and R. (n.d.). *Student's t-test in R and by hand: how to compare two groups under different scenarios?*. Retrieved from https://shorturl.at/fsNO2

[5] Analytics Vidhya. (n.d.). *K-Fold Cross Validation Technique and its Essentials*. Retrieved from https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

[6] TC Technology Knowledge Base. (n.d.). *Plot data in R-Studio*. Retrieved from https://teacherscollege.screenstepslive.com/a/1130011-plot-data-in-r-studio

[7] Kanstrén, T. (n.d.). *A Look at Precision, Recall, and F1-Score*. Retrieved from https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

[8] Vitalflux. (n.d.). *Accuracy, Precision, Recall & F1-Score - Python Examples*. Retrieved from https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

[9] SDS Club. (n.d.). *Linear Regression vs Multiple Regression: Know the Difference*. Retrieved from https://sdsclub.com/linear-regression-vs-multiple-regression-know-the-difference/

[10] JMP. (n.d.). *Multiple Linear Regression*. Retrieved from https://www.jmp.com/en_nl/learning-library/topics/correlation-and-regression/multiple-linear-regression.html

[11] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (pp. 986–996). Springer Berlin Heidelberg.