Time Series and Logistic Regression

Reyna Vargas Antonio
School of Computing
National College of Ireland
Dublin, Ireland
x23127635@student.ncirl.ie

***Abstract*—Analyzing the time series of Grass Minimum Temperature by degrees Celsius, from the dataset weather_revised using different methods of forecasting. The purpose is to recognize the pattern of the series time from 2019 to 2022 and to apply the best model for performing the prediction of future values in the data set test, which contains values from January 2023 to October 2023. The time series models reported in this project are Simple Time Series Models, Exponential Smoothing, and ARIMA/ SARIMA. On the other hand, analyzing, preprocessing, and a dataset that contains information about cardiac information, the purpose of this analysis is to evaluate if it is affected the presence or absence of cardiac conditions by factors like age, weight, gender, or fitness score. The most popular method for solving problems that involve quantitative variables is Linear Regression or Multiple Regression, even though not always happen, there are many areas where a predictor is categorical, so it can use Logistic Regression to provide valuable predictions.**

Keywords—Time Series, prediction, Temperature, Exponential Smoothing, ARIMA, Logistic Regression, Cardiac Condition.

I. Introduction

A Time Series (Joseph) is a set of observations sequentially in time, such as every hour or every month, that are collected by businesses and are growing larger and wider by the minute. Even the analysis of time series is performed for different purposes, the main ones are classification, prediction, sequence detection, filtering, and simulation (Gharehbaghi). Nowadays, there are various machine learning models and training methods that can help to perform the time series and improve the learning process, which means, less handmade manipulation and more process executes.

Another method applied in this project is logistic regression (David W. Hosmer Jr.), which is the most frequently used in the analysis of categorical data, the goal of this kind of analysis is to find the best fitting model that can interpret the relationship between an outcome variable and a set of independent variables. Logistic Regression tells the probability of response is correct and then it uses a predefined threshold to assign classes.

II. System design

A. Methodology

The KDD methodology (Delen, Real-World Data Mining: Applied Business Analytics and Decision Making) is a comprehensive end-to-end process that encompasses various steps to convert the data into knowledge.

Data Selection: Identify and select the relevant data for analysis.

Data Preprocessing: This step involves cleaning data, removing outliers, and removing missing values, inconsistencies, or irrelevant data.

Data Transformation: Transform the data into a format that data mining techniques can analyse, involves reducing the data dimensionality, normalizing the data, etc.

Data Mining: Apply various data mining techniques to the transformed data to discover hidden patterns, trends, relationships, and insights.

Pattern Evaluation: To evaluate the discovered patterns to determine their usefulness and relevance.

Knowledge Representation: This step involves representing the knowledge extracted from the data in a way that can easily be understood and used through visualizations, reports, or meaningful insights into the data.

III. Time Series Analysis

A. Data Selection

The objective is to analyze one of the variables in the data set weather.csv, which one contains a daily time series of historical weather data. The information is from January 1st , 1942 to October 31st, 2023. The variable for this analysis is *gmin- Grass Minimum Temperature (C).*

B. Data Preprocessing

The initial dataset contains 29,889 observations and 9 variables, according to the instructions in the report the dataset is reduced to two columns and 29,889 observations.

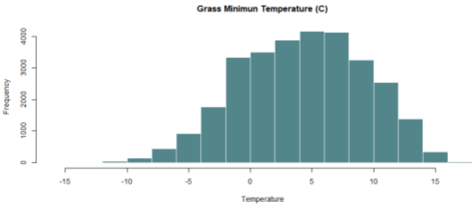


Figure 1. Histogram of Grass Minimum Temperature (C)

The figure above shows the distribution of temperature from 1942 to October 2023.

- The dataset is converted to time series by ts function, starting from 1ts January 1942 to 31st October 2023, with a daily frequency.

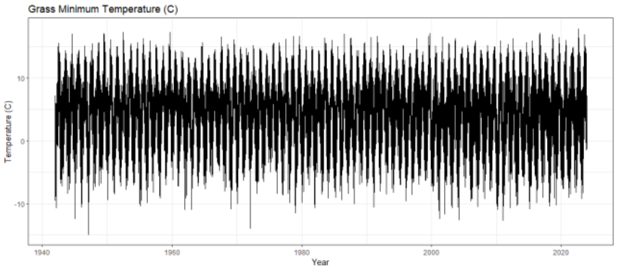


Figure 2. Time series of Grass Minimum Temperature (C) from 1942 to 2023

At the first view, the time series shows a horizontal pattern, with a constant mean and a variability constant over time.

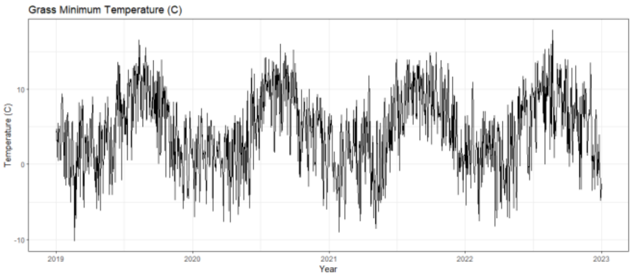


Figure 3. Train Time Serie

The figure above shows that the dataset was divided into train datasets, which contain values from 2019 to 2013, the pattern shows a seasonality pattern, it looks like in the second quarter the temperature starts to increase and the last quarter decreases, the stationary pattern is better to see.

Below, the image shows the time series of the test dataset, which contains values from January 2023 to October 2023.

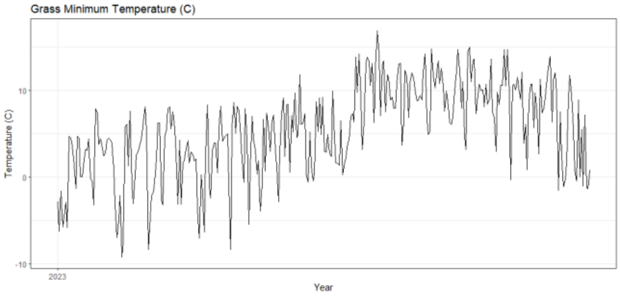


Figure 4. Test Time Serie

- **Seasonal Decomposition**

Using the function of seasonal decomposition, it can be seen that the time series has a small trend in the year 2022, a seasonality pattern, and cyclic behavior. All this one, it can be seen easily by the decomposition plot and ggtsdisplayplot function.

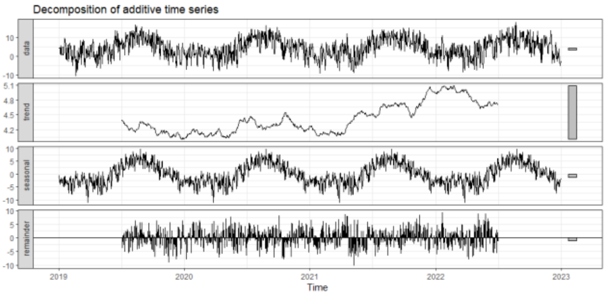


Figure 5. Plot of Seasonal Decomposition

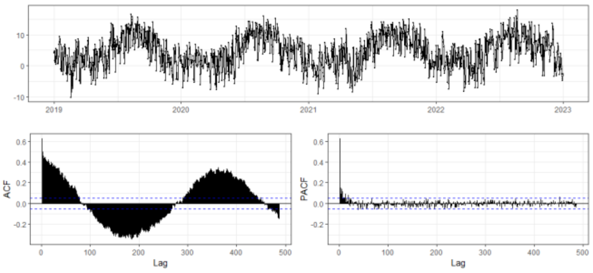


Figure 6. ACF, PACF and Time Serie

1. ***Simple Time Series Models:***

Simple Moving Average: It is an easy technique to identify trends and smooth them. This is one of the simplest methods of smoothing a time series for it cannot see a big change in the pattern in the figure above.

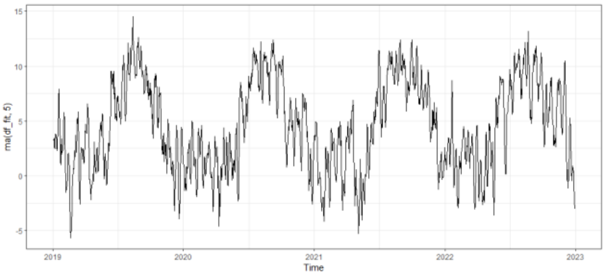
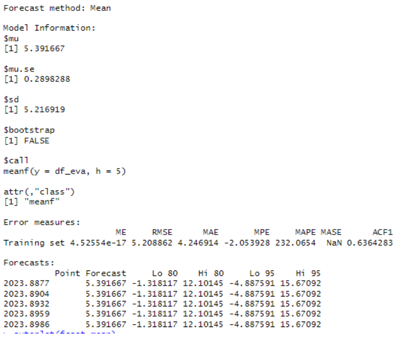


Figure 7. ma(time serie, 5)



```
Forecast method: Mean

Model Information:
$mu
[1] 5.391667

$mu.se
[1] 0.2898288

$sd
[1] 5.216919

$bootstrap
[1] FALSE

$call
meanf(y = df_eva, h = 5)

attr(,"class")
[1] "meanf"

Error measures:
                       ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 4.52554e-17 5.208862 4.246914 -2.053928 232.0654  NaN 0.6364283

Forecasts:
          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2023.8877       5.391667 -1.318117 12.10145 -4.887591 15.67092
2023.8904       5.391667 -1.318117 12.10145 -4.887591 15.67092
2023.8932       5.391667 -1.318117 12.10145 -4.887591 15.67092
2023.8959       5.391667 -1.318117 12.10145 -4.887591 15.67092
2023.8986       5.391667 -1.318117 12.10145 -4.887591 15.67092
```

Figure 8. Summary of Forecast SMA

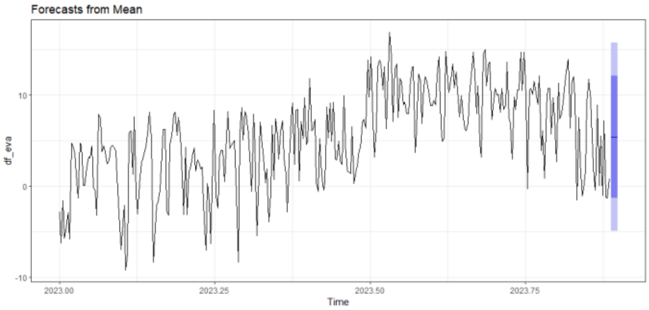
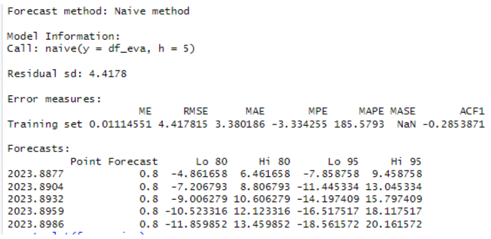


Figure 9. Time Serie 2023 + Forecast SMA

2. ***Naïve Model:*** (Random Walk) It is a forecasting method that only uses last year's actual value data as a forecast for the actual year and so on.



```
Forecast method: Naive method

Model Information:
Call: naive(y = df_eva, h = 5)

Residual sd: 4.4178

Error measures:
                     ME     RMSE      MAE       MPE     MAPE MASE       ACF1
Training set 0.01114551 4.417815 3.380186 -3.334255 185.5793  NaN -0.2853871

Forecasts:
          Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
2023.8877            0.8  -4.861658  6.461658  -7.858758  9.458758
2023.8904            0.8  -7.206793  8.806793 -11.445334 13.045334
2023.8932            0.8  -9.006279 10.606279 -14.197409 15.797409
2023.8959            0.8 -10.523316 12.123316 -16.517517 18.117517
2023.8986            0.8 -11.859852 13.459852 -18.561572 20.161572
```

Figure 10. Summary of Forecast Naïve Model

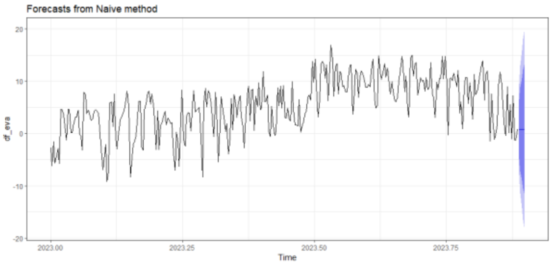
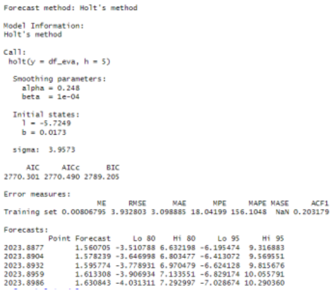


Figure 11. Time Serie 2023 + Forecast Naïve model

3. Exponential Smoothing Model: It is a method of forecasting a moving average with a sophisticated weight but still easy to use.



```
Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = df_eva, h = 5)

  Smoothing parameters:
    alpha = 0.248
    beta  = 1e-04

  Initial states:
    l = -5.7249
    b = 0.0173

  sigma:  3.9573

     AIC     AICc      BIC
2770.301 2770.490 2789.205

Error measures:
                     ME     RMSE      MAE      MPE     MAPE MASE     ACF1
Training set 0.00806795 3.932803 3.098885 18.04199 156.1048  NaN 0.203179

Forecasts:
          Point Forecast     Lo 80    Hi 80     Lo 95     Hi 95
2023.8877       1.560705 -3.510788 6.632198 -6.195474  9.316883
2023.8904       1.578239 -3.646998 6.803477 -6.413072  9.569551
2023.8932       1.595774 -3.778931 6.970479 -6.624128  9.815676
2023.8959       1.613308 -3.906934 7.133551 -6.829174 10.055791
2023.8986       1.630843 -4.031311 7.292997 -7.028674 10.290360
```

Figure 12. Summary of Forecast Exponential Smoothing

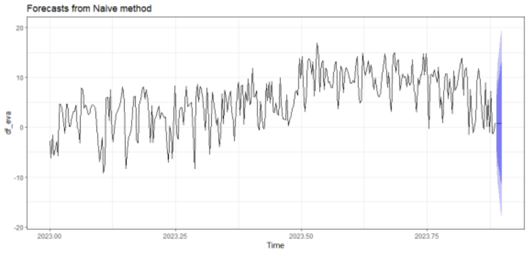


Figure 13. Time Serie 2023 + Forecast Exponential Smoothing

4. **ARIMA:** an approach to forecasting, predicted values are a linear function of recent actual values and recent errors of prediction (residual).

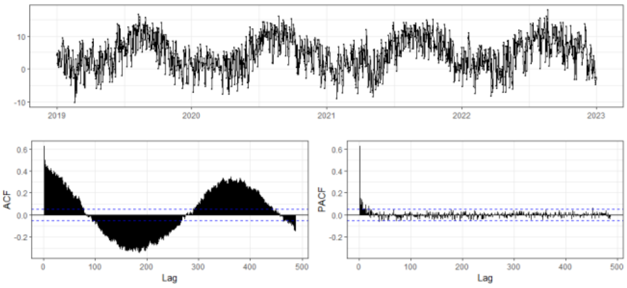


Figure 14. ACF, PACF and Time Serie

Figure 15 is a ACF on non-stationary, and th figure 16, is like ACF on difference/ stationary). Its sample autocovariance function (ACF) also decays to zero very slowly, meaning past values have a long lasting impact on current values.

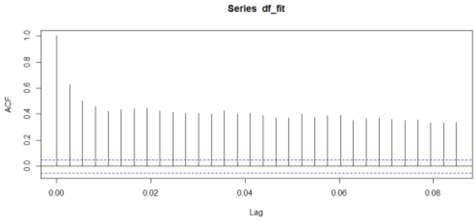


Figure 15. ACF

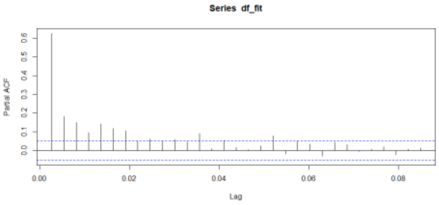


Figure 16. PACF

- Applying the first difference, the model gets a stationary pattern with means close to zero value.

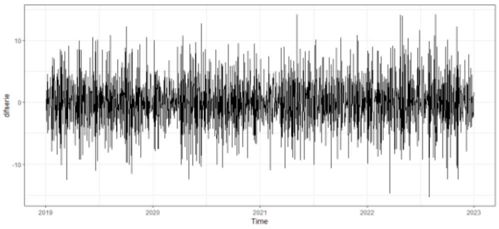


Figure 17. The first difference in the time series

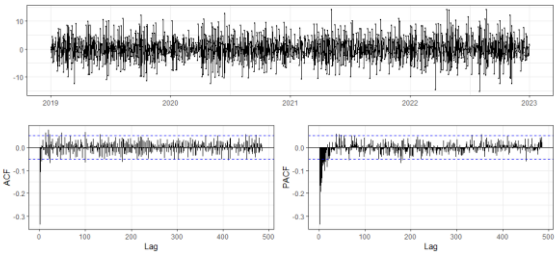


Figure 18. ACF, PACF, Time Series with first difference

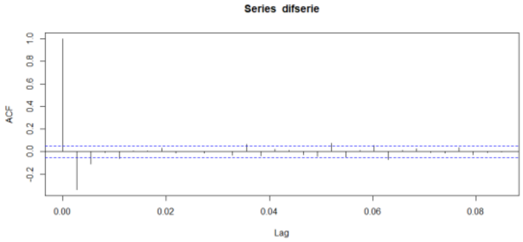


Figure 19. ACF

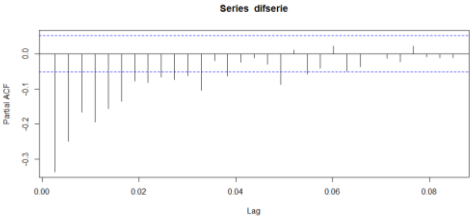
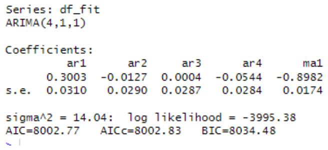


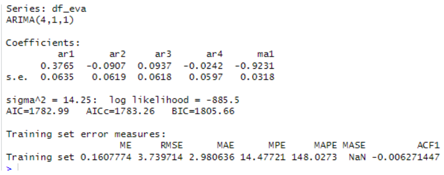
Figure 20. ACF

The prediction was based on Arima(4,1,1).



```
Series: df_fit
ARIMA(4,1,1)

Coefficients:
         ar1      ar2     ar3      ar4      ma1
      0.3003  -0.0127  0.0004  -0.0544  -0.8982
s.e.  0.0310   0.0290  0.0287   0.0284   0.0174

sigma^2 = 14.04:  log likelihood = -3995.38
AIC=8002.77   AICc=8002.83   BIC=8034.48
```

Figure 21. Autoarima



```
Series: df_eva
ARIMA(4,1,1)

Coefficients:
         ar1      ar2     ar3      ar4      ma1
      0.3765  -0.0907  0.0937  -0.0242  -0.9231
s.e.  0.0635   0.0619  0.0618   0.0597   0.0318

sigma^2 = 14.25:  log likelihood = -885.5
AIC=1782.99   AICc=1783.26   BIC=1805.66

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE MASE         ACF1
Training set 0.1607774 3.739714 2.980636 14.47721 148.0273  NaN -0.006271447
>
```

Figure 22. Prediction by Autoarima

The residuals are therefore correlated and withi a mean =0, normally distributed, The ADF is stationary.

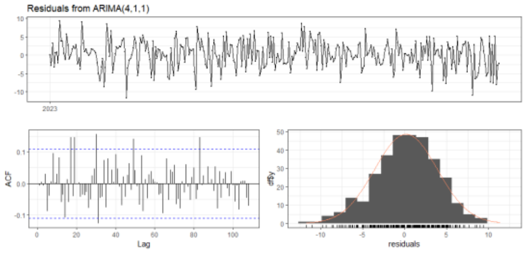


Figure 23. Residuals

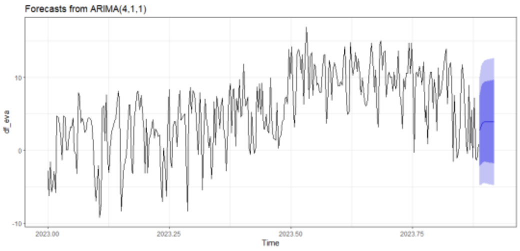
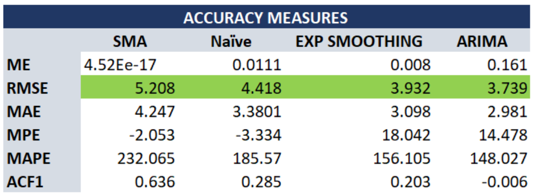


Figure 24. Prediction Plot

PATTERN EVALUATION

Comparing the accuracy measures among the models, it can be seen that the best model is ARIMA(4,1,1) with a RMSE of 3.7 and also it has the lower average error between the predictions and actual temperature in the dataset, which is 2.981, the model with lower variance between differently scaled data is Simple Moving Average (-2.05). The best model for predicting the Grass Minimum Temperature for the next days, after 31th October will be the Arima Model. However, it is recommended to make the predictions by periods or cycles, as it will help to keep the pattern of the Time series.



ACCURACY MEASURES				
	SMA	Naïve	EXP SMOOTHING	ARIMA
ME	4.52Ee-17	0.0111	0.008	0.161
RMSE	5.208	4.418	3.932	3.739
MAE	4.247	3.3801	3.098	2.981
MPE	-2.053	-3.334	18.042	14.478
MAPE	232.065	185.57	156.105	148.027
ACF1	0.636	0.285	0.203	-0.006

Table 1. Accuracy Measures

IV. Logistic Regression

A. Data Selection

The file cardiac.csv is a dataset containing information about 100 participants with their names, age, weight, gender, fitness score, and the predictor variable cardiac condition, this variable allows to evaluate if the factors before mentioned take part of the presence or absence of the cardiac condition. Into the dataset there are 100 observations and 5 variables.

B. Data Preprocessing

1) Some changes in the dataset were like moving the variable "casino" to "Row. names" due to the values being the same.

2) No missing values, the figure below shows 0% of missing values.

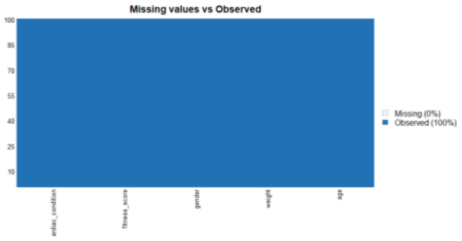


Figure 25. Missmap of missing values

3) Visualizations of the numerical variables, the histograms of the numerical variables do not show any pattern of normal distribution, even though it could see some outliers.

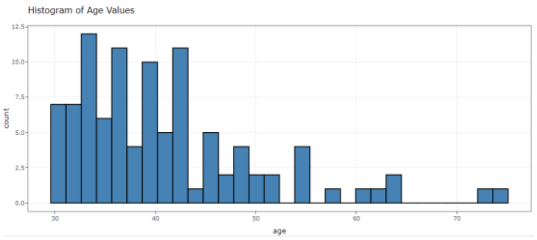


Figure 26. Histogram of Age Values

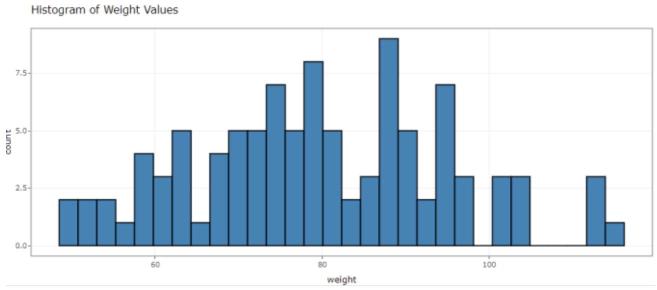


Figure 27. Histogram of Weight values

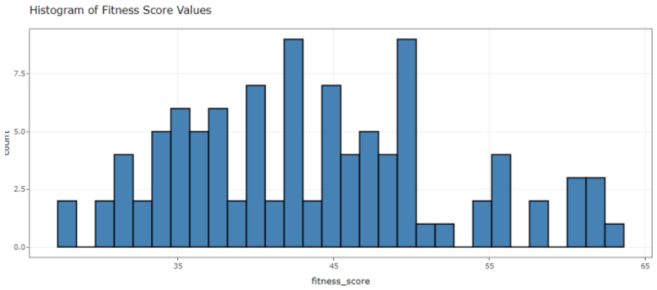


Figure 28. Histogram of fitness_score

4) Visualizations of categorical data, there are double of man than women, and 65% of the observations allow to cardiac condition “Absent”.

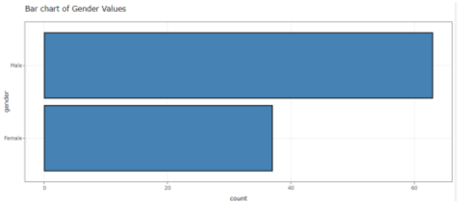


Figure 29. Barplot of variable gender

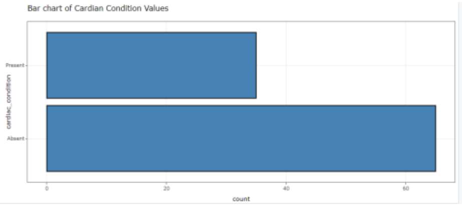


Figure 30. Barplot of variable cardiac condition

5) The variables do not show any relationship between them, additionally, the values are distributed by cardiac condition variable, the colour orange is “absent” of cardiac condition and green to “Present” cardiac condition.

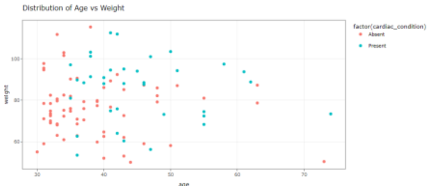


Figure 31. Scatterplot of Age vs Weight

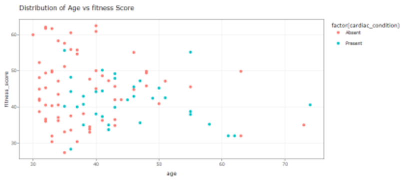


Figure 32. Scatterplot of Age vs Fitness_score

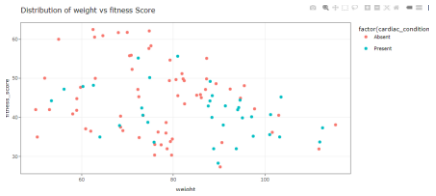


Figure 33. Scatterplot of weight vs fitness_score

6) Remove outliers, applying interquartile Range (IQR) from the dataset with just four outliers.

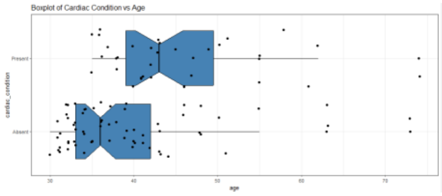


Figure 34. Outliers in Age variable

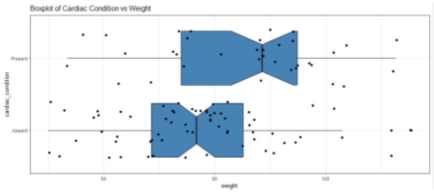


Figure 35. Outliers in weight variable

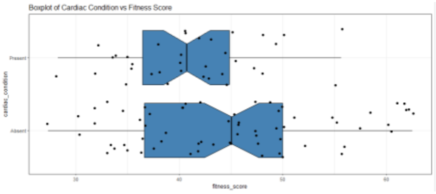


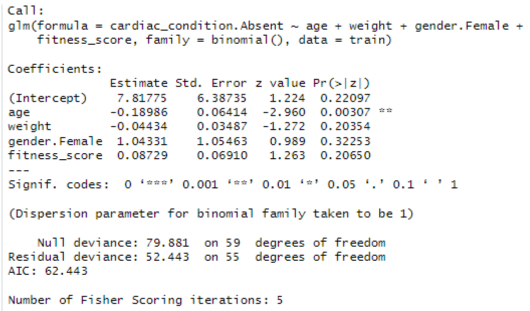
Figure 36. Outliers in fitness score variable

C. Data Transformation

Dummy Variables: The variables gender and cardiac_condition were transformed into binary variables for their analysis of the model. So, it has gotten the variables gender.Female, gender.Male, cardiac_condition.Absent and cardiac_condition.Present.

D. Data Mining

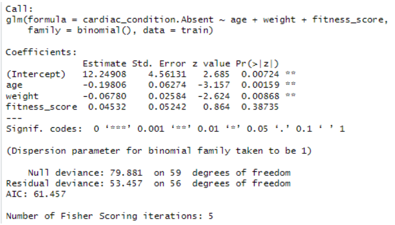
The preprocessed dataset is split into train and test for modeling by Logistic Regression. The figure below applies the variables age, weight, gender.Female and fitness_score for modeling the variable cardiac_condition.Absent. The model shows that the variable gender.The female has a high P-value that is not influencing the model.



```
Call:
glm(formula = cardiac_condition.Absent ~ age + weight + gender.Female +
    fitness_score, family = binomial(), data = train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    7.81775    6.38735   1.224  0.22097
age           -0.18986    0.06414  -2.960  0.00307 **
weight        -0.04434    0.03487  -1.272  0.20354
gender.Female  1.04331    1.05463   0.989  0.32253
fitness_score  0.08729    0.06910   1.263  0.20650
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.881  on 59  degrees of freedom
Residual deviance: 52.443  on 55  degrees of freedom
AIC: 62.443

Number of Fisher Scoring iterations: 5
```

Figure 37. Model 1 by Logistic Regression

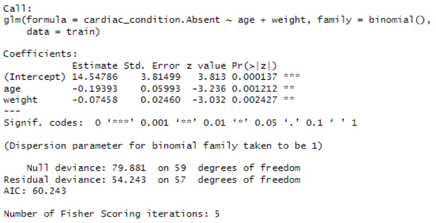
The second model has improved the significance of weight in the model, which means that only the variable fitness_score is not giving high significance to the model.



```
Call:
glm(formula = cardiac_condition.Absent ~ age + weight + fitness_score,
    family = binomial(), data = train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.24908    4.56131   2.685  0.00724 **
age           -0.19806    0.06274  -3.157  0.00159 **
weight        -0.06780    0.02584  -2.624  0.00868 **
fitness_score  0.04532    0.05242   0.864  0.38735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.881  on 59  degrees of freedom
Residual deviance: 53.457  on 56  degrees of freedom
AIC: 61.457

Number of Fisher Scoring iterations: 5
```

Figure 38. Model 2 by Logistic Regression

In the third model, the variables age and weight give relevant significance to the model, hence, the final model is done.

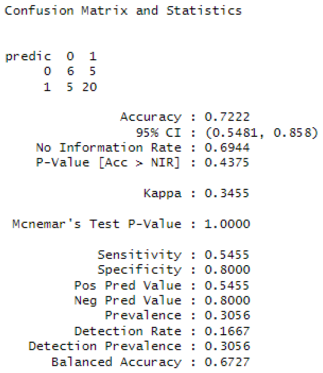


```
Call:
glm(formula = cardiac_condition.Absent ~ age + weight, family = binomial(),
    data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 14.54786    3.81499   3.813 0.000137 ***
age         -0.19393    0.05993  -3.236 0.001212 **
weight      -0.07458    0.02460  -3.032 0.002427 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.881  on 59  degrees of freedom
Residual deviance: 54.243  on 57  degrees of freedom
AIC: 60.243

Number of Fisher Scoring iterations: 5
```

Figure 39. Model 3 by Logistic Regression

E. Pattern Evaluation

The model has an accuracy of 72.22% vs the prediction, True Positive(TP) of 6, True Negatives (TN) of 20 predicted absent, False Negatives of one non-absent cardiac condition and 5 of False Positive (FP) absent. Sensitivity of 54.55% (Recall or True Positive Rate).



```
Confusion Matrix and Statistics


predic  0  1
     0  6  5
     1  5 20

               Accuracy : 0.7222
                 95% CI : (0.5481, 0.858)
    No Information Rate : 0.6944
    P-Value [Acc > NIR] : 0.4375

                  Kappa : 0.3455

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.5455
            Specificity : 0.8000
         Pos Pred Value : 0.5455
         Neg Pred Value : 0.8000
             Prevalence : 0.3056
         Detection Rate : 0.1667
   Detection Prevalence : 0.3056
      Balanced Accuracy : 0.6727
```

Figure 40. Confusion Matrix table

Finally, the ROC Curve visualizes 67.27% of the representation of classifier’s performance across all possible decision thresholds.

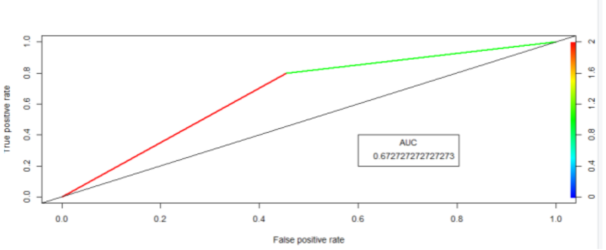


Figure 41. ROC Curve

Also, it applied Principal Component Analysis for dimension-reduction technique in the dataset by 3 principal components, this technique was not apply in the model because from the beginning the variables are uncorrelated.

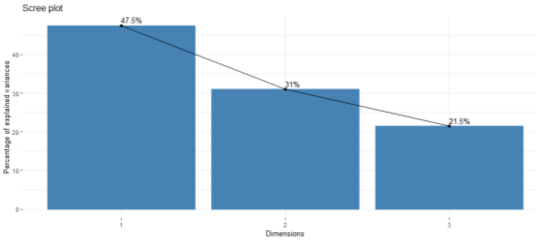


Figure 42. PCA

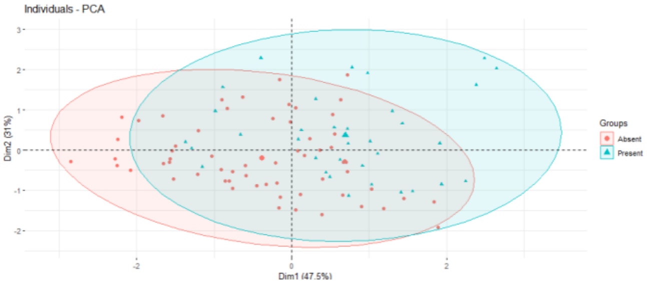


Figure 43. Visualization of components.

V. Conclusion

This project has been a full experience of learning, to know that the topic involves different areas, and there are many techniques for forecasting but at the same time, it is important to be clear on how to evaluate the model and to interpret the outcomes, one of the most difficult stuffs was to understand the evaluation of the model. On the other hand, the logistic regression model was a bit better to understand, thanks to the notes and the activities seen in the laboratories.

References

David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant. Applied Logistic Regression, Third Edition. Wiley, 2013.

Delen, Dursun. Predictive Analytics: Data Mining, Machine Learning and Dta Science for Practiotioners. Pearson FT Press, 2020.

—. Real-World Data Mining: Applied Business Analytics and Decision Making. Pearson, 2014.

Gharehbaghi, Arash. Deep Learning in Time Series Analysis. CRC Press, 2023.

Joseph, Manu. Modern Time Series Forecasting with Python. Packt Publishing, 2022.