# Analysis for Statistic: Time Series for Weather, Logistical Regression for Cadiac

Etinosa Eghaghe
*MSc in Data Analytics*
National Collegde of Ireland
x23138548@student.ncirl.ie

***Abstract*—**This report provides a thorough examination of two different statistical approaches used in two separate domains, which are Weather data using time series analysis and Cardiac health data using logic regression. Using the relevant datasets, the goal is to extract significant insight, patterns, forecast, and predictive models from the respective datasets using R programming language. The time series was conducted by using weather data from one of the Ireland weather stations(www.met.ie). It spans from 1942 t0 2023. After creating the time series, we work with a subset from 2019 to 2023. Simple, exponential models and ARIMA models were developed and evaluated. For the cardiac dataset with attribute of 100 participants, using the robust methodology of statistic a logic regression was developed. The assumptions were observed, and the goals were met, further advance statistical analysis can still be used to enhance and update these methods (time series and logistic regression)

***Keywords—Time Series, R Programming Language, EDA ARIMA, Logic Regression.***

## 1. INTRODUCTION

The aim of this analysis is to carry out a time series analysis, estimate and report suitable model. A weather dataset containing 9 variables is used, specifically focusing on the mean weed speed. That is to forecast Mean Wind Speed using time series forecasting techniques.
A time series can be thought of numbers, along with some information about what times those numbers were recorded [1]. The stationary is one the assumption of time series

On the other hand, this report presents an exploratory data analysis (EDA) and the development of a logic regression model using cardiac dataset. We estimate a binary logical regression to help understand the relationship between various variables. The objective is to understand the relationship between variables in the dataset and create a predictive model for the cardiac_condition.. a logical regression was created and evaluated.
Logic regression is a mathematical modelling approach which can be used to describe the relationship of many X's to a dichotomous dependent variable [2]. Several studies have used logic regression to explore the determinants of cardiac conditions [3]

## 2. METHODOLOGY

A detail time series and logic regression analysis, modeling and forecasting process with weather dataset and cardiac dataset.
The process for the Time series involved data exploration, cleaning, and preprocessing. Followed by the creation of a time series. Simple forecasting models (Average,Naïve,Seasonal, Drift), exponential smoothing models (SES, holt, Holt-Winters), and ARIMA models were implemented and evaluated Model comparisons were conducted based on accuracy metrics to identify the optimal forecasting models .

For the logic regression analysis and modeling, began with Same process exploration, visualization, cleaning and preprocessing, data transformation. Next building of logic regression model to predict a present or absence of cardiac condition. Evaluated and conclusion.

## TIME SERIES ANALYSIS

### 3 DATA EXPLORATION

**Dataset 1: Weather**
Began by installing the necessary packages and libraries needed for the analysis, modelling and forecast. Then followed by:
**3.1 Loading the Dataset(weather_reverse):** The dataset is loaded to jupyter notebook, from the location of storage.
**3.2 Descriptive Statistic:** Here the features of the dataset was examined, the dataset contains 9 variable and 29889 rows an initial exploration conducted to understand its structure and main variable, which include:
Checking the overview, head, summary, and dimension of the dataset.

```
wdsp.Mean.Winc          date
Min.   : 0.0    01/01/1942:    1
1st Qu.: 6.8    01/01/1943:    1
Median : 9.6    01/01/1944:    1
Mean   :10.2    01/01/1945:    1
3rd Qu.:13.0    01/01/1946:    1
Max.   :35.5    01/01/1947:    1
                (Other)   :29883
```

**Figure 1: Subset Summary**

### 4 DATA CLEANING AND PREPROCESSING

**4.1 Missing values:** This is a crucial stage in analysis because the decision made can influence the outcome of the analysis positively or negatively. For this time series analysis there

were 5 missing values in column 'gmin' and 2 missing values in 'evap', the columns were dropped because not needed for this time series analysis. Also, the columns deemed unnecessary for the time series analysis were removed.

```
                                                  date
                                                     0
 maxtp.Maximum.Air.Temperature...degrees.C.
                                                     0
 mintp.Minimum.Air.Temperature...degrees.C.
                                                     0
gmin.Grass.Minimum.Temperature...degrees.C.
                                                     5
             rain.Precipitation.Amount...mm.
                                                     0
                cbl..Mean.CBL.Pressure.hpa.
                                                     0
               wdsp.Mean.Wind.Speed...knot.
                                                     0
      pe.Potential.Evapotranspiration...mm.
                                                     0
                    evap.Evaporation..mm.
                                                     2
```

**4.2 Data Conversion:** The date column was converted to suitable date type to be used in further analysis. If the date is not in appropriate format can cause issue in creating time series. And wdsp column renamed. Figure 3

```
      date                wdsp
 Min.   :1942-01-01   Min.   : 0.0
 1st Qu.:1962-06-17   1st Qu.: 6.8
 Median :1982-12-01   Median : 9.6
 Mean   :1982-12-01   Mean   :10.2
 3rd Qu.:2003-05-17   3rd Qu.:13.0
 Max.   :2023-10-31   Max.   :35.5
```

**Figure 3: After Conversion and Renaming**

**4.3Visualization**

**Histogram:** The variables were visualized to get a better understanding of the distribution and any potential patterns or trends. Histogram generated for numeric variable As seen below in Figure 4.

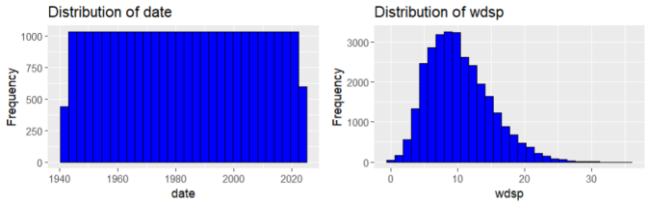


**Figure 4: Visualization of Distribution**

**4.4 Outliers:** outliers examined and corrected by using the interquartile Range (IQR) method. To get a better outcome in the time series analysis. Show in Figure 5 and 6 repectively.

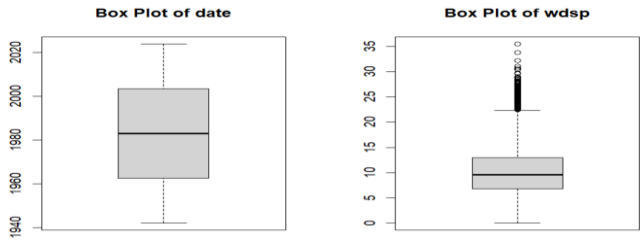


**Figure 5: Boxplot Before Outlier Removal**

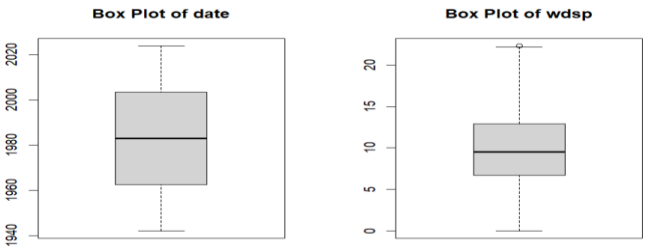


**Figure 6: Boxplot after Outlier Removal**

**4,5 Creating of Time Series**

Now a time series object was created for Mean Wind Speed, spanning from January 1942 to October 2023 with a frequency of 12. The dataset was then split into training sets start from 2019 and end in 2022, while testing sets 2023,01 to 2023, 10 for model evaluation. And plotted. Showed in Figure 7, 8 and 9.

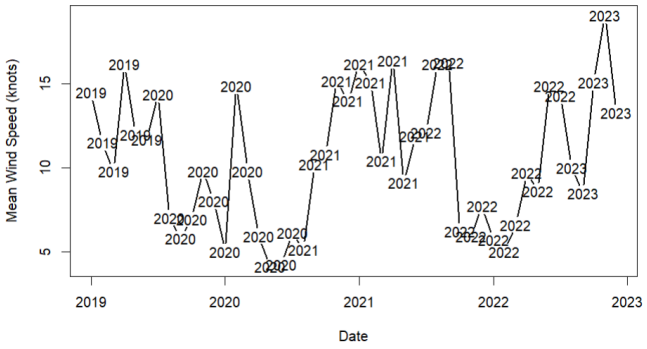


**Figure7: Train Sets**

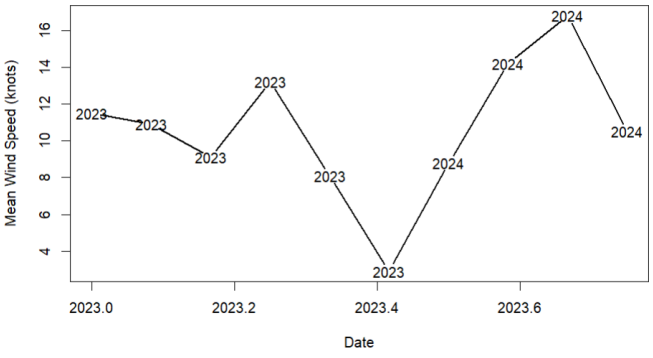


**Figure 8: Test Sets**

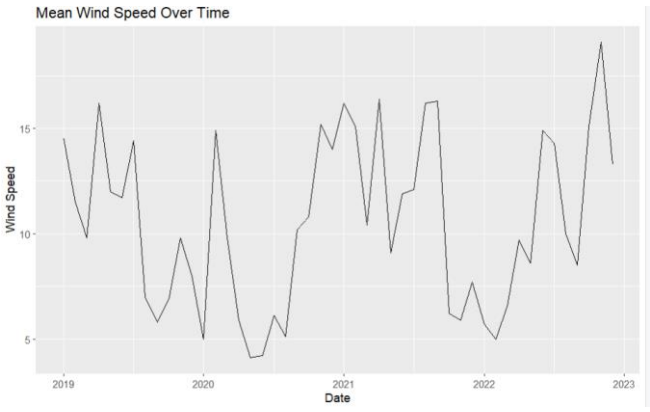


**Figure 9: Visualization of Raw Time Series**

## 5 MODELLING AND EVALUATION

### 5,1 Simple Model:

We fitted several basic models, Average(Figure10), Naïve(Figure11), Seasonal (Figure12), and Drift(Figure13) to the training set. Plotting of the model forecasts and computation of accuracy metrics were done. The model that performed the best at forecasting was determined to be the Average model. And then the Average model was used to evaluate the in the test data set.

| Model | RMSE | MAE |
|---|---|---|
| avg_model | 4.031094 | 3.484722 |
| naive_model | 4.046327 | 3.170213 |
| seasonal_model | 7.135008 | 6.266667 |
| drift_model | 4.046247 | 3.168583 |

**Table 1: Model Comparison Result**

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|---|
| Training set | 3.553059e-16 | 3.606938 | 2.72 | -24.17397 | 42.68041 | NaN | 0.2901614 |

**Average Model Evaluation Result**

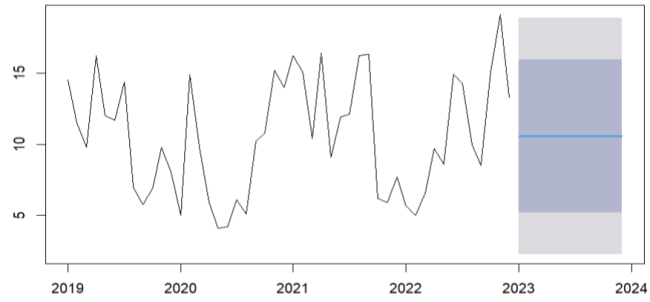


**Figure 10: Average Model**

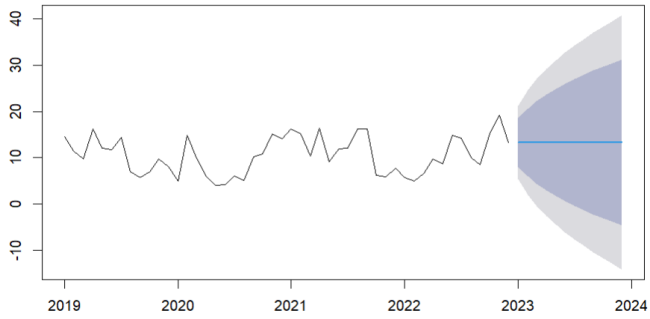


**Figure 11: Naïve Model**

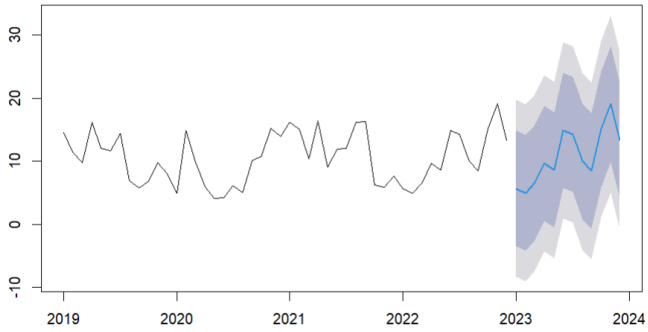


**Figure 12: Seasonal Model**

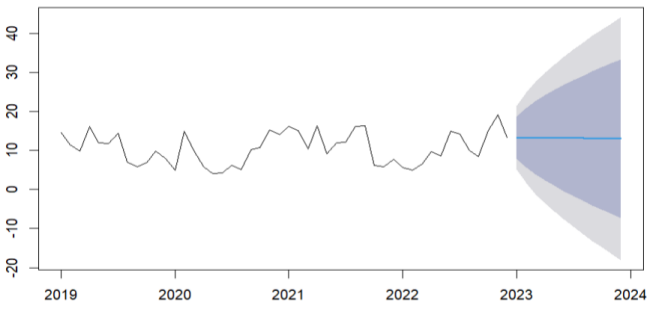


**Figure 13: Drift Model**

### 5.2 Exponential Models:

This is a type of forecasting technique that used past weights data from past period with exponentially reducing importance in the forecast that the recent data have more weight in the moving average [4]. This includes Simple Exponential Smoothing (ses), Holt, and Holt-Winters were applied. Model performances were compared, and the best model emerges for forecasting.

| Model | AIC | AICc | BIC |
|---|---|---|---|
| ANN | 318.4085 | 318.9540 | 324.0221 |
| AAN | 322.8090 | 324.2376 | 332.1650 |
| AAA | 339.5574 | 359.9574 | 371.3678 |

Table 2: ETS

| Model | RMSE | MAE | MAPE |
|---|---|---|---|
| ses | 3.738389 | 3.062714 | 33.29526 |
| Holt | 3.754018 | 3.062915 | 33.17630 |
| Holt-Winters | 3.480894 | 2.882052 | 31.80799 |

**Model Comparison Result**

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| -0.0003957713 | 3.607118 | 2.720107 | -24.17938 | 42.68324 | NaN | 0.2901626 |

**Table 3: ETS(A,N,N) Test Result**

**5.2.1 Simple Exponential Smoothing**: Is also known as a single exponential model. Simple Exponential Smoothing (SES) is used to project future data points based on historical observations. It is a widely used and rather easy approach, especially when a basic and clearly understood forecasting model is required. Different observations are given varying weights by SES, with more recent observations being given larger weights. Result from fit RMSE: 3.738, MAE: 3.063

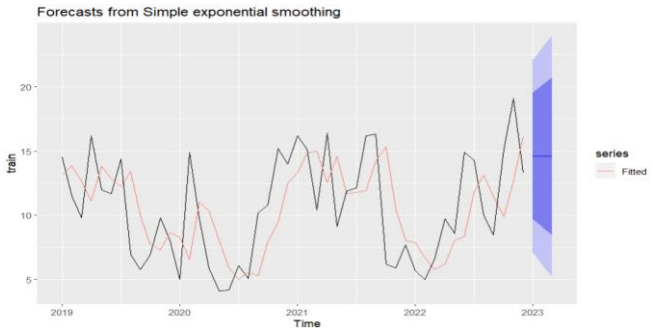


**Figure 14: Simple Exponential Smoothing**

5.2.2 Holt **Exponential Smoothing**:

Double exponential smoothing, sometimes referred to as SES-plus, is a variation of SES that adds a trend component to the level component. When there is a distinct pattern in the time series data but no seasonality, this approach might be helpful. Two smoothing parameters are used in the Holt's exponential smoothing method: β for the trend and α for the level (which are the same as in SES). The result from this RMSE: 3.754, MAE: 3.063

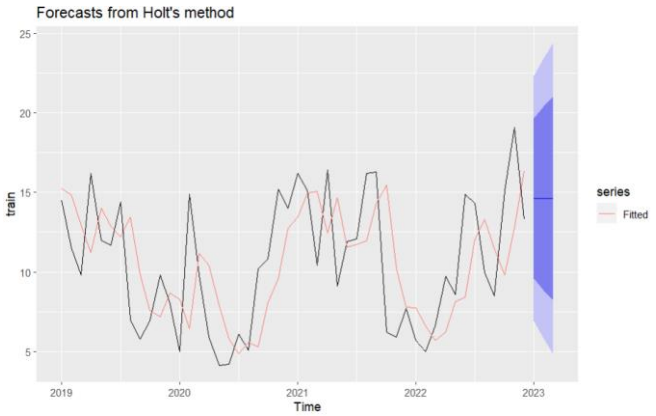


**Figure 15: Holt's Method**

**5.2.3 Holt-Winter Exponential Smoothing:**

This is an extension of holt's exponential smoothing that includes a seasonal component. It is suitable for time series data that exhibits trend and seasonality and can be used for both additive and multiplicative seasonality. Has three components: Level, Trend, and Seasonal. The result obtain from this analysis RMSE: 3.281 MAE: 2.882

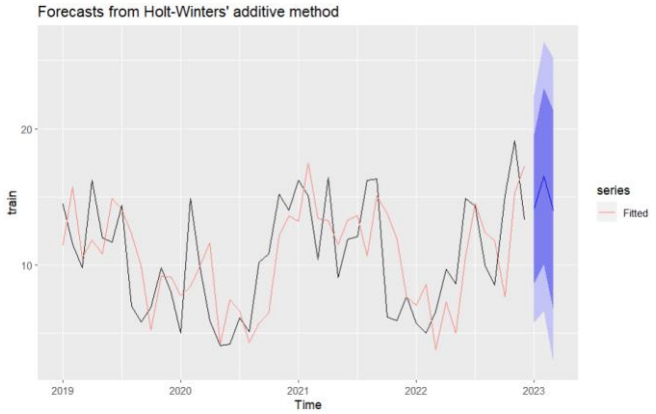


**Figure 16: Holt-Winter Additive Method**

**5.3 ARIMA Model:**

The begins with conducting Augmented Dickey-Fuller test to check for stationarity, applying Differencing to achieve stationarity. 4 ARIMA models were fitted, and the atu.arima function was used to identify the best model. To get a clear understanding of the best model, all the model performances were compared. The ARIMA model with the lowest RMSE was selected for forecasting.

```
        Augmented Dickey-Fuller Test

data:  train
Dickey-Fuller = -2.6288, Lag order = 3, p-value = 0.323
alternative hypothesis: stationary
```

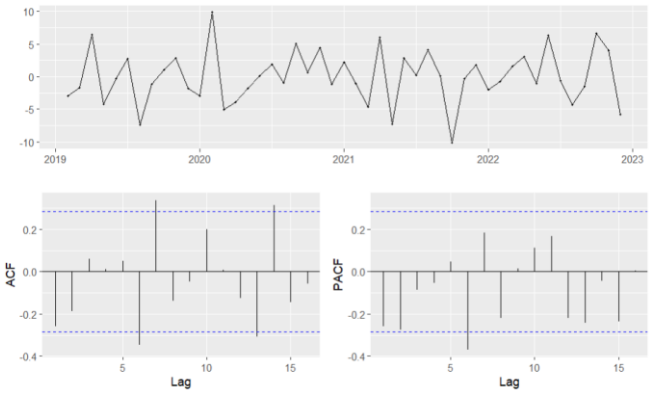


**Figure 17: Differencing to Achieve Stationarity**

| Model | RMSE | MAE | MAPE |
|---|---|---|---|
| model | 3.857733 | 3.097695 | 33.06685 |
| model2 | 3.491994 | 2.980392 | 33.48984 |
| model3 | 3.493595 | 2.953626 | 32.69254 |
| model4 | 3.700114 | 2.981050 | 32.04109 |

**Table 4: ARIMA Model Comparison Result**

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| -0.02635026 | 3.451981 | 2.700324 | -22.48338 | 40.98501 | NaN | 0.1644754 |

**Table 5: ARIMA MODEL2 Model Evaluation Result**

**5.4 Overall Model Comparison and Evaluation**

The best models from the three sections of modelling from Simple, Exponential, and ARIMA models. Which is Average model form simple models, Holt-Winters from Exponential models, model2 from ARIMA models. Was compared based on RMSE, MAE, and MAPE. Model2 from ARIMA was selected as the optimal model for forecasting with RMSE: 3.451981 based on the evaluation metrics, using indexing from the R programming language.

| Model | RMSE | MAE | MAPE |
|---|---|---|---|
| avg_model | 3.606938 | 2.720000 | 42.68041 |
| weatherwdsp_ses | 3.738389 | 3.062714 | 33.29526 |
| model2 | 3.451981 | 2.700324 | 40.98501 |

**Table 6: Overall Model Comparison**

## LOGISTICAL REGRESSION

### 3 DATA EXPLORATION

**Dataset 2: Cardiac**

**3.1 Loading the Dataset(Cardiac):** The necessary packages, including tidyverse, caret, pROC, were installed and loaded. Then work directory and data was loaded from the device to jupyter note book it was on csv format. Upon loading the dataset.

**3.2 Descriptive Statistics:** Here the summary and viewing of the data were carried out to know the relevant feature of the dataset. The dataset consists of 100 rows (instances) and 6 features, which are: caseno, age, weight, gender, fitness score, and cardiac condition. For this analysis cardiac condition column is the dependent variable and other as the independent variable.

| Variable | Data Type |
|---|---|
| caseno | Integer |
| age | Integer |
| weight | Numeric |
| gender | Factor |
| fitness_score | Numeric |
| cardiac_condition | Factor |

**Table 7: Variable Types**

## 4 DATA CLEANING AND PREPROCESSING

**4.1 Missing values Handling:** Next was to check if some data was missing because missing values affect the accuracy of the analysis. There are several causes of missing data, which can occur during data collection or absence of information. For the dataset used for this analysis there were no missing value after examining the dataset using missing function (is.na() in R). Also, a check for duplicate was done and there were no duplicate in the dataset. Below is a subset.

**4.2 Visualization**

**Histogram:** The variables were visualized to get a better understanding of the distribution and any potential patterns or trends, the visualization plays a crucial role to understanding the structure and characteristic of the dataset. Geom_histogram generated for numeric variable (caseno, age, weight,fitness_score) and Geom_bar plot for variables(gender, cardiac_condition) From cardiac_condition variable it is seen that the absent percentage is higher than the precent. As seen below.

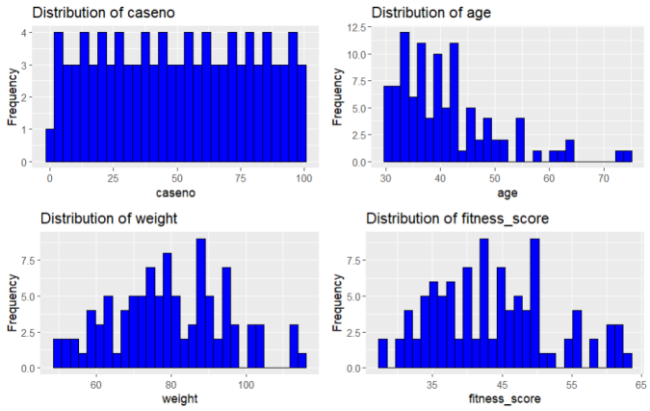


**Figure 18: Numerical Variables**

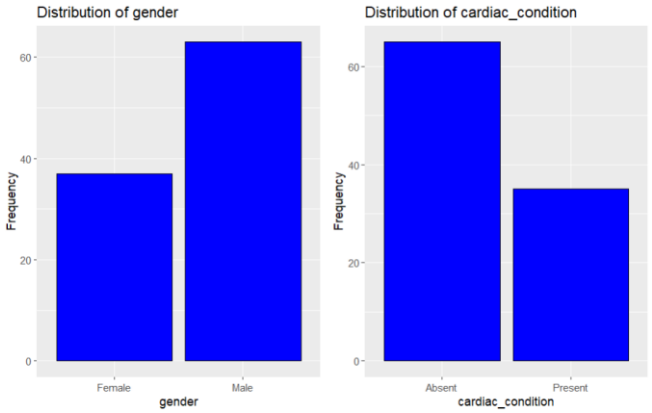


**Figure 19: Categorical Variables**

**4.3 Outliers Handling:**

Box plots were employed to detect potential outlier in numeric variables contributing to a robust data quality assessment. And linearity also checked.

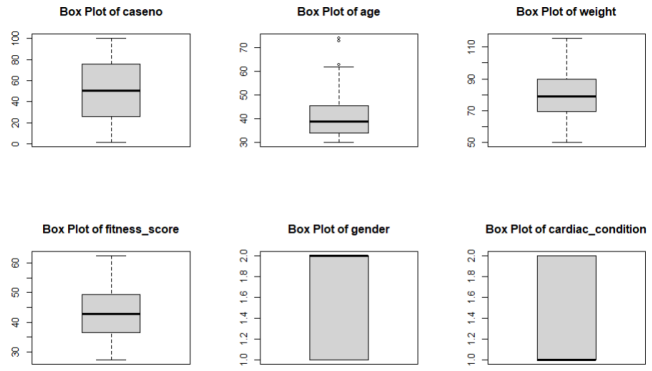


**Figure 20: Box Plot**

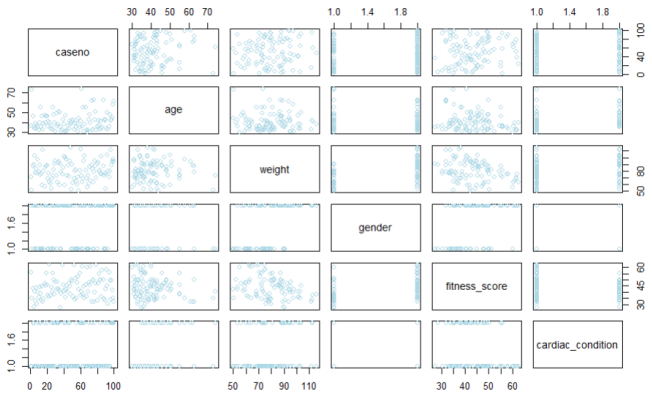


**Figure 21: Linearity checked.**

**4.4 Standardization:** This can also be refers to as normalization used to scale, numeric variables to get similar scale to prevent certain variables from disproportionately influencing the outcome of model training process. The variables are standardized to a scale of mean 0 and standard deviation of 1. The process was done by applying min-max normalization to ensure that there is consistency in scale.

$$Z = (X - mean(X)) / std(X)$$

## 5 MODEL DEVELOPMENT

A random seed set using the specified seed number. The dataset was split into two with a ratio of 75 percent for training and 25 percent for testing, resulting in train set to be 76 row and test 26 rows, 6 columns. This is a statistical method to ensure that model performance can be examined on data that it is not trained on to get better accuracy. Training data is used to train the model and test sets are you to evaluate the model performance.

**Logistic regression:** Is a way of classifying thing that are most time used to resolve problems with only two options [5] model was developed using cardiac condition as the dependent variable and other variables (caseno, age, weight, fitness score, gender) as independent variables. The model summary was generated to get a comprehensive overview of the performance of the model. With AIC:91.392 . as show in figure

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.249181   0.701172  -3.208  0.00134 **
caseno       -0.007851   0.274726  -0.029  0.97720
age           0.721149   0.308063   2.341  0.01924 *
weight       -0.224508   0.388654  -0.578  0.56350
genderMale    2.448090   0.940365   2.603  0.00923 **
fitness_score -0.897262   0.458731  -1.956  0.05047 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.898  on 75  degrees of freedom
Residual deviance: 79.392  on 70  degrees of freedom
AIC: 91.392
```

**Coefficients**

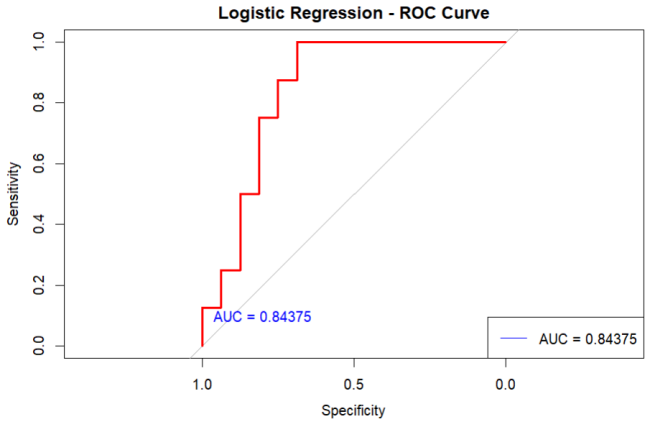


**Figure 22: ROC Curve**

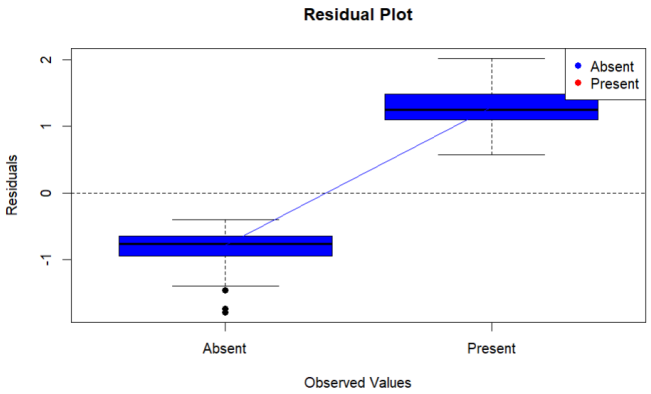


**Figure 23: Residual Plot**

## MODEL EVALUATION

Exhaustive examination of the level of the cardiac_condition variable in the test set was undertaking to comprehend the distribution of target classes.

Prediction on the test set was made using the logic regression model with resulting predicted labels transform into factors.

Detail confusion matrix was generated to provide a granular understanding of the model performance delineating trues positives, trues negatives, false positives and false negative.

Cross-validation results are obtain to asses the model performance figure below shows the result.

```
Confusion Matrix and Statistics

          Reference
Prediction Absent Present
   Absent      13       2
   Present      3       6

               Accuracy : 0.7917
                 95% CI : (0.5785, 0.9287)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.1383

                  Kappa : 0.5455

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.8125
            Specificity : 0.7500
         Pos Pred Value : 0.8667
         Neg Pred Value : 0.6667
             Prevalence : 0.6667
         Detection Rate : 0.5417
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.7812

       'Positive' Class : Absent
```

**Confusion Matrix and Statistics**

```
Generalized Linear Model

76 samples
 5 predictor
 2 classes: 'Absent', 'Present'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 60, 61, 61, 61, 61
Resampling results:

  Accuracy   Kappa
  0.6841667  0.2983165
```

**Cross Validation**

## 6. CONCLUSIONS

**Dataset 1: Weather (Time Series)**
In the time series analysis, the Model2 from ARIMA model demonstrates the best forecasting performance for Mean Wind Speed. This report serves as a detailed walkthrough of the data analysis and model selection process, enabling accurate and informed decision making for future forecasting endeavors.

**Dataset 2: Cardiac (Logic Regression)**
The logic regression model successfully leveraged features to predict cardiac condition.
The comprehensive confusion matrix offers valuable insights into the model efficacy laying the groundwork for further refinement optimization.
In conclusion, this report offers thorough exploration of the dataset, encompassing detailed analysis and the development and evaluation of a logic regression model for predicting cardiac conditions. The iterative nature of the analysis allows for ongoing improvements and insights into the predictive capabilities of the model.

## 7 RECOMMENDATIONS

Adding more characteristics or experimenting with advance statistical techniques might help forecasting models become even more refined, it is advised to regularly update the dataset and reevaluation of the forecasting models to maintain accuracy and relevance.

## REFERENCE


[1] P. Bhowmick, S. Das, and K. Mazumdar, Cognitive Cardiac Rehabilitation Using IoT and AI Tools (Advances in Medical Diagnosis, Treatment, and Care). IGI Global, 2023.
[2] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2018.
[3] D. G. Kleinbaum, Logistic Regression: A Self-Learning Text (Statistics for Biology and Health). Springer New York, 2013.
[4] J. S. Raj, I. Perikos, and V. E. Balas, Intelligent Sustainable Systems: Proceedings of ICISS 2023 (Lecture Notes in Networks and Systems). Springer Nature Singapore, 2023.
[5] J. K. Sharma, Business Statistics (Always Learning). Dorling Kindersley, 2012.