Time Series Analysis for Dublin Airport Weather and Logistic Regression for Cardiac Patient

Muddassir Ahmed
School of Computing
National College of Ireland
Dublin, Ireland
x23138688@student.ncirl.ie

***Abstract*—In this research paper we are performing times series analysis for weather dataset which is related to Dublin airport. This dataset consists of many attributes such as wind speed, air temperature and evaporation. We are calculating mean wind speed (knots) of dataset. This dataset is taken from Dublin Airport for year of 1942 to 2023. In this paper we are also doing logistic regression on cardiac dataset which contains 100 subject's details such as cardiac condition, fitness score, gender etc. This research shows relationship between different factors of variable and check cardiac disease is present or absent in subject. We applied logistic regression to analyzed the factors which connected to heart disease. Weather forecasting is checked by MSE and RMSE. Visualization also perform to better understanding of the dataset. Cardiac dataset is provided for logistic regression problem. It is used for supervised learning algorithms. It predicts the probability of binary output. Cardiac dataset include age, gender, blood pressure and physical activity. Cardiac condition showing either present or absent in dataset.**

***Index Terms*—Time Series Analysis, Exponential Smoothing, ARIMA/SARIMA, Logistic Regression, Dimensionality Reduction Techniques, Forecasting and Prediction.**

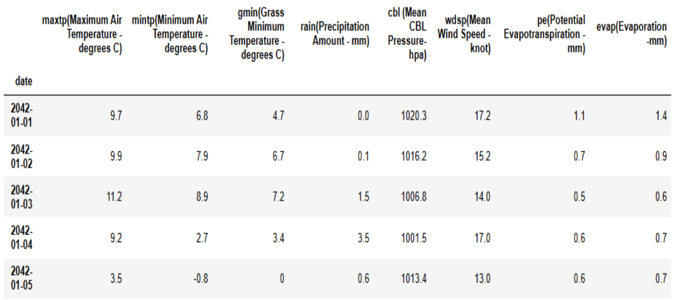
I. Introduction

Time Series Analysis

In this research our aim to apply time series model for weather dataset which is taken from Dublin airport Ireland from January 1, 1942 to October 31, 2023 which contains historical overview. Daily measurements such as wind speed, air temperature, precipitation and evaporation attributes are available in the dataset. We are calculating mean wind speed (knots) of dataset. Time Series analysis is used to analyse the data which is collected over the time. Three models such as simple time series model, exponential smoothing, and ARIMA/SARIMA will use for this problem. Weather forecasting will be performing on the year 2019 to 2022 and 2023 data will be used for performance evaluation. There are many models available for this type of problem such as MAE, RMSE.

II. Data Analysis

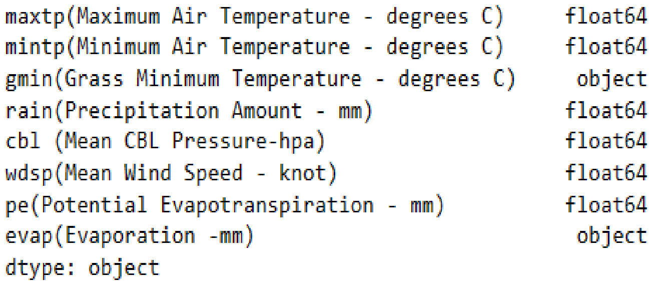
This section consists of data exploration, weather dataset of 29890 rows of data and it consist of these variables such as date, Maximum Air Temperature - degrees C, Minimum Air Temperature - degrees C, Grass Minimum Temperature - degrees C, Precipitation Amount - mm, Mean CBL Pressure-hpa, Mean Wind Speed - knot, Potential Evapotranspiration – mm and Evaporation -mm. For better understanding of dataset below diagram shows heads of dataset and first 5 rows of the dataset.



	maxtp(Maximum Air Temperature - degrees C)	mintp(Minimum Air Temperature - degrees C)	gmin(Grass Minimum Temperature - degrees C)	rain(Precipitation Amount - mm)	cbl (Mean CBL Pressure-hpa)	wdsp(Mean Wind Speed - knot)	pe(Potential Evapotranspiration - mm)	evap(Evaporation -mm)
date								
2042-01-01	9.7	6.8	4.7	0.0	1020.3	17.2	1.1	1.4
2042-01-02	9.9	7.9	6.7	0.1	1016.2	15.2	0.7	0.9
2042-01-03	11.2	8.9	7.2	1.5	1006.8	14.0	0.5	0.6
2042-01-04	9.2	2.7	3.4	3.5	1001.5	17.0	0.6	0.7
2042-01-05	3.5	-0.8	0	0.6	1013.4	13.0	0.6	0.7

Fig. 1. Image

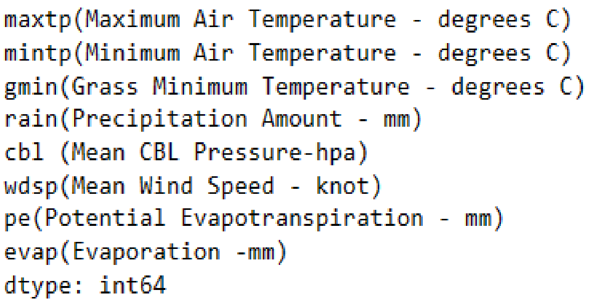
Now diagram shows data type of the dataset. Grass minimum temperature and evaporation has object data type and remaining variables has float values.



```
maxtp(Maximum Air Temperature - degrees C)      float64
mintp(Minimum Air Temperature - degrees C)      float64
gmin(Grass Minimum Temperature - degrees C)      object
rain(Precipitation Amount - mm)                 float64
cbl (Mean CBL Pressure-hpa)                     float64
wdsp(Mean Wind Speed - knot)                    float64
pe(Potential Evapotranspiration - mm)           float64
evap(Evaporation -mm)                            object
dtype: object
```

Fig. 2. Image

In this step, we are checking null values of dataset. Task is performed in python language and there are zero null values in dataset.



```
maxtp(Maximum Air Temperature - degrees C)      0
mintp(Minimum Air Temperature - degrees C)      0
gmin(Grass Minimum Temperature - degrees C)     0
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)                           0
dtype: int64
```

Fig. 3. Image

A. Pre-Modelling Assessment

For checking Mean Wind Speed(knot), dataset is filtered for the year of 2019 to 2022. Data is showing in month and years for calculation of mean wind speed. Plotting time series for each year. Blue lines show the data of 2019 to 2022 and it goes high in the month of march in every year. Orange line goes high in February and all other years showing in below graph. The highest temperature goes in each year in February.

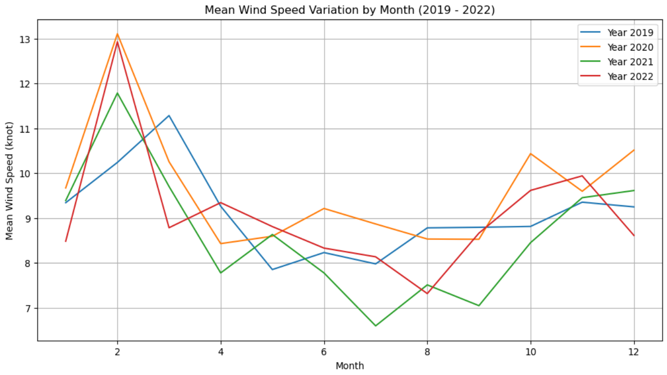
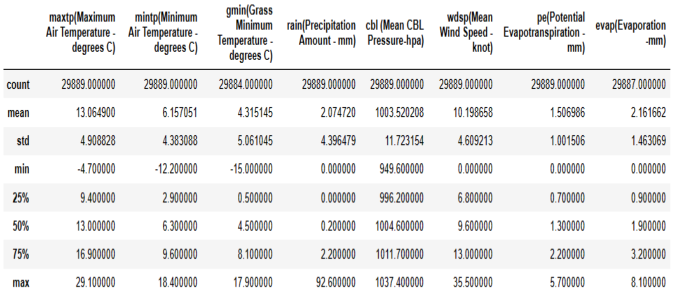


Fig. 4. Image

In this step we are converting object columns to float. Below diagram showing descriptive analysis in which contain count, mean, std, min, max, 25 percent, 50 percent, 75 percent of dataset.



	maxtp(Maximum Air Temperature - degrees C)	mintp(Minimum Air Temperature - degrees C)	gmin(Grass Minimum Temperature - degrees C)	rain(Precipitation Amount - mm)	cbl (Mean CBL Pressure-hpa)	wdsp(Mean Wind Speed - knot)	pe(Potential Evapotranspiration - mm)	evap(Evaporation -mm)
count	29889.000000	29889.000000	29884.000000	29889.000000	29889.000000	29889.000000	29889.000000	29887.000000
mean	13.064900	6.157051	4.315145	2.074720	1003.520208	10.198658	1.506986	2.161662
std	4.908828	4.383088	5.061045	4.396479	11.723154	4.609213	1.001506	1.463069
min	-4.700000	-12.200000	-15.000000	0.000000	949.600000	0.000000	0.000000	0.000000
25%	9.400000	2.900000	0.500000	0.000000	996.200000	6.800000	0.700000	0.900000
50%	13.000000	6.300000	4.500000	0.200000	1004.600000	9.600000	1.300000	1.900000
75%	16.900000	9.600000	8.100000	2.200000	1011.700000	13.000000	2.200000	3.200000
max	29.100000	18.400000	17.900000	92.600000	1037.400000	35.500000	5.700000	8.100000

Fig. 5. Image

B. Outlier Detection

In this step we identified outliers of the dataset and using boxplot for visualize of the dataset. We added gridlines for better readability of outliers and detected outliers based on Interquartile Range (IQR) by using descriptive statistics. Then we have calculated upper and lower bound for outliers. This is the result of the outliers.

pe(Potential Evapotranspiration - mm),
Number of Outliers: 98
evap(Evaporation -mm),
Number of Outliers: 21

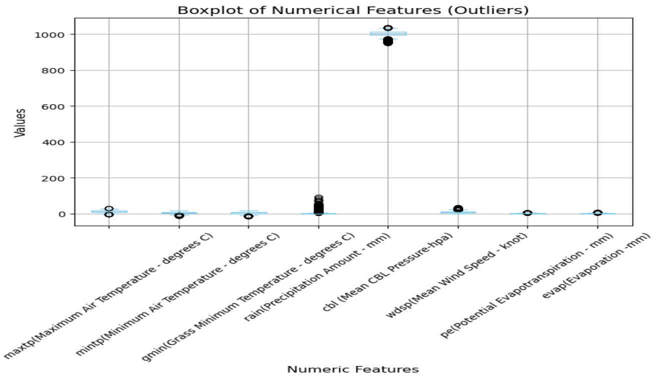
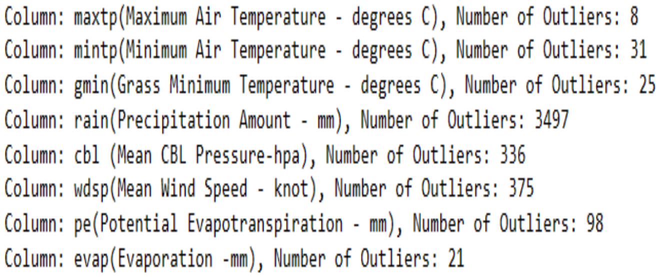


Fig. 6. Image

Column wise displaying outliers in below diagram



```
Column: maxtp(Maximum Air Temperature - degrees C), Number of Outliers: 8
Column: mintp(Minimum Air Temperature - degrees C), Number of Outliers: 31
Column: gmin(Grass Minimum Temperature - degrees C), Number of Outliers: 25
Column: rain(Precipitation Amount - mm), Number of Outliers: 3497
Column: cbl (Mean CBL Pressure-hpa), Number of Outliers: 336
Column: wdsp(Mean Wind Speed - knot), Number of Outliers: 375
Column: pe(Potential Evapotranspiration - mm), Number of Outliers: 98
Column: evap(Evaporation -mm), Number of Outliers: 21
```

Fig. 7. Image

C. Visualisation

In this graph showing data behaviour from 1970 to 2022. It is clearly displaying mean wind speed time series.

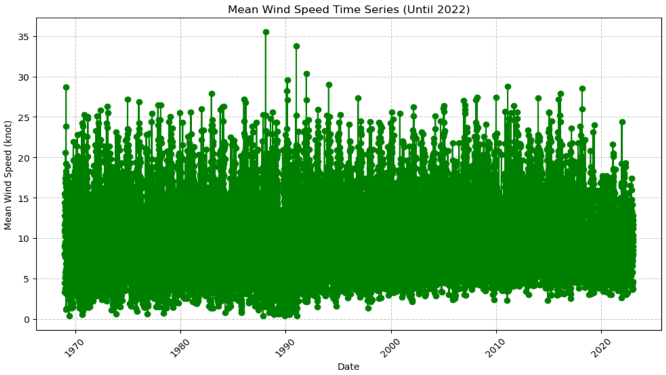


Fig. 8. Image

Perform seasonal decomposition on the 'wdsp' variable, which identified patterns. Plot shows the decomposed components using a different visual style. It is divided into four parts, original data, seasonality, trend and residual data. It is showing overall behaviours of the dataset and showing time series structure.

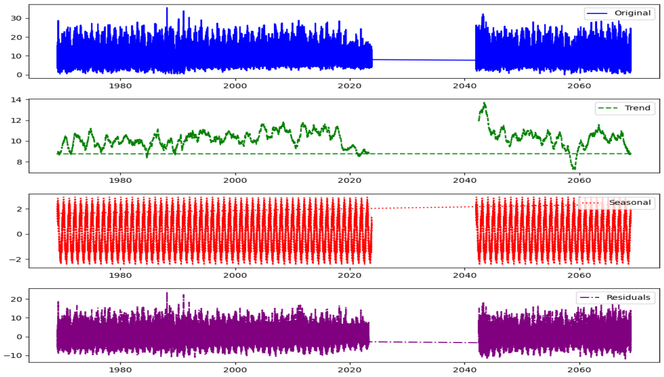


Fig. 9. Image

Now Calculating Autocorrelation Function (ACF) for 'wdsp' variable. ACF shows correlation of mean wind speed observation which is approximately 0.60. This is previous observation for current mean wind speed

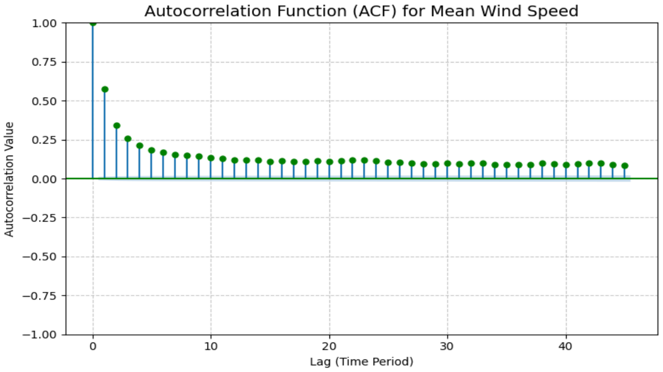
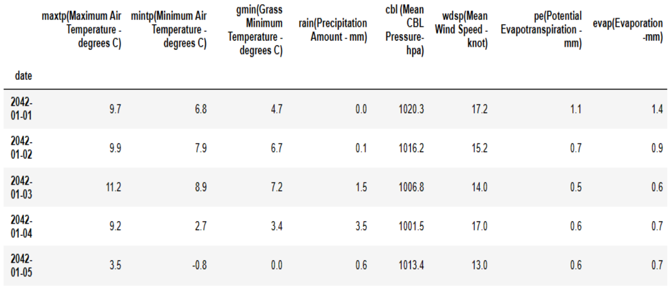


Fig. 10. Image

Again, displaying head of the variables and first five rows for updated dataset.



	maxtp(Maximum Air Temperature - degrees C)	mintp(Minimum Air Temperature - degrees C)	gmin(Grass Minimum Temperature - degrees C)	rain(Precipitation Amount - mm)	cbl (Mean CBL Pressure-hpa)	wdsp(Mean Wind Speed - knot)	pe(Potential Evapotranspiration - mm)	evap(Evaporation -mm)
date								
2042-01-01	9.7	6.8	4.7	0.0	1020.3	17.2	1.1	1.4
2042-01-02	9.9	7.9	6.7	0.1	1016.2	15.2	0.7	0.9
2042-01-03	11.2	8.9	7.2	1.5	1006.8	14.0	0.5	0.6
2042-01-04	9.2	2.7	3.4	3.5	1001.5	17.0	0.6	0.7
2042-01-05	3.5	-0.8	0.0	0.6	1013.4	13.0	0.6	0.7

Fig. 11. Image

III. Modeling

A. 1: Simple Moving Average

The dataset is split from year 2019 to 2022 for training and we will use 2023 year data for testing purpose. Simple moving average is applying for training purpose and checking mean wind speed. SMA is showing in red and providing representation of the training data.
Model Description:
The Simple Moving Average (SMA) model is used to calculate the average of previous observations for future value prediction.

Results:
The SMA model is represented graphically by red line that represents the average wind speed. Green lines showing testing data and blue lines showing training of the dataset

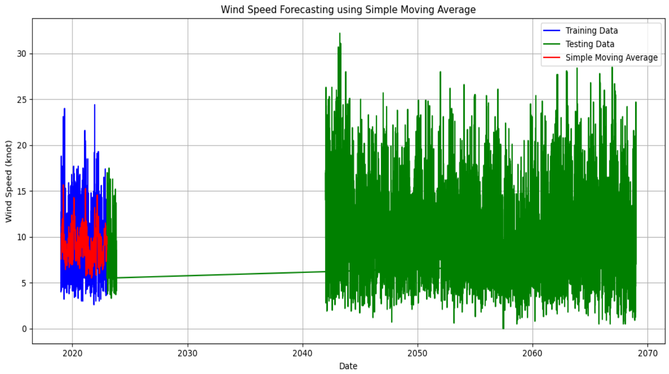


Fig. 12. Image

B. 2: Exponential Smoothing

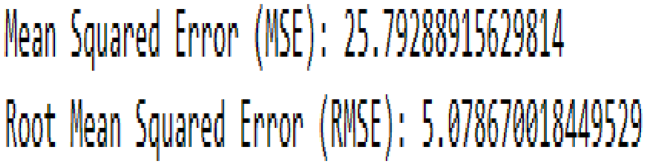
This model is used to forecast the wind speed. Mean square error and root mean square error is used to check accuracy of the model. Red lines showing exponential smoothing model. The obtained MSE value (25.79) and RMSE value (5.08) giving insights into the model's predictive performance.

Model Description:

The Exponential Smoothing is used for time series forecasting method which is assigns decreasing weights to previous observations.

Results:

The results of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are calculated to check the accuracy of the model.



```
Mean Squared Error (MSE): 25.79288915629814
Root Mean Squared Error (RMSE): 5.078670018449529
```

Fig. 13. Image

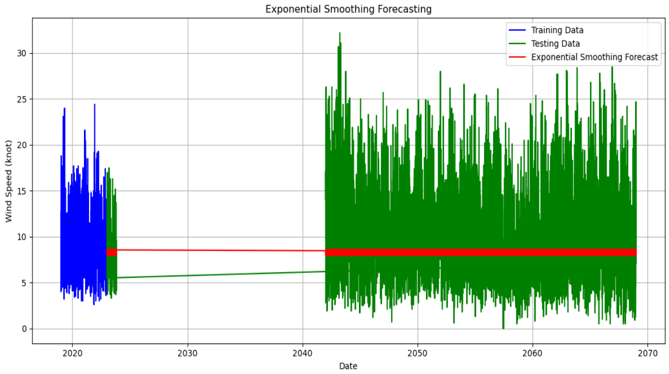


Fig. 14. Image

C. 3: SARIMA

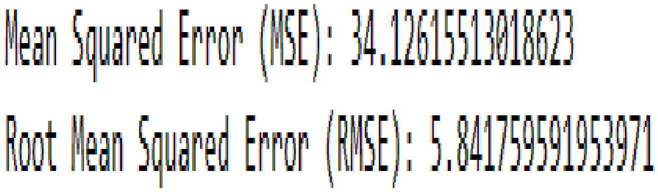
The Seasonal Autoregressive Integrated Moving Average (SARIMA) model we are using in our research for prediction of the wind speed. MSE and RMSE can calculate for performance evaluation. SARIMA forecast data is showing in red lines, testing data showing in green lines and blue lines are showing training data. Result of MSE (34.13) and RMSE (5.84) is giving accuracy of forcasting.

Model Description:

Seasonal Autoregressive Integrated Moving Average (SARIMA) model is next form of ARIMA mdel.

Results:

Red lines show SARIMA forecast and green lines showing testing data and blue lines showing training of the dataset. MSE value is 34.12 and RMSE is 5.84 as showing in below table.



```
Mean Squared Error (MSE): 34.12615513018623
Root Mean Squared Error (RMSE): 5.841759591953971
```

Fig. 15. Image

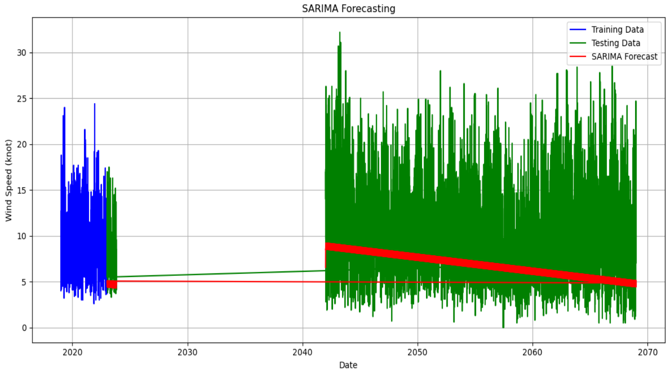


Fig. 16. Image

D. 3: Mean Forecasting Model

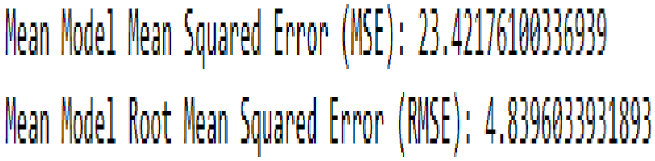
This model is use for calculating the mean of the training data and again it is used for the length of the test data. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values are (23.42 and 4.84) which is showing the performance of the model.

Model Description:

We use this model when mean of the training data is used as the forecast for the future.

Results:

Green line showing mean forecast and orange showing testing data and blue showing training of the data. Graph and results are showing below.



```
Mean Model Mean Squared Error (MSE): 23.42176100336939
Mean Model Root Mean Squared Error (RMSE): 4.8396033931893
```

Fig. 17. Image

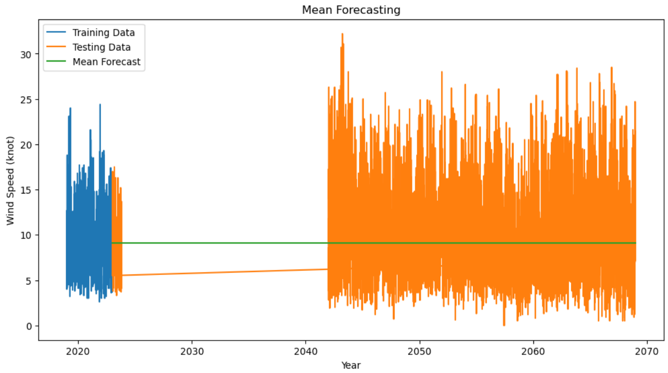


Fig. 18. Image

IV. Evaluation Results

A. Simple Moving Average (SMA)

SMA is baseline smoothing training data tread. SMA=12.01

B. Exponential Smoothing

This model capture wind speed patterns of the dataset. These matrices are used to quantify predictive accuracy. Evaluation Metrics: MSE = 25.79, RMSE = 5.08.

C. SARIMA

Evaluation Metrics: MSE = 34.13, RMSE = 5.84.

D. Mean Forecasting Model

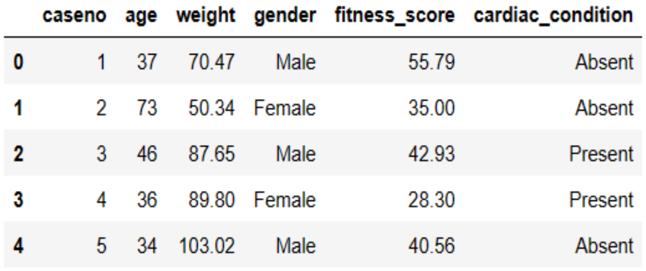
Evaluation Metrics: MSE = 23.42, RMSE = 4.84.

V. Logistic Regression

Cardiac dataset is provided for logistic regression problem. It is used for supervised learning algorithms. It predicts the probability of binary output. [3] Cardiac dataset include age, gender, blood pressure and physical activity. Cardiac condition showing either present or absent in dataset [1].

VI. Description of Dataset

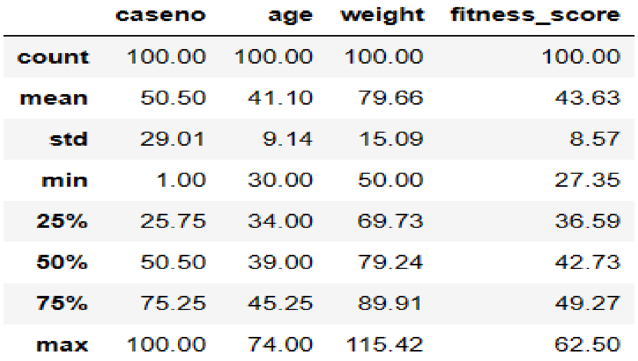
This dataset is showing variable names and first five rows of the dataset. Gender and cardiac condition is categorical data and remaining all variables are float or numeric data.



	caseno	age	weight	gender	fitness_score	cardiac_condition
0	1	37	70.47	Male	55.79	Absent
1	2	73	50.34	Female	35.00	Absent
2	3	46	87.65	Male	42.93	Present
3	4	36	89.80	Female	28.30	Present
4	5	34	103.02	Male	40.56	Absent

Fig. 19. Image

Complete descriptive of dataset is given below. There are total 100 rows which are showing in total count. Mean, std, min, max, 25 percent, 50 percent and 75 percent of dataset is also visible.

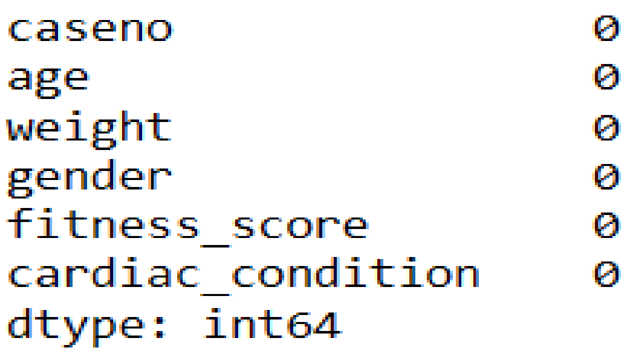


	caseno	age	weight	fitness_score
count	100.00	100.00	100.00	100.00
mean	50.50	41.10	79.66	43.63
std	29.01	9.14	15.09	8.57
min	1.00	30.00	50.00	27.35
25%	25.75	34.00	69.73	36.59
50%	50.50	39.00	79.24	42.73
75%	75.25	45.25	89.91	49.27
max	100.00	74.00	115.42	62.50

Fig. 20. Image

A. Null values checking

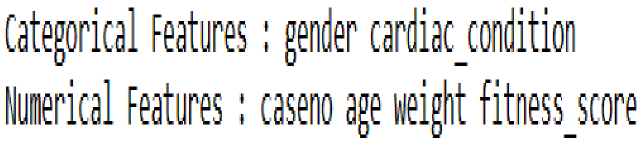
After checking the description of the dataset, next task is to check null values of the dataset. There are zero null values found in our dataset. Picture is given below.



```
caseno               0
age                  0
weight               0
gender               0
fitness_score        0
cardiac_condition    0
dtype: int64
```

Fig. 21. Image

Now we are checking which variables have numeric features and which are categorical features.



```
Categorical Features : gender cardiac_condition
Numerical Features : caseno age weight fitness_score
```

Fig. 22. Image

B. Visualization

Cardiac condition is showing present or absent in below diagram. Absent count is greater than 60 and present count is greater than 30 as per provided dataset. Bar plot is showing graphical picture of the dataset.

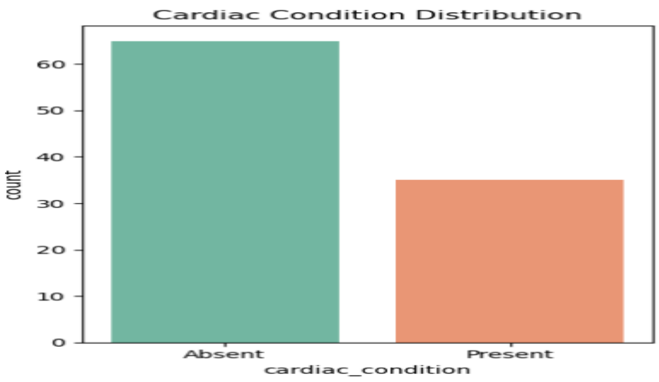


Fig. 23. Image

In this graph it is showing age wise cardiac conditions. Present cardiac patient are higher than 40 and absent are less than 40.

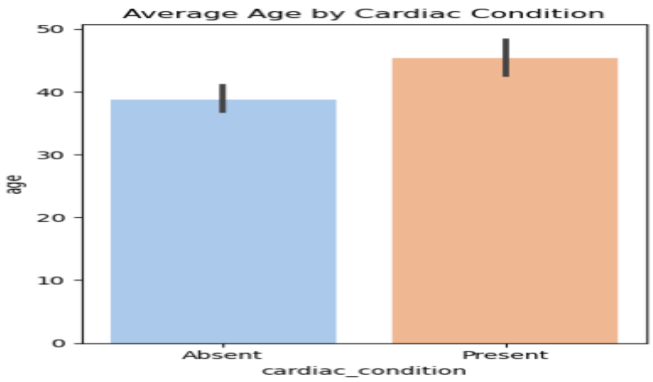


Fig. 24. Image

This diagram is showing average weight wise cardiac condition. There are more than 80 person which lies under present condition due to weight and less than 80 which are absent.

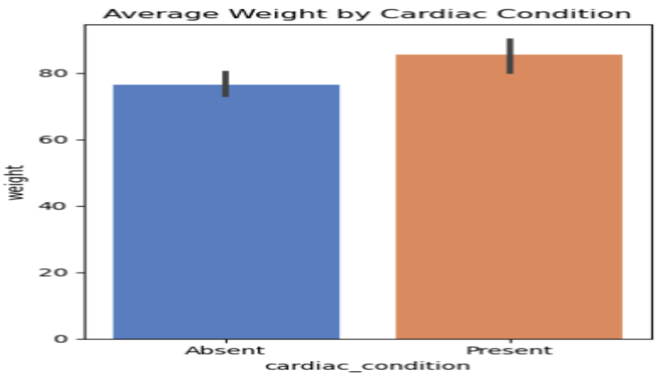


Fig. 25. Image

Pie plot showing Counting of occurrences of each category in cardiac condition, bar plot showing count. There is 65 percent ratio for absent cardiac condition and 35 percent ratio for present distribution in dataset.

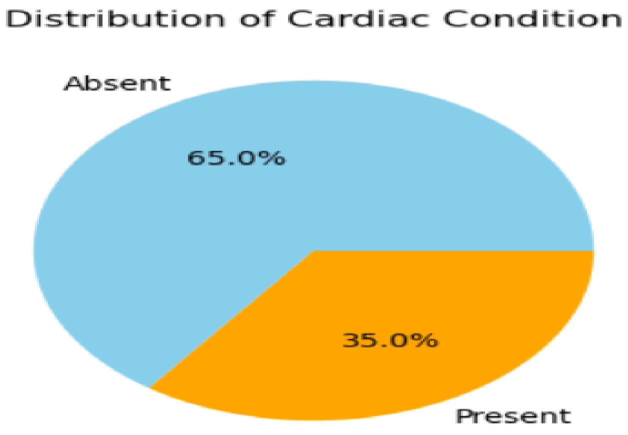


Fig. 26. Image

Below is Counting occurrences of cardiac conditions by gender. Skyblue color is showing present in gender wise and blue color is showing absent. Maximum cardiac patients are from female in dataset.

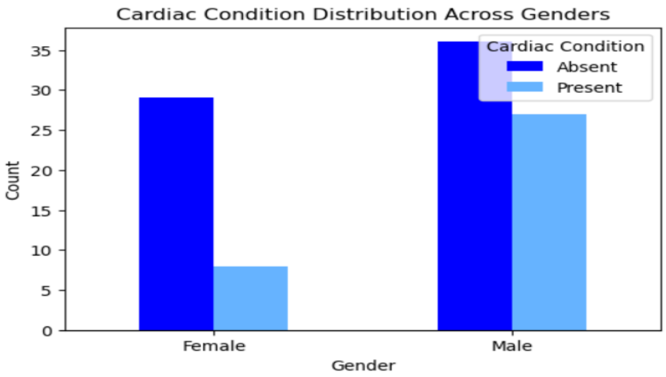
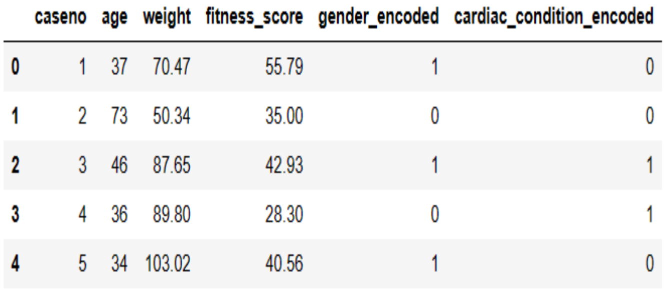


Fig. 27. Image

Label encoding applied in categorical features such as cardiac condition and gender. Now they are showing either 0 or 1 form.



	caseno	age	weight	fitness_score	gender_encoded	cardiac_condition_encoded
0	1	37	70.47	55.79	1	0
1	2	73	50.34	35.00	0	0
2	3	46	87.65	42.93	1	1
3	4	36	89.80	28.30	0	1
4	5	34	103.02	40.56	1	0

Fig. 28. Image

For this graph age distribution is visualize, lower age has maximum number of count. When age increase count gradually decreased.

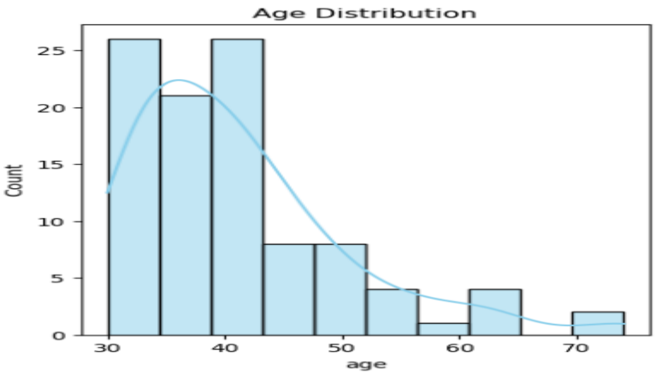


Fig. 29. Image

VII. Exploratory Data Analysis

Boxplot of numerical feature is showing outliers. There are total 120 count in y axis and four numerical variables are present in x axis. In x axis there are caseno, age, weight, and fitness score. Maximum number of the outliers are showing in age variable. There are zero outliers in other remaining three variables.

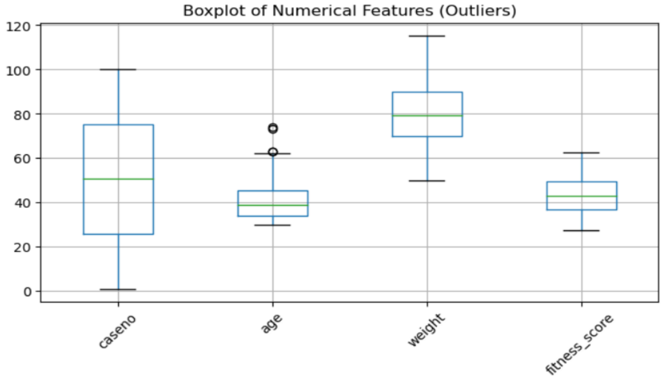
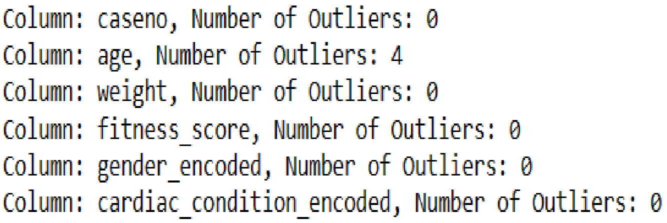


Fig. 30. Image

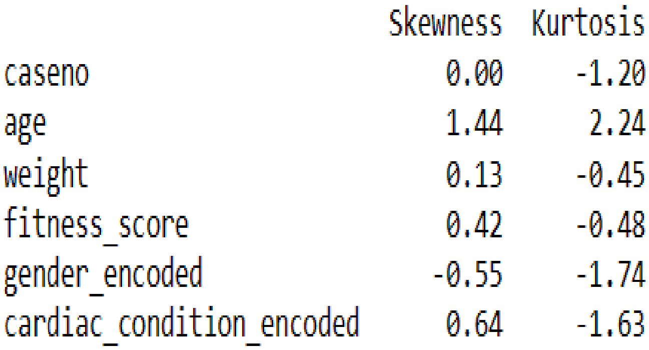
Below diagram shows calculation of the outliers.



```
Column: caseno, Number of Outliers: 0
Column: age, Number of Outliers: 4
Column: weight, Number of Outliers: 0
Column: fitness_score, Number of Outliers: 0
Column: gender_encoded, Number of Outliers: 0
Column: cardiac_condition_encoded, Number of Outliers: 0
```

Fig. 31. Image

Now we have computed skewness and kurtosis of the features from dataset. Age is showing 1.44 skewness maximum and gender is showing minus 0.55 which is less. Same kurtosis is also showing in below diagram.



```
                           Skewness  Kurtosis
caseno                         0.00     -1.20
age                            1.44      2.24
weight                         0.13     -0.45
fitness_score                  0.42     -0.48
gender_encoded                -0.55     -1.74
cardiac_condition_encoded      0.64     -1.63
```

Fig. 32. Image

Correlation heatmap is also using for better understanding of the dataset. Now our aim is to prepare balance dataset for training purpose [2]. We use seaborn and matplotlib for plotting of dataset. We are dropping cardiac condition, fitness score and caseno from dataset for creating x and y side balance. Further code is divided into training and testing purpose [4]. Prediction of cardiac condition is very crucial part in this research.

VIII. Modelling and evaluation

Variance Inflation Factors (VIF) is applied for checking multiple correlations. Const value in VIF is 50.12, gender is 1.38, age value is 1.01 and weight is 1.39. All these values checked using VIF in python code. Age has least value as compare to others variables.

Variable: const, gender encoded, age, weight

VIF: 50.12, 1.38, 1.01, 1.39

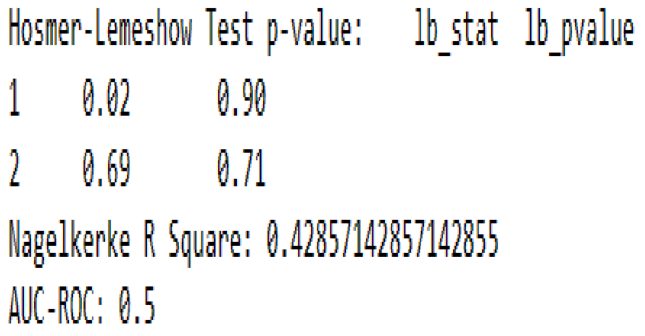
A. Model 1

Baseline model is used to predict the outcome of the provided data. First it is used for analysis purpose then we can check performance of the dataset. There are two types of baseline models which are used for classification. First one is Majority class classifier and second is Random classifier. Now we use baseline model which is predicting most frequent class in dataset. Baseline accuracy model is found 0.4286.

Baseline Model Accuracy: 0.4286

B. Model 2

Now we are applying logistic regression model which is use for supervised machine learning that complete binary classification tasks use to predict the outcome of the dataset. Result of the accuracy is is 0.6814 for training dataset and 0.7383 for testing purpose. Nagelkerke R square value shows 0.42 and AUC is 0.5. P values is also calculated in this model.



```
Hosmer-Lemeshow Test p-value:   lb_stat  lb_pvalue
1     0.02      0.90
2     0.69      0.71
Nagelkerke R Square: 0.42857142857142855
AUC-ROC: 0.5
```

Fig. 33. Image

C. Logistic regression model

Showing training accuracy which get predictions for training and testing data. Also showing testing accuracy which get accuracy for training and testing.

Training accuracy : 0.7243
Testing accuracy : 0.6943

Now plotting the results in Plot.

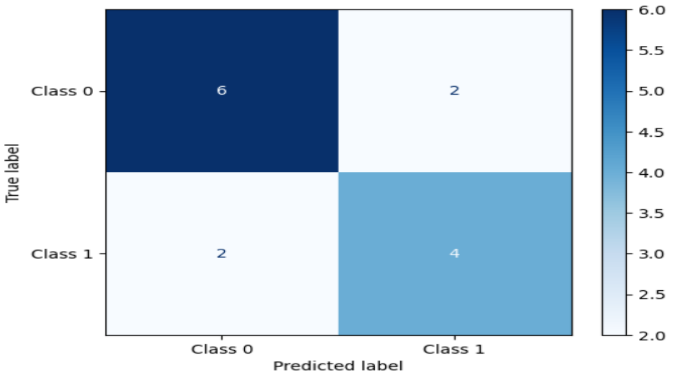
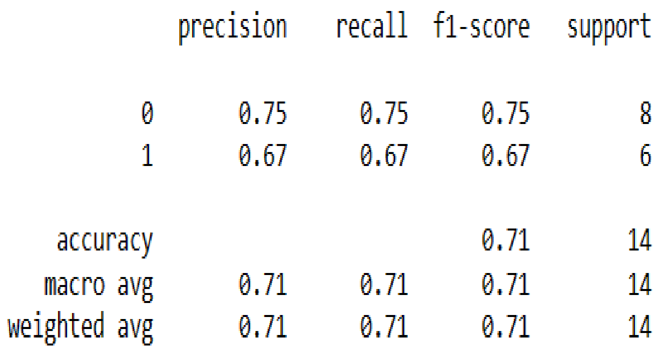


Fig. 34. Image



```
              precision    recall  f1-score   support

           0       0.75      0.75      0.75         8
           1       0.67      0.67      0.67         6

    accuracy                           0.71        14
   macro avg       0.71      0.71      0.71        14
weighted avg       0.71      0.71      0.71        14
```

Fig. 35. Image

IX. Final model performance

This results showing that all three models has some different calculation variables for outcome. Nagelkerke R square value shows 0.42 and AUC is 0.5. P values is also calculated in this model. Baseline Model Accuracy is 0.4286. Logistic regression Training accuracy is 0.7243 and Testing accuracy is 0.6943. It is showing lowest accuracy is got from nagelkerke and maximum accuracy is found in logistic regression. There are many factors which can be used to improve accuracy and performance of the model but logistic regression is performed very well.

X. Conclusion

The time series analysis and modeling efforts different key insights such as performance. Exponential smoothing is lower RMSR (5.08) as compare to SARIMA (5.84). Which is showing best accuracy predictive analysis. Mean forecasting model achieved RMSE (4.84) that is best baseline. Exponential smoothing gives best for accuracy that is used for wind speed calculation. SARIMA provide higher forecasting errors which is challenge in capturing the dataset.

References

[1] G Ambrish, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Kiran Mensinkal, et al. Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, 3(1):127–130, 2022.

[2] Sushmitha Kothapalli and SG Totad. A real-time weather forecasting and analysis. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 1567–1570. IEEE, 2017.

[3] Katherine Meadows, Richard Gibbens, Caroline Gerrard, and Alain Vuylsteke. Prediction of patient length of stay on the intensive care unit following cardiac surgery: a logistic regression analysis based on the cardiac operative mortality risk calculator, euroscore. *Journal of cardiothoracic and vascular anesthesia*, 32(6):2676–2682, 2018.

[4] Luís Sanhudo, Joao Rodrigues, and Enio Vasconcelos Filho. Multivariate time series clustering and forecasting for building energy analysis: Application to weather data quality control. *Journal of Building Engineering*, 35:101996, 2021.