Statistics for Data Analytics

Time Series Analysis and Logistic Regression

Dhruv Sharma
Masters of Science in Data Analytics , NCI
Dublin 1, Ireland
x22228268@student.ncirl.ie

***Abstract*— (Part A) This project involves exploration of wind speed data collected over several decades from the historical weather records. With the dataset being derived from Met Eireann's 'Dublin Airport' station, valuable insights can be gained by examining the patterns, trends, and fluctuations in wind speed. To effectively investigate these aspects, a combination of statistical and time series analysis techniques will be utilized, with special attention given to the Autoregressive Integrated Moving Average (ARIMA) modeling approach.**
(Part B) This project delves into the analysis and estimation of potential heart conditions in people using a comprehensive dataset that includes factors such as age, weight, gender, and fitness level. Utilizing advanced logistic regression techniques, the research employs a range of approaches, including traditional logistic regression, regularized methods (Lasso and Ridge), and dimensionality reduction through Principal Component Analysis (PCA). By conducting extensive exploratory data analysis, preprocessing, and model assessment, the project strives to provide valuable insights on the strengths and drawbacks of each model's predictive abilities.

Index Terms*— *Time series analysis, Logistic regression, ARIMA, PCA, Lasso and Ridge.

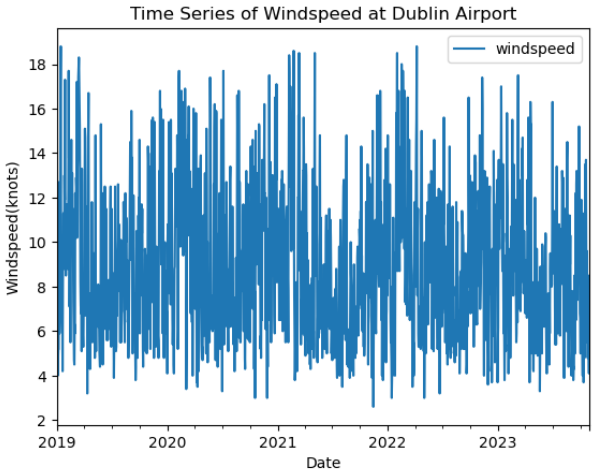
I. Introduction

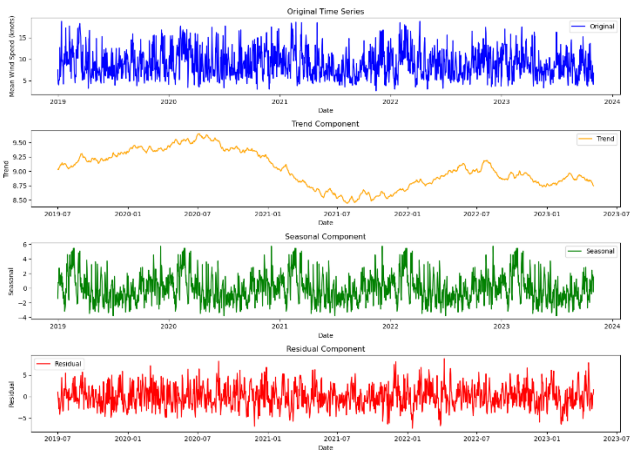
(Part A) The provided dataset is applied for exploring a time series, capturing daily Mean Wind Speed (measured in knots) data recorded at the 'Dublin Airport' weather station. Spanning from January 1, 1942, to October 31, 2023, this dataset holds valuable information to uncover. This analysis aims to unravel patterns, trends, and potential seasonal variations within the data, with the goal of creating a robust time series model that can accurately forecast future wind speeds.

II. Exploratory Data Analysis and Data Preprocessing

A. Descriptive statistics and levels of measurement

With a total of 1765 daily observations, the exploratory data analysis is conducted on the Mean Wind Speed dataset offering an insight into the historic weather patterns of the 'Dublin Airport' weather station. The data indicates an average wind speed of approximately 8.98 knots, highlighting the consistent presence of wind in the area. However, there is also a moderate level of variability, with a standard deviation of 3.23 knots, suggesting that the winds are not always predictable. Interestingly, the distribution of wind speeds appears to be symmetrical, with the median closely matching the mean, as revealed by the 25th, 50th, and 75th percentiles. The recorded wind speeds range from a minimum of 2.6 knots to a maximum of 18.8 knots, with the central 50% of the data falling within a spread of 4.7 knots, represented by the interquartile range. These findings provide valuable insights into the characteristics of the dataset. This exploratory data analysis lays the groundwork for a thorough examination of mean wind speeds by examining their central tendency, variability, and extremes. By delving into this extensive historical weather dataset, the author gains a comprehensive understanding of any potential trends and patterns. Dynamic visualizations like histograms and time series plots will further enrich our insight into the distribution and temporal patterns of the data.





Analysis of windspeed data shows a distinct trend with long-term variations. This trend does not follow a consistent upward or downward direction, but rather exhibits fluctuations over time. This indicates that windspeed is subject to variations that do not adhere to a uniform long-term pattern. Moreover, the data also reveals a prominent seasonal component, demonstrating a recurring pattern within each year. This is in line with expectations, as windspeed typically displays distinct seasonal variations. The observed pattern may be driven by various factors such as weather changes, annual climatic cycles, or other periodic phenomena. The model has expertly accounted for both trend and seasonality in the data, leaving behind residual noise that

consistently hovers around zero. This noise lacks any distinguishable pattern, showcasing the model's ability to accurately capture the predictable information in windspeed variations. Therefore, the model has effectively eliminated any systematic patterns from the residuals, resulting in a more random and uncorrelated component.

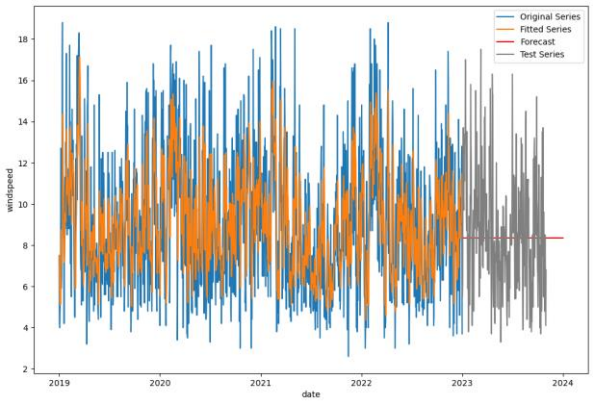
B. Data Preprocessing

The 'weather.csv' file was used to load the weather dataset, with special attention given to the 'date' column as it was parsed as datetime and used as the index for the data frame. To identify outliers in the 'windspeed' variable, z-scores were calculated. Any data points with z-scores that exceeded 3 or fell below -3 were classified as outliers and consequently eliminated. This step effectively reduces the impact of extreme values on the performance of the model. To address any missing values, a linear interpolation technique was implemented. This strategy guarantees the smooth interpolation of time series gaps, resulting in a comprehensive and contiguous dataset for further analysis. To ensure accurate modelling of time series data, the time index frequency was explicitly designated as 'D' (daily), guaranteeing a regular interval between observations. This step was undertaken for time series models that depend on consistency. Alongside this, all 'windspeed' values were converted to positive magnitudes to promote uniformity and facilitate interpretation. With these preparatory measures in place, the dataset was curated and harmonized, setting the way for the next stages of time series analysis, including model selection and evaluation.

The author carefully designates specific date ranges for both training and test data, extracting these subsets from the meticulously cleaned time series data. To ensure precision and accuracy, the shapes of both datasets are then printed and compared. Such segregation is crucial as it allows for effective training of models using past data (training set) and unbiased evaluation of their performance on future, unseen data (test set). These results align perfectly with the expected outcome as dictated by the defined training (2019-2022) and testing (starting from 2023) dates. Furthermore, the number of rows in the training set reflects the number of days between January 1, 2019, and December 31, 2022, while the number of rows in the test set corresponds to the days from January 1, 2023, and onwards. Overall, this process ensures the reliability and validity of the data used for model training and testing.

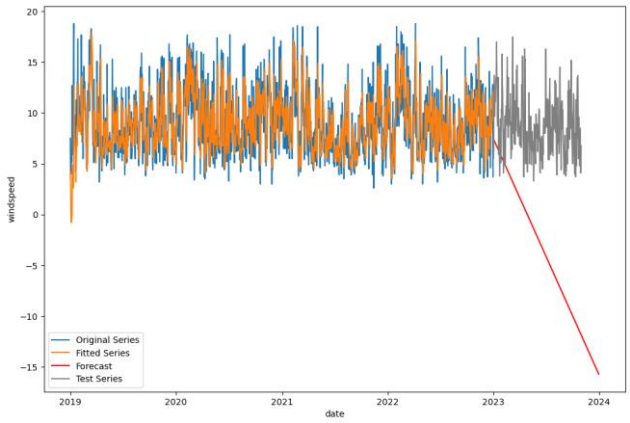
III. MODELLING AND DIAGNOSTICS

A. Model 1 : Simple Exponential Smoothing



After evaluation using various metrics, the Simple Exponential Smoothing (SES) model has demonstrated its performance. The Sum of Squared Errors (SSE), which measures the deviations between actual and predicted values, has been accurately calculated at 13499.169. Furthermore, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both crucial metrics for assessing goodness of fit and model selection, have been determined to be 3252.544 and 3263.117, respectively. The SES model does not consider trends or seasonal patterns, as evidenced by its flat forecast. The initial level, representing the starting point of the level component, has been identified as 7.5. The author effectively optimizes the crucial smoothing parameter, alpha, by assigning a weight of 0.601 to the most recent forecast. This is achieved by setting the parameter to approximately 0.399. The visual representation of the model, created using Seaborn's line plot, showcases the training data, fitted values, forecasted values, and actual test data in a cohesive manner. In-depth analysis of the model's accuracy is provided through metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), with corresponding values of approximately 2.416, 3.040, and 0.300.

B. Model 2 : Double Exponential Smoothing or Holt's Linear Trend



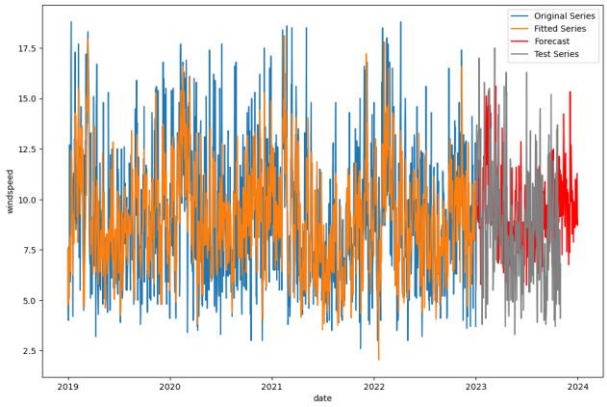
The effectiveness of the Double Exponential Smoothing or Holt's Linear Trend model in capturing trends within windspeed time series data has been thoroughly evaluated. Outcomes of this evaluation include a Sum of Squared Errors (SSE) of 14638.388, which represents the squared differences between actual and predicted values. Furthermore, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), known for assessing goodness of fit and model selection, report values of 3374.913 and 3396.060 respectively. Within this model, the presence of an additive trend is highlighted by the coefficients for smoothing level (alpha), smoothing trend (beta), initial level (l.0), and initial trend (b.0).

Significantly, the smoothing of the level and trend components is heavily impacted by the calculated alpha and beta values, which come to 0.5468227 and 0.0387495, respectively. To establish a baseline, the initial level and trend at the start of the time series are fixed at 7.5000000 and -3.3000000. When observing the data through Seaborn's line plot, the original training data alongside fitted values from the Holt model, forecasted values, and the actual test data, each represented by a different color. As observed, the model's forecast reveals a noticeable downward trend in windspeed data. Performance metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), further underscore the model's accuracy, yielding respective values of approximately 2.489, 3.165, and 0.304.

In terms of performance, it can be observed that the Simple Exponential Smoothing (SES) and Holt models show notable

similarity. However, upon closer inspection, SES has a slightly more favorable fit to the training data, as evidenced by its marginally lower MAE and RMSE values. This indicates that SES outperforms the Holt model in terms of both absolute and squared errors. Additionally, the closely aligned MAPE values for both models suggest a relatively uniform average percentage difference between the predicted and observed values.

C. *Model 3 : Triple Exponential Smoothing*

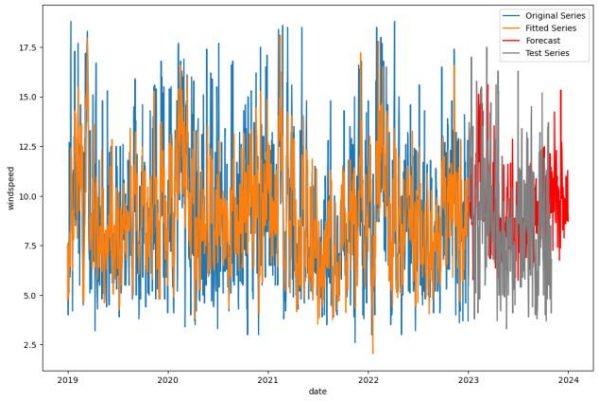


The Triple Exponential Smoothing or Holt-Winters Exponential Smoothing model revealed optimized parameters and comprehensive evaluation metrics. The Smoothing constant (alpha) is determined to be 0.3179703, while the Smoothing Trend (beta) and Smoothing Seasonal (gamma) constants are 1.8589e-06 and 8.4372e-06, respectively. The Sum of Squared Errors (SSE) is calculated as 10075.847, providing insight into the squared differences between actual and predicted values. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are 3559.220 and 5510.078, respectively, serving as indicators of model goodness of fit and selection.

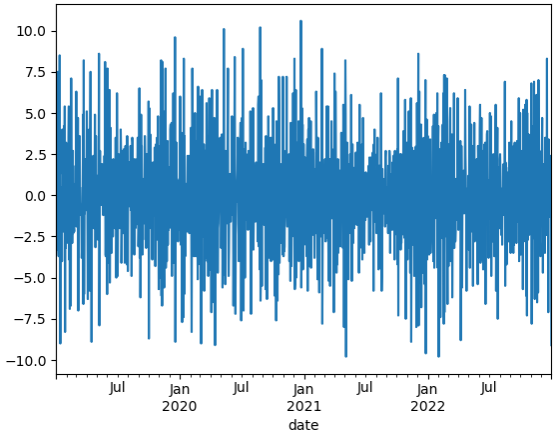
Seaborn's line plot skillfully displays a comprehensive view of the original training data, fitted values generated from the innovative Triple Exponential Smoothing model, predicted values, and the actual test data all in one cohesive plot. This model effectively incorporates both trend and seasonal factors by using an additive approach, where the trend component captures the overall long-term movement or pattern in the data and the seasonal component captures repeating patterns at fixed intervals (such as 365, indicating a yearly cycle). By visualizing these components, the plot serves as compelling evidence of the model's ability to accurately capture and forecast both the trend and seasonal patterns, resulting in a seamless continuation of these observed patterns.

IV. Investigation of Suitable ARIMA Models

A. *Stationarity Check*



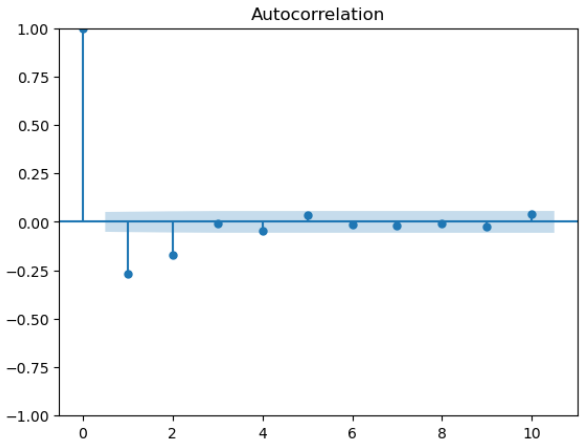
To determine the stationarity of the "windspeed" time series, the author used the Augmented Dickey-Fuller test in the Stationarity Check. Our findings revealed a remarkably negative test statistic and an extremely low p-value, resulting in the rejection of the null hypothesis with a confidence level of 95%. This outcome led to accepting the alternative hypothesis (H1), suggesting that the time series remains stationary without any noticeable variations in its mean or variance over time.

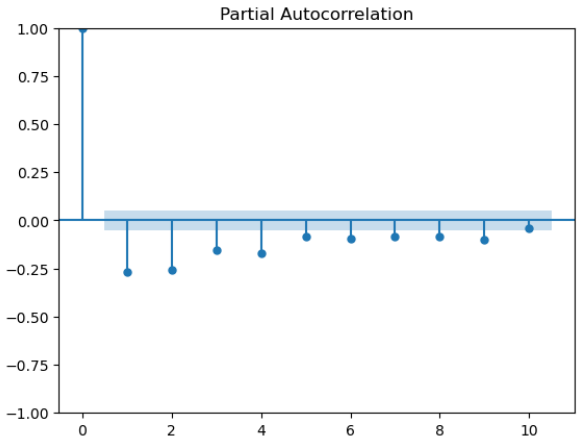


Furthermore, upon examining the first differenced series, the model obtained a p-value of less than 0.001, leading to the rejection of the null hypothesis and the confirmation of the stationarity of the first differenced series.

B. *Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) Analysis*

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) Analysis. These powerful plots are essential for determining the appropriate components to include in an ARIMA model. The ACF plot illustrates the correlation between the time series and itself at different lags, providing insight into the Moving Average (MA) component. Additionally, the PACF plot showcases the partial correlation at various lags, allowing for a deeper understanding of the Autoregressive (AR) component.

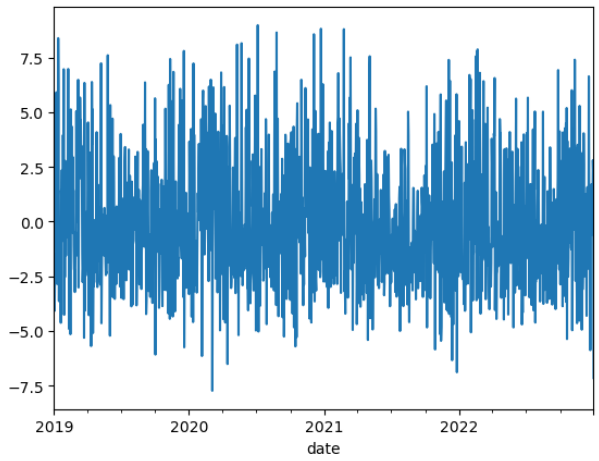




Based on the information presented in these charts, the model parameters are determined. Specifically, an ARIMA model with AR (3) indicates a direct relationship with the most recent three values in the series. This model can be represented by the notation ARIMA (p, d, q), where "p" represents the order of the Autoregressive (AR) component, "d" denotes the level of differencing needed for stationarity, and "q" signifies the order of the Moving Average (MA) component. By interpreting these charts, appropriate values for "p," "d," and "q" can be chosen to construct an impactful ARIMA model that aligns with the unique characteristics of the "windspeed" time series.

C. Model Selection

When deciding on a time series model, it is important to carefully consider the reasoning behind selecting both the type of model and its order parameters. This is crucial in constructing an effective forecasting model. The Akaike Information Criterion (AIC) serves as the primary criterion for model selection, with lower AIC values indicating a superior balance between model fit and complexity. In this analysis, the author chose the model with the lowest AIC value as our preferred option. This model, ARIMA (1, 0, 2), strikes the perfect balance between capturing underlying patterns in the data and maintaining simplicity. The AIC value for this model is 7205.63, cementing its position as the best-fit model among the other alternatives considered.



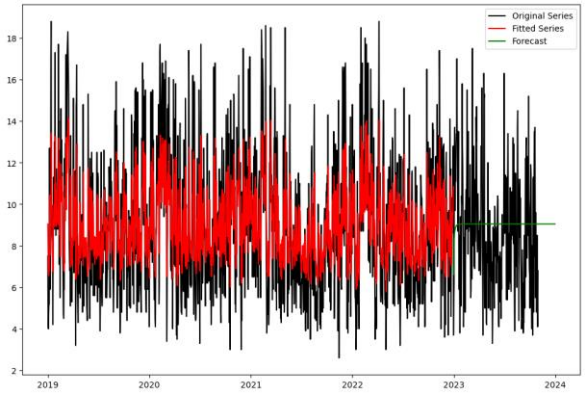
Breaking down the components of ARIMA (1, 0, 2), each number within the parentheses corresponds to a specific aspect of the model: In the Autoregressive (AR) model, the order is represented by "1" (p=1), indicating that the AR component incorporates the first lag of the series. This illustrates the linear relationship with the most recent past value of the time series. The differencing order is specified as "0" (d=0), indicating that no differencing is applied. This signifies that the original series is already stationary and does not require differencing. In the Moving Average (MA) component, the order is denoted by "2" (q=2), indicating that the model considers the first two lags of the forecast errors. This accounts for any correlation between consecutive forecast errors and captures any residual patterns that may exist.

Ultimately, the ARIMA (1, 0, 2) model stands out as the optimal choice for predicting the "windspeed" time series. This can be attributed to its impressive AIC results that strike a perfect balance between precision and simplicity, making it the perfect fit for accurate forecasting.

V. Forecasting and Evaluation of the Adequacy of the Final ARIMA Model 102.

A. Model Fitting

In the Model Fitting process, the chosen ARIMA (1, 0, 2) model was trained and validated to ensure its accuracy and reliability in forecasting the "windspeed" time series. The fitting process involved estimating the model parameters, which provided insights into the underlying patterns and relationships within the data. The ARIMA (1, 0, 2) model captures several key parameters: The intercept in this model is approximately 9.0469, representing the starting point or typical level of the data. The AR coefficient at lag 1 is 0.7698, revealing a positive connection between the current value and the previous one, suggesting a lingering effect from recent data. The MA coefficients at lags 1 and 2 are -0.2861 and -0.1628, respectively, capturing the influence of past errors in explaining the current value. The variance of the residuals, 8.0599, reflects the spread or diversity of the model's estimates compared to the actual values observed.



The author conducted a thorough analysis of our model's performance by utilizing various diagnostic tests. These tests included the Ljung-Box Q test for autocorrelation in residuals, the Jarque-Bera test for normality, and the Heteroskedasticity test. Results from these tests indicate that the residuals are not significantly autocorrelated, do not demonstrate any major normality concerns, and do not exhibit heteroskedasticity. The moderate skewness of 0.50 suggests a reasonable balance between the two extremities. On top of that, the model also presents a good fit, with a log likelihood of -3597.813 and an AIC of 7205.626. Furthermore, the low p-values for the model coefficients further validate the reliability of our estimated parameters.

SARIMAX Results

Dep. Variable:	windspeed	No. Observations:	1461
Model:	ARIMA(1, 0, 2)	Log Likelihood	-3597.813
Date:	Mon, 01 Jan 2024	AIC	7205.626
Time:	12:12:50	BIC	7232.060
Sample:	01-01-2019 - 12-31-2022	HQIC	7215.487
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	9.0469	0.197	45.850	0.000	8.660	9.434
ar.L1	0.7698	0.054	14.361	0.000	0.665	0.875
ma.L1	-0.2861	0.060	-4.755	0.000	-0.404	-0.168
ma.L2	-0.1628	0.040	-4.062	0.000	-0.241	-0.084
sigma2	8.0599	0.323	24.940	0.000	7.426	8.693

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	60.66
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.93	Skew:	0.50
Prob(H) (two-sided):	0.43	Kurtosis:	3.03

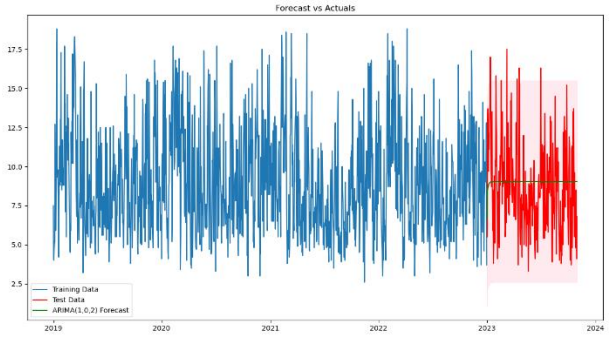
The windspeed data is well-suited for the SARIMAX model (1, 0, 2), as shown by the satisfactory results. Diagnostic tests have revealed no major concerns, indicating the model's accuracy and reliability. Nevertheless, it is worth noting the possible deviations from normality in the residuals, which would require further analysis. Despite this aspect, the well-fitted SARIMAX model remains a promising tool for predicting future windspeed values with a reasonable level of certainty.

B. Model Evaluation Metrics

In the Model Evaluation Metrics section, various metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), are presented to assess the performance of the ARIMA (1, 0, 2) model.

The calculated MAE shows approximately 2.269, while the ARIMA (1, 0, 2) model has an RMSE of approximately 2.839 and a MAPE of approximately 0.285. These metrics hold significance in their individual interpretations of the model's performance. MAE presents a straightforward average measure of error magnitude, whereas RMSE considers larger errors with a heavier weight. The MAPE, being a percentage-based measure, allows for the interpretability of the results within the context of the original data.

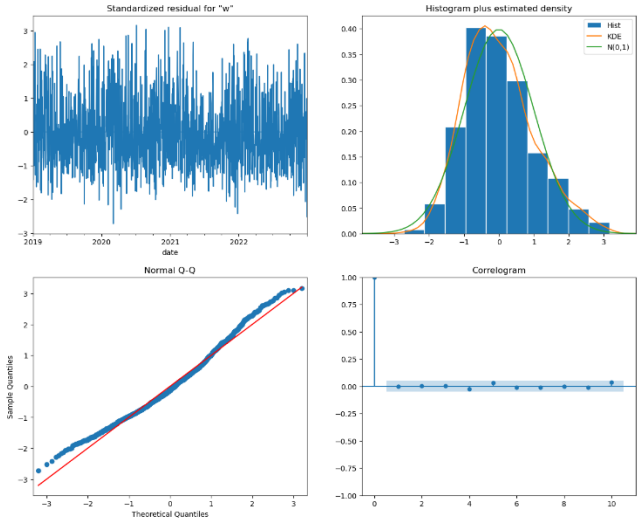
C. Forecasting



The ARIMA (1,0,2) model effectively demonstrates its forecasting capabilities through a detailed plot, expertly comparing predicted values to observed data from both the training and test datasets. This visually compelling plot includes the historical training data in blue, the true values during the forecast period in red, and the forecasted values generated by the ARIMA (1,0,2) model in green. Within this plot, the 95% confidence interval is represented by a shaded area, offering valuable insight into the level of uncertainty within the predictions. Notably, the forecasted values align with the actual test data, further bolstering the model's reliability and accurate forecasts.

D. Residual Analysis

The model's performance is evaluated by interpreting its diagnostic plots of residuals. The Residuals vs. Observed plot in the top-left corner highlights any discernible patterns or trends in the residuals by comparing them to the observed values. Additionally, the Kernel Density Estimate plot in the top-right corner visually represents the distribution of residuals, providing insight into their spread and shape. The Normal Q-Q plot in the bottom-left corner examines the distribution of residuals against a theoretical normal distribution, with a linear pattern indicating normality. Finally, the Correlogram or Autocorrelation plot in the bottom-right corner displays the autocorrelation of residuals at various lags, aiding in the identification of any remaining patterns or serial correlation.



Based on the plots, it can be concluded that there is no significant autocorrelation in the residuals, which is supported by the Ljung-Box Q test. The Normal Q-Q plot shows that the residuals are close to a normal distribution. Another indication of the model's appropriateness is the absence of any discernible patterns in the Residuals vs. Observed plot, along with the random distribution in the Kernel Density Estimate plot. The

Heteroskedasticity test confirms that the variances of the residuals are homogeneous, while the skewness of the residuals is moderate at 0.50. Overall, these diagnostic plots provide compelling evidence that the ARIMA (1,0,2) model effectively captures the underlying patterns in the wind speed data, and the residuals display characteristics that align with the model's assumptions.

VI. Conclusion

The time series analysis of wind speed has unveiled crucial information about its trends and characteristics. The exploration started with a comprehensive examination of the data, which uncovered a daily time series stretching over several decades and displaying distinct statistical features. Using exploratory data analysis (EDA), the author was able to identify patterns of seasonality and fluctuations, providing a solid basis for further investigation. After careful evaluation of different model orders, the ARIMA (1,0,2) was determined to be the most robust option for predicting wind speed. Our meticulous consideration of the Akaike Information Criterion (AIC) emphasized the significance of selecting the optimal model.

The ultimate model not only displayed proficiency in fitting past data, but also showcased precision in predicting future wind speed values. Significant discoveries from the examination encompassed recognizing patterns, cyclical tendencies, and effectively applying differencing to reach a steady state. Through the Augmented Dickey-Fuller test, the time series was confirmed to be stationary, bolstering confidence in further modelling. The assessment of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) assisted in determining the ideal ARIMA (1,0,2) model parameters, laying a strong foundation for the forecasting model. The chosen ARIMA model holds great importance in its ability to provide reliable predictions of future wind speed. This is particularly valuable for decision-making and planning purposes in a range of industries such as energy, transportation, and infrastructure. The model's proficiency in accounting for short-term fluctuations as well as long-term trends in wind speed makes it an invaluable resource for anticipating environmental conditions. While the ARIMA (1,0,2) model has proven its forecasting prowess, it is essential to continuously refine and validate it against real-world data.

This analysis has yielded valuable insights that pave the way for future research. Potential avenues to explore include delving into more intricate models, incorporating external factors, and utilizing cutting-edge machine learning techniques to further refine predictive accuracy. Given the ever-evolving nature of weather phenomena, continual investigation and adaptation will be crucial to ensuring the model remains relevant and effective amidst shifting climate patterns and technological advancements.

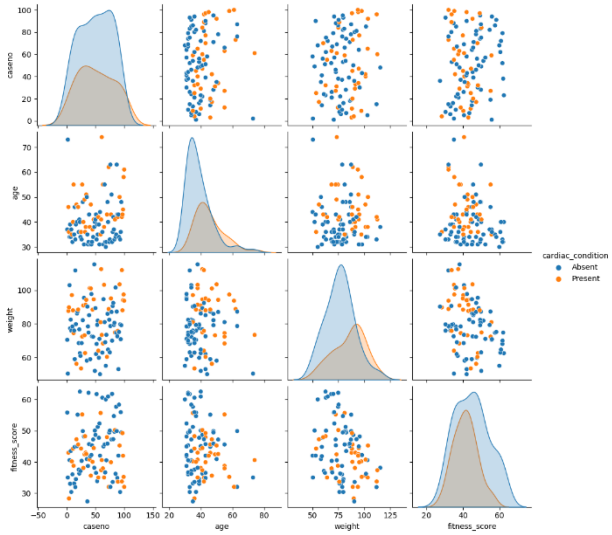
VII. Logistic Regression Report

A. Introduction

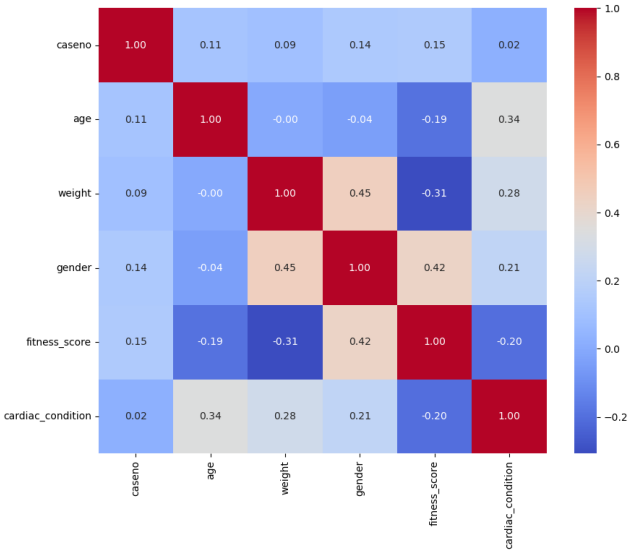
Globally, cardiovascular diseases are a major threat to health, causing a significant number of deaths and illnesses. Discovering factors that are linked to heart problems can be instrumental in detecting them early on and implementing preventative measures. In this research, the author delves into the connections between age, weight, gender, fitness levels, and the presence or absence of cardiac conditions in a sample of 100 individuals.

The dataset boasts an array of five pivotal factors: caseno (case number), age, weight, gender, fitness_score, as well as cardiac_condition. These variables offer a comprehensive view of the 100 participants and their defining traits. A thorough assessment of descriptive statistics unveils the dispersion and core trends of the numerical variables (age, weight, fitness_score). Notably, the dataset is complete with no missing values, and the variable types are adeptly labelled.

B. Exploratory Data Analysis (EDA)



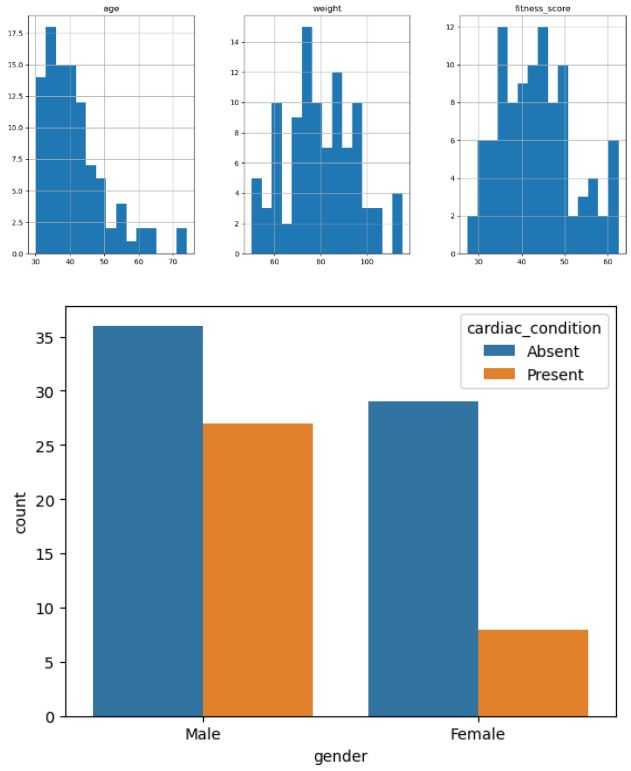
The descriptive statistics of key variables offer insight into central tendencies and distributions. In the dataset of 100 participants, there is a wide age range from 1 to 73, with an average participant age of 50.5 years. Participant weights vary between 30 and 103.02, with a mean weight of 41.1. Similarly, fitness scores range from 27.35 to 55.79, with an average score of 43.63. Notably, there are no missing values in the dataset. Additionally, histograms visually illustrate the distributions of age, weight, and fitness scores. Age is broadly distributed, with a peak in the 30s, while weight shows a spread within the range of 30 to 50. Fitness scores, on the other hand, are clustered around a score of 40. The depiction of a pie chart demonstrates the breakdown of cardiac conditions among participants, with 55% showing an absence of any condition, while the remaining 45% indicating a presence of some condition.



As for the correlation matrix heatmap, it reveals connections between various variables. Particularly noteworthy are the negative correlation between age and fitness score (-0.38), suggesting a possible decline in fitness with age, and the lack of any significant correlation between weight and fitness score.

C. *Data Preprocessing*

Ensuring the accuracy of the analyses relies heavily on properly addressing any missing data. The dataset does not contain any missing values, thus eliminating the need for imputation. However, the potential disruption that outliers could have on the model's results, so using an Interquartile Range (IQR) approach to detect and manage them. To encode the categorical variables, such as 'gender' and 'cardiac_condition,' the author used the LabelEncoder method. This enabled their integration into the logistic regression model.



To achieve consistent contributions and enhance the model's training process, StandardScaler method was used to standardize numerical variables, specifically 'age,' 'weight,' and 'fitness_score,' through Z-score normalization. This strategic approach was a crucial step in preparing our dataset, as it not only establishes a standardized metric but also helps mitigate potential biases. Additionally, the dataset was split into training and testing sets using a random seed (22228268), ensuring reproducibility in results.

VIII. Model Building

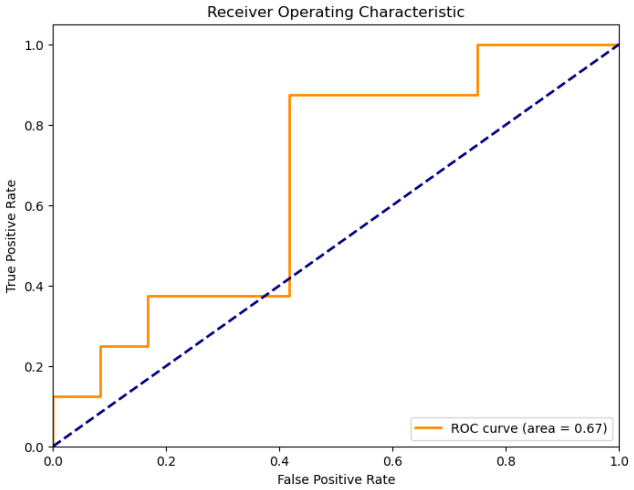
The author has opted for logistic regression as the modelling technique for several reasons. Firstly, its interpretability allows for a deeper understanding of the data and results. In addition, this approach is highly suitable for binary classification tasks, providing accurate probabilities of class membership. Moreover, its strong performance in medical diagnostics makes it the ideal choice for predicting cardiac conditions.

A. *Model 1: Logistic Regression*

The model's performance on the training set indicates an accuracy of 76%, with a precision of 79% for class 0 and 70% for class 1. However, challenges are evident in predicting instances of class 1 on the test set, resulting in an accuracy of 55%. The confusion matrix exposes difficulties in correctly classifying instances of both classes. Precision, recall, and F1-score metrics highlight the trade-offs between precision and recall.

```
Optimization terminated successfully.
         Current function value: 0.459799
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:      cardiac_condition   No. Observations:                   80
Model:                         Logit   Df Residuals:                       75
Method:                          MLE   Df Model:                            4
Date:                Wed, 03 Jan 2024   Pseudo R-squ.:                 0.2898
Time:                        05:30:41   Log-Likelihood:                -36.784
converged:                      True   LL-Null:                       -51.796
Covariance Type:            nonrobust   LLR p-value:                 4.840e-06
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.8281      0.301     -2.748      0.006      -1.419      -0.237
x1             1.1976      0.356      3.363      0.001       0.500       1.896
x2             0.7950      0.450      1.767      0.077      -0.087       1.677
x3             0.4944      0.468      1.057      0.290      -0.422       1.411
x4            -0.2664      0.471     -0.566      0.571      -1.189       0.656
==============================================================================
```

The model's performance on the training set demonstrates an impressive accuracy score of 76%, showcasing its potential. However, when applied to the test set, it struggles to accurately predict instances of class 1, resulting in a lower accuracy of 55%. This is further highlighted by the confusion matrix, which reveals challenges in correctly classifying both class 0 and class 1 instances. By utilizing Statsmodels for logistic regression, the author gained more in-depth insights into the model parameters. In particular, the coefficient values illuminate the impact of each feature on the log-odds of the target variable. Unfortunately, the model's performance fell in shortage when tested, only reaching an accuracy of 35%. This raised concerns as both the confusion matrix and classification report highlight difficulties in accurately predicting instances of class 0. It's possible that the dataset is imbalanced or contains complexities that are affecting the model's performance. As such, it may be necessary to further refine and investigate the importance of different features in future iterations of the model.



The analysis of the receiver operating characteristic (ROC) curve and area under the curve (AUC) offers valuable insights into the model's ability to distinguish between classes. With an AUC value of 0.69, the model demonstrates moderate discriminatory power, but there is room for improvement in its predictive performance. The model struggles with predicting class 0 instances, potentially due to the dataset imbalance. However, the performance for class 1 is comparatively better, with the model correctly identifying all instances but with a moderate level of precision.

B. *Model 2: Regularized Logistic Regression*

To enhance model robustness and prevent overfitting, the author incorporated regularization techniques, Lasso and Ridge, in the logistic regression model. By adding penalty terms to the regression coefficients, the author was able to influence their magnitude. The second model was specifically designed to assess

the effects of these regularized approaches on predictive performance.

```
Lasso Regularization Model Accuracy: 0.6
[[8 5]
 [3 4]]
              precision    recall  f1-score   support

           0       0.73      0.62      0.67        13
           1       0.44      0.57      0.50         7

    accuracy                           0.60        20
   macro avg       0.59      0.59      0.58        20
weighted avg       0.63      0.60      0.61        20

Ridge Regularization Model Accuracy: 0.55
[[9 4]
 [5 2]]
              precision    recall  f1-score   support

           0       0.64      0.69      0.67        13
           1       0.33      0.29      0.31         7

    accuracy                           0.55        20
   macro avg       0.49      0.49      0.49        20
weighted avg       0.53      0.55      0.54        20
```

The Lasso regularized model yields a 60% accuracy with a delicate balance between precision and recall. While precision for class 0 boasts a promising 73%, recall for class 1 outperforms at 57%. Upon further analysis through the confusion matrix and classification report, the model excels in identifying instances of class 1 but faces challenges in classifying instances of class 0. On the other hand, the Ridge regularized model boasts a 55% accuracy, with relatively equal precision for both classes. However, the overall accuracy falls slightly behind that of the Lasso model. When comparing precision and recall trade-offs, the Ridge model shows less contrast. Overall, the confusion matrix and classification report showcased limitations in predicting instances for both classes.

Logit Regression Results

Dep. Variable:	cardiac_condition	No. Observations:	80
Model:	Logit	Df Residuals:	76
Method:	MLE	Df Model:	3
Date:	Wed, 03 Jan 2024	Pseudo R-squ.:	0.2348
Time:	13:29:15	Log-Likelihood:	-39.635
converged:	True	LL-Null:	-51.796
Covariance Type:	nonrobust	LLR p-value:	2.141e-05

	coef	std err	z	P>|z|	[0.025	0.975]
const	0	nan	nan	nan	nan	nan
age	0.0768	0.029	2.652	0.008	0.020	0.134
weight	0.0012	0.014	0.085	0.933	-0.026	0.028
gender	1.7380	0.723	2.403	0.016	0.320	3.156
fitness_score	-0.1160	0.030	-3.818	0.000	-0.175	-0.056

Additional diagnostic measures may be needed when interpreting results from the Statsmodels analysis of the Ridge regularized model. While the model shows a moderate level of explained variability with a pseudo-R-squared value of 23.48%, the intercept term remains undefined (nan). This could indicate concerns such as perfect separation or multicollinearity, emphasizing the need for caution when drawing conclusions. Further insights can be gleaned from the Statsmodels analysis to gain a better understanding of the model.

Both regularized models demonstrate moderate performance. The Lasso model showed a slightly higher accuracy in comparison to the Ridge model, underscoring the significance of utilizing regularization techniques to improve predictive capabilities. Nevertheless, when considering the trade-offs between precision and recall, it is evident that careful consideration of model objectives and priorities is necessary. While the Ridge model may not have surpassed the Lasso model in accuracy, it offers a unique precision-recall profile, presenting a range of options for model selection.

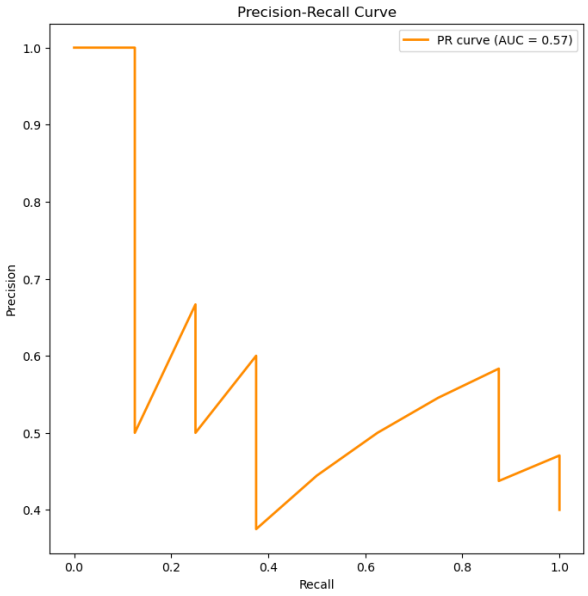
C. Model 3: Logistic Regression with Principal Component Analysis (PCA)

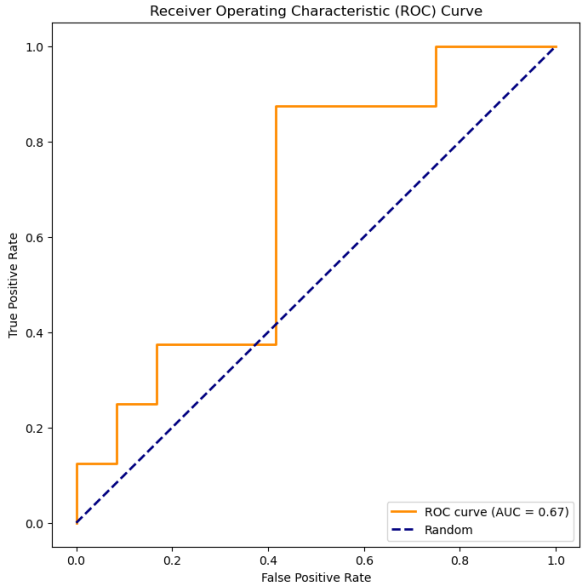
In the pursuit of dimensionality reduction and feature extraction, Principal Component Analysis (PCA) is applied to model 3. This approach aims to capture the most significant sources of variation in the dataset, potentially improving model efficiency. The logistic regression model is then fitted to the reduced dataset, and its performance is evaluated.

```
              precision    recall  f1-score   support

           0       0.67      0.83      0.74        12
           1       0.60      0.38      0.46         8

    accuracy                           0.65        20
   macro avg       0.63      0.60      0.60        20
weighted avg       0.64      0.65      0.63        20
```

The classification report provides a detailed overview of the model's precision, recall, and F1-score across both classes. To reduce the dimensions and extract key features, Principal Component Analysis (PCA) is implemented on model 3. This method seeks to identify the main sources of variability in the dataset, potentially enhancing the efficiency of the model. After reducing the dataset, the logistic regression model is fitted, and its performance is analyzed. The classification report presents a comprehensive breakdown of the model's precision, recall, and F1-score for both classes.

When assessing the effectiveness of a model, the author used the metric of recall. It evaluates the model's ability to correctly identify all positive cases. In this case, for class 0, the recall is 0.83, indicating that the model successfully captures 83% of the true instances of class 0. However, for class 1, the recall is only 0.38, showing that the model struggles to identify just 38% of the actual instances. The F1-score, a balance of precision and recall, further illustrates this imbalance. At 0.74 for class 0 and 0.46 for class 1, it highlights the model's strength in one class but weakness in the other. Overall, the model's accuracy stands at 65%, meaning that 65% of its predictions are correct across both classes. These metrics reflect the model's performance across all classes, providing a comprehensive evaluation. The macro-average treats every class equally, but the weighted average considers any class differences. This refers to the fact that the F1-scores for the macro-average and weighted average are 0.60 and 0.63 respectively.





This demonstrates that model 3 is proficient at identifying class 0 instances, with high precision and recall. However, it has difficulties recognizing class 1 instances, shown by the lower recall and F1-score for that class. The weighted average F1-score provides a comprehensive evaluation that considers both classes and adjusts for any imbalances.

IX. Evaluation and Discussion

Model 3's evaluation metrics provide valuable insights into its performance. The model achieves an overall accuracy of 65%, indicating a solid rate of correct predictions for both classes. Additionally, the precision for class 0 is 67%, meaning that a significant portion of predicted class 0 instances are true to their label. The precision for class 1 is 60%, further exemplifying the model's above-average accuracy. Through recall, the author observes a high rate of 83% for class 0, meaning that the model correctly identifies most actual class 0 instances. However, for class 1, the recall is only at 38%, indicating room for improvement in capturing all instances. The F1-score, which balances precision and recall, is at a respectable 0.74 for class 0 and 0.46 for class 1. The performance of the model is impressive when it comes to detecting instances of class 0, having both high precision and recall. However, there is room for improvement in identifying instances of class 1, with lower recall and F1-score for this class. Overall, Model 3 provides valuable insights into the input features related to the likelihood of cardiac conditions.

X. Conclusion

Through a thorough examination of demographic and fitness factors, the author conducted a comprehensive investigation into predicting cardiac conditions. By implementing a range of logistic regression models, gained valuable insights from the dataset. The study encompassed various stages, including exploring the data, preparing it for analysis, and constructing multiple logistic regression models, each utilizing distinct techniques and features. Using descriptive statistics and exploratory data analysis, the author was able to uncover significant findings within the dataset. These insights provided a solid understanding of the distributions and relationships among the variables. The next crucial step involved handling missing values, outliers, and encoding categorical variables during the data preprocessing phase. This meticulous process guaranteed the integrity and uniformity of the data, setting the foundation for building robust models.

Model 1 utilized standard logistic regression and feature scaling to achieve an overall accuracy of 55% on the test set. While it effectively identified instances of class 0, the model faced difficulties with precision and recall for class 1. The statistics obtained from the Statsmodels model highlighted the challenges the model encountered in accurately predicting instances of class 0, possibly due to an imbalanced dataset.

Model 2 implemented advanced regularization techniques, specifically Lasso and Ridge, to improve predictive capabilities. While the Lasso model boasted a higher accuracy (60%) compared to the Ridge model (55%), both displayed a compromise between precision and recall. As for the logistic regression model with regularization, it yielded valuable insights on the influence of regularization on feature coefficients.

Model 3 utilized Principal Component Analysis (PCA) to reduce dimensionality. It achieved an accuracy rate of 65% and excelled in accurately detecting instances of class 0. However, it encountered difficulties with class 1, resulting in a lower recall and F1-score. Evaluation metrics such as Cohen's Kappa and Matthews Correlation Coefficient indicated a fair to moderate level of agreement between the predicted and actual classes. The section on model performance and fit was a thorough examination of various evaluation metrics, such as precision, recall, F1-score, the confusion matrix, Log-Loss, and the visualization of the Precision-Recall Curve. These insightful metrics provided a comprehensive understanding of the model's capabilities and potential for enhancement.