TABA Statistics : Time Series Analysis and Logistic Regression

Komal Prakashchandra Patra
M.Sc. Data Analytics
National college of Ireland
Dublin, Ireland
x22210369@student.ncirl.ie

***Abstract*—The research study is divided into two sub-section, which are Time series analysis and Logistic Regression. Part A of the study involved performing time series analysis on the 'weather' dataset to analyse the decomposition of the components and forecast the data using five different models. The forecasting for the time series was done for an year (365 days). The best model was determined by comparing all of the time series models with one another. The best fit model for the dataset was Simple Exponential smoothing (SES). The "cardiac" dataset was used for our logistic regression analysis in part B, with the target variable being cardiac condition. In order to optimize the hyperparameter tuning, the logistic model was executed using the Logit function from Statsmodel in conjunction with gridSearchCV, which produced an accuracy of 76%.**

***Index Terms*—Exponential Smoothing, ARIMA, Moving Average, Holt's Linear, Trend, Seasonal, Accuracy, Root Mean square error, Akaike Information criterion (AIC), Bayesian Information criterion (BIC), Logistic model, Precision, Recall**

PART A: TIME SERIES ANALYSIS

I. Data Description and Data Understanding

Time series analysis and forecasting is the most popular statistical approach, which refers to a series of data concerning a specific time interval. When analyzing data points, time series analysis plays a significant role in order to investigate and identify seasonal fluctuations, autocorrelations, and trends over a period of time. Forecasting the time series helps to predict future values based on historical observations. The weather dataset consists of features like maximum and minimum air temperature, precipitation amount, evapotranspiration and potential evapotranspiration, and mean wind speed from January 1942 to October 2023. This data is the historical data on weather by Met Eireann from one of Ireland's important stations, i.e., Dublin Airport, consisting of 29889 records. Mean wind speed (*wdsp*) is the feature on which we have conducted this time series analysis study. In addition to offering insights into historical atmospheric conditions, an understanding of mean wind speed historical patterns serves as the foundation for potential trend prediction in the future. Long-term shifts in wind dynamics can be found through trend analysis, which offers essential insights into the bigger picture of climate change and its possible effects. Precise forecasting of wind speed help energy suppliers anticipate future wind patterns, which facilitates more effective planning for wind-powered electricity production. Forecasting for the model has been done using different models.

II. Data Analysis

Parsing temporal data, such as timestamps, is essential when working with data because it makes the representation and interpretation of the data easier for software and applications to understand. The data is parsed using the function `to_datetime()`. A problem with two-digit year formats emerged during the parsing process, especially when moving to the 21st century. For example, "42" was interpreted as either 2042 or 1942. A corrective measure has been implemented to correct parsing ambiguity by subtracting 100 years from a date with a year greater than 2023. As a result, the year has been extracted and added to a new column after the data was parsed in the format "%d-%b-%y". Through direct row location during filtering, sorting, or querying, indexing improves the efficiency of data retrieval and speeds up the process. In order to guarantee that the DataFrame is indexed according to the given dates, the `set_index()` method sets the 'date' column as an index, uniquely identifying each row. After that, the data was plotted for mean wind speed to see if any consistent trends emerged while maintaining a yearly frequency; however, no cyclic data were discovered. Once the data is filtered and

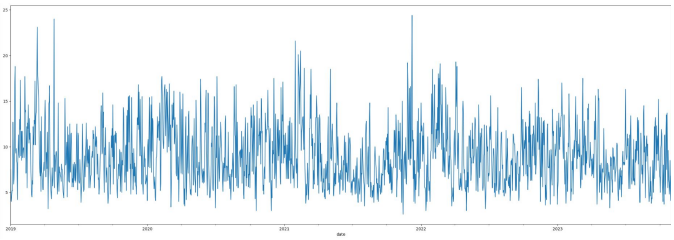

Fig. 1. Time Series Plot for Mean Wind Speed (wdsp)

reshaped, the next step in the time series analysis process is to investigate the underlying patterns and relationships between the data and other variables, such as wind speed. Plotting the ACF and PACF allowed us to examine the correlation of our raw data, which gave us important information about its nature and evolution over time. The primary objective behind the decomposition is to obtain the understanding of the underlying pattern and variation within a time. More precisely,

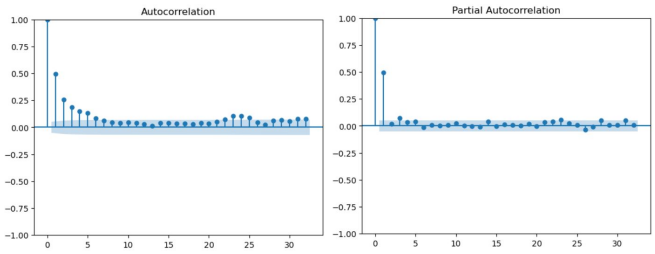

Fig. 2. ACF and PACF for raw data

they are made up of four components that are connected to different kinds of temporal fluctuations. Seasonal components, observables, trends, and random elements make up a time series. In order to gain a deeper understanding , breaking down a time series into its constituent parts, one can more thoroughly comprehend and examine the underlying trends, cyclic variation, seasonality, and random variations, which can provide light on the general behaviour of the time-dependent data. Since there is no obvious seasonality or recurring pattern in

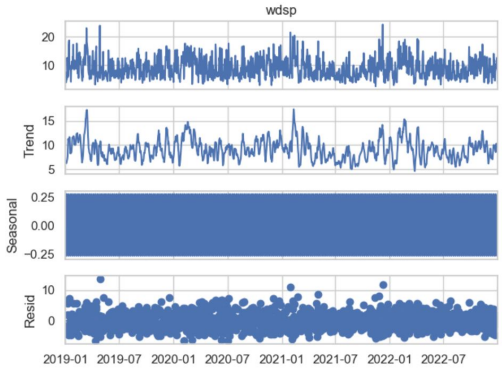

Fig. 3. Decomposition of time series for mean wind speed

the time series, it is likely that the seasonal component in Fig. 3 is flat. Decomposition can be divided into two categories: multiplicative and additive. We will be analyzing the data using the Additive time decomposition model because it does not show exponential seasonal fluctuation. The multiplicative

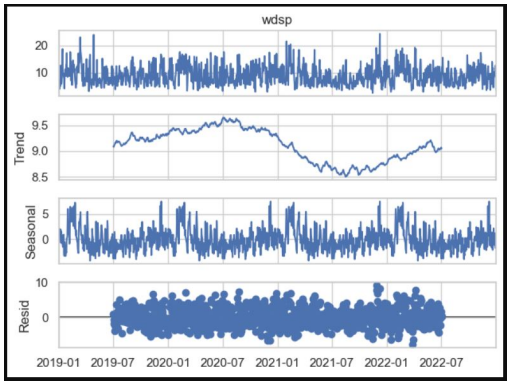

Fig. 4. Decomposition using Additive time series model for mean wind speed

model, on the other hand, performs best when the seasonal variance increases over time. By performing the additive time series model seen in Fig. 4, we are now able to observe the trend and seasonal pattern. The data's residual component is what is left over after the trend and seasonal components have been eliminated.

III. Model Building

Based on accuracy and the AIC/BIC curve, the optimal model will be chosen in compliance with the previously mentioned analysis. The future mean wind speed for 2023 will be predicted using a model that has lower AIC and root mean square, and we will base our decision on the output of these models.

A. MOVING AVERAGE:

1. Simple Moving Average

The moving average smoothens the curve by averaging adjacent values over a window (a predetermined time horizon) with respect to the mean, which is how it is used to identify trend patterns. According to the mean model, the average of all the past events can serve as the most accurate indicator of what will occur tomorrow. The rolling mean with set window sizes has been computed using the rolling method on the 'wdsp' column in order to smooth out short-term fluctuations and emphasise longer-term trends. A smaller window responds to changes faster but might be more noisy; a larger window offers a more stable but lagging average. As we have mean

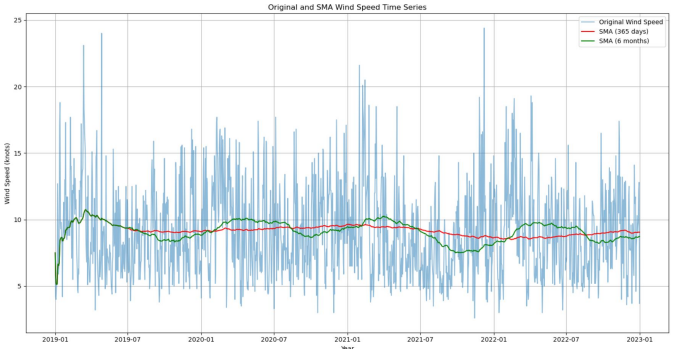

Fig. 5. Simple Moving Average(SMA) for 6 months and 1 year

wind speed value for each year with a 365 days data, we have calculated the SMA with two window size i.e. one for 6 months (182 days) and the other one as 1 year (365 days). We have observed from the given Fig. 5. that the SMA corresponding to yearly window provides more smoothened curve compared to SMA of 6 months.

2. Weighted Moving Average

Each observation is given a particular weight or frequency in a weighted moving average, which is calculated by increasing the weight of the most recent observation in relation to the earlier observations. By giving different weights to historical values, the WMA method emphasises recent observations more than historical ones. The RMSE value that we obtained was 3.364.

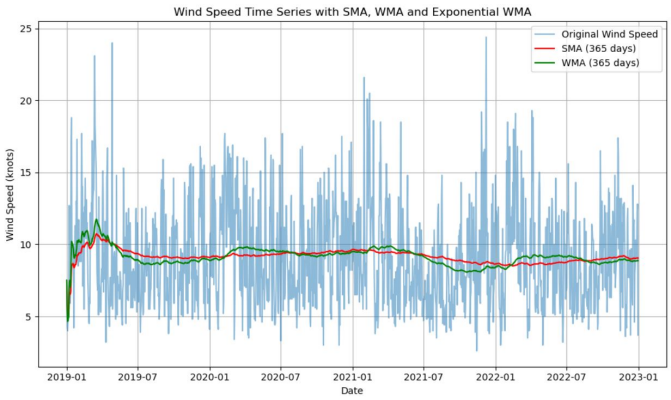

Fig. 6. Comparison of Simple Moving Average(SMA) and Weighted Moving Average (WMA) for 1 year(365 days)

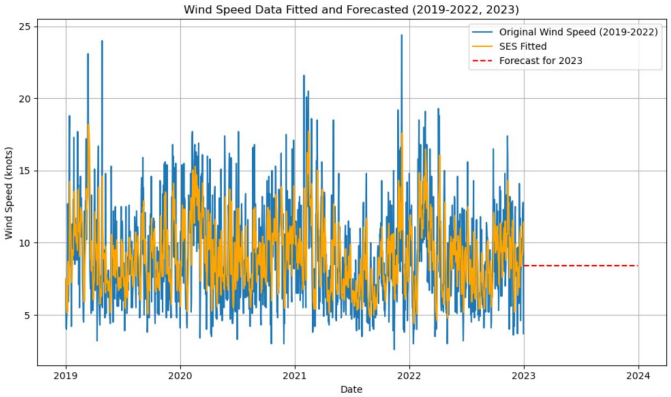

Fig. 8. Forecasting for SES

B. EXPONENTIAL SMOOTHING

Exponential smoothing predicts future values by leveraging a weighted average of all past values in the series. The goal of this simple and popular model is to smooth out a time series by forecasting it without any trend or seasonality. We are going to investigate three different forms of exponential smoothing.

1. Simple Exponential smoothing:

SES is best suited for forecasting the time series data with no trend as well as no seasonality. We have fitted a single exponential smoothing using the wdsp column in the dataframe. The reason that we have chosen this approach is because the data does not exhibit any systematic patterns or irregularities, and we wanted a simple forecast that was based on the previous forecast and the most recent observation. Using the legacy heuristic initialization method, we have fitted

SimpleExpSmoothing Model Results

Dep. Variable:	wdsp	No. Observations:	1461
Model:	SimpleExpSmoothing	SSE	14550.374
Optimized:	True	AIC	3362.102
Trend:	None	BIC	3372.676
Seasonal:	None	AICC	3362.129
Seasonal Periods:	None	Date:	Sun, 31 Dec 2023
Box-Cox:	False	Time:	08:17:34
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.3913006	alpha	True
initial_level	7.5000000	l.0	False

Fig. 7. Summary for SES

the model for the years 2019–2022, setting the smoothing parameters initially. An RMSE value of 3.156 and an AIC value of 3362.102 are obtained after the model is executed, and Fig. 8. displays a forecast for the next 365 days, or a complete year.

Due to the SES model's inability to identify seasonality or trends in the data, the forecast is flat.

2. Holt's Linear Model

Also referred as Double Exponential Smoothing (DES). This model is capable of fitting time series that have a trend but no seasonality. When there is a steady upward or a downward trend in the data and a forecast that takes into account both the current level and the trend is required, DES is helpful. Level and trend are its smoothing parameters; their values are independent of one another. Similar to the level-smoothing constant , the trend-smoothing constant can be interpreted in the same way. After the model was fitted, we saw that the

Holt Model Results

Dep. Variable:	wdsp	No. Observations:	1461
Model:	Holt	SSE	15750.683
Optimized:	True	AIC	3481.911
Trend:	Additive	BIC	3503.059
Seasonal:	None	AICC	3481.969
Seasonal Periods:	None	Date:	Sat, 30 Dec 2023
Box-Cox:	False	Time:	15:01:39
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.5402752	alpha	True
smoothing_trend	0.0377368	beta	True
initial_level	7.5000000	l.0	False
initial_trend	-3.3000000	b.0	False

Fig. 9. Summary for Holt's Linear Model

RMSE was 10.781 and the AIC was 3481.911, both of which are significantly higher. After projecting the time series for

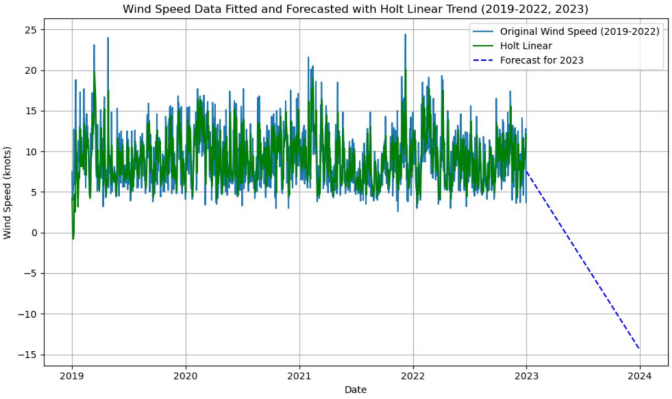

Fig. 10. Forecasting for Holt's Linear Trend (DES)

2023 shown in Fig. 10., we saw that the forecast was tracking

the trend, although in downwards direction. Therefore, As the error is large it shows that this model is not the suitable model.

3. Triple Exponential Smoothing

Winter's exponential smoothing is a method for adjusting time series that show trend, level, and seasonality. There are two different types of variations: the multiplicative is chosen when seasonal variations fluctuate in proportion to the series level, while the additive is chosen when seasonal variations stay mostly constant throughout the series. Upon model fitting,

ExponentialSmoothing Model Results

Dep. Variable:	wdsp	No. Observations:	1461
Model:	ExponentialSmoothing	SSE	10880.590
Optimized:	True	AIC	3671.482
Trend:	Additive	BIC	5622.340
Seasonal:	Additive	AICC	3924.948
Seasonal Periods:	365	Date:	Sun, 31 Dec 2023
Box-Cox:	False	Time:	09:57:14
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.2934835	alpha	True
smoothing_trend	6.2862e-07	beta	True
smoothing_seasonal	3.5981e-06	gamma	True

Fig. 11. Summary for Triple Exponential smoothing

the values of α, β, and γ were found to be 0.293, 6.2862×10^{-7}, and 3.5981×10^{-6}, respectively. Furthermore, RMSE is 3.507 and AIC is 3671.482. The model that we predicted for the test data is displayed in fig. Since we used the additive

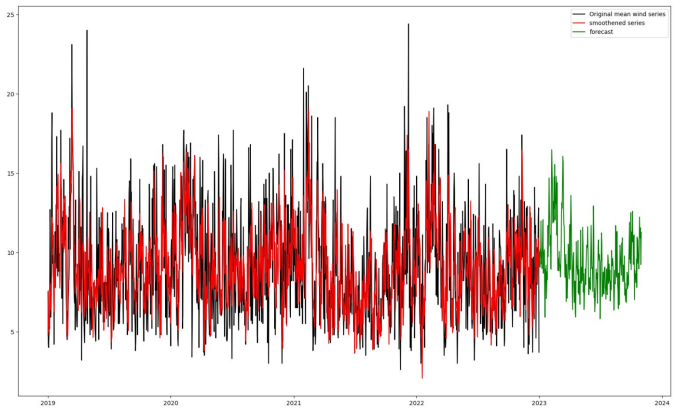

Fig. 12. Forecasting for Triple Exponential smoothing

method with seasonal period as yearly i.e. 365 days and the season is constant throughout the train, it is evident from the plot that the model has somewhat captured both the trend and the season.

C. ARIMA Model

Auto-Regression Integrated Moving Average is referred to as ARIMA. Auto Regression is the regression within a variable. ARIMA models aim to characterize the autocorrelations within the data, as opposed to exponential smoothing models, which explain the seasonality and trend in the data. Since the data must be stationary in order for ARIMA to work, we first verified whether the data is stationary. In order for our model to predict the mean and variance, which will remain constant in subsequent periods, the mean and standard deviation must remain constant for the data to be considered stationary. In general, There will not be any long-term consistent patterns in a stationary time series. In order to establish the null hypothesis, which claims that the data are stationary, and the alternative claims that they are not, we first constructed a statistical hypothesis. The data are stationary, as evidenced by the fact that the pvalue is less than the selected significance, such as 0.05. Therefore, we do not reject the null hypothesis. Consequently, we can

```
----------------------------------------
Results of Dickey-Fuller test:
Test statistic                -1.575764e+01
p-value                        1.202256e-28
Lags used                      2.000000e+00
Number of observations         1.458000e+03
Confidence Interval (1%)      -3.434843e+00
Confidence Interval (5%)      -2.863524e+00
Confidence Interval (10%)     -2.567826e+00
dtype: float64
```

Fig. 13. Dickey-Fuller test for validating Stationary data

skip the Differences step, which turns non-stationary data into stationary data. In particular, The ARIMA is specifically denoted as p,d,q where p represents Auto-regression, d as differences and q as moving average. Using PACF for p and ACF for q, the Auto Correlation yields the relationship between a variable and time lag correlation. The order of the AR and MA components is shown in the ACF and PACF plot seen in Fig. 14. With a sharp decline in PACF

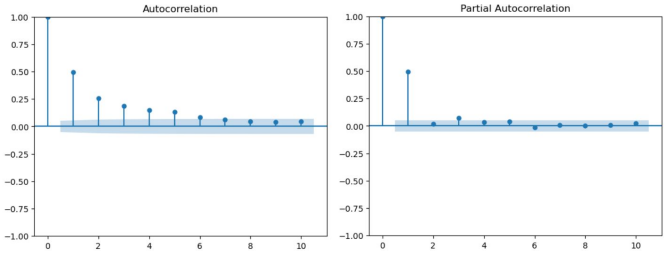

Fig. 14. ACF and PACF for ARIMA

indicating direct dependence and a gradual decline in ACF indicating a continuous correlation with historical values, the time series' pattern points to a first-order autoregressive model (AR1). However, to confirm the most suitable model, we have also fitted close variations of this model, such as $(1,0,0)(1,0,1)(1,0,2)(1,1,0)(1,1,1)(1,1,2)(2,1,0)(2,1,1)$ $(1,2,1)(2,2,1)(0,1,1)(1,0,5)$. AIC is lowest for the ARIMA $(1,0,2)$. Moreover, the `auto_arima()` function recommended the same sequence as the most suitable model. The residual plot, which can be used to check the distribution using the `plot_diagnostic()` function and to see the significant difference from white noise, is shown below in Fig. 16. According to the Ljung-Box test results, the time series is not significantly autocorrelated up to lag 10 as shown

SARIMAX Results

Dep. Variable:	wdsp	No. Observations:	1461
Model:	ARIMA(1, 0, 1)	Log Likelihood	-3656.885
Date:	Sat, 30 Dec 2023	AIC	7321.770
Time:	15:01:42	BIC	7342.918
Sample:	01-01-2019 - 12-31-2022	HQIC	7329.659
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	9.0952	0.179	50.922	0.000	8.745	9.445
ar.L1	0.5299	0.044	11.919	0.000	0.443	0.617
ma.L1	-0.0477	0.050	-0.947	0.344	-0.146	0.051
sigma2	8.7402	0.305	28.662	0.000	8.142	9.338

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	130.10
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	0.93	Skew:	0.64
Prob(H) (two-sided):	0.44	Kurtosis:	3.71

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Fig. 15. Summary for ARIMA model

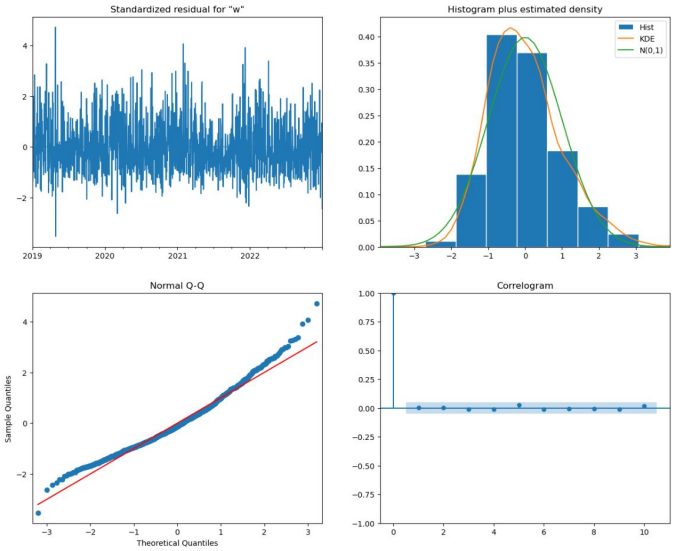

Fig. 16. Residual plot for ARIMA

in Fig. . The observed test statistics are consistent with the null hypothesis that there is no autocorrelation, as indicated by the high p-values. Thus, this is a good finding because it shows that the residuals or errors in the ARIMA time series model do not show any visible temporal patterns. We have projected the ARIMA model for ORDER$(1, 0, 2)$ on the test, which is shown in Figure **??**. Since the data has been divided into training and test sets, following evaluation, we obtained an RMSE value of 2.926 for ARIMA(1,0,2) and an AIC value of 7313.196.

IV. Conclusion

In this comprehensive time series analysis, we explored various forecasting models to capture and understand the underlying patterns in the data. The objective was to identify the most suitable model for predicting future values with the least amount of error. Starting with the basic SMA and WMA models, we observed that the WMA outperformed SMA. This indicated that the weighted approach, which assigns different weights to historical values, captured the underlying trends more effectively. Moving on to SES, the model demonstrated improved performance as the lower AIC value suggested a

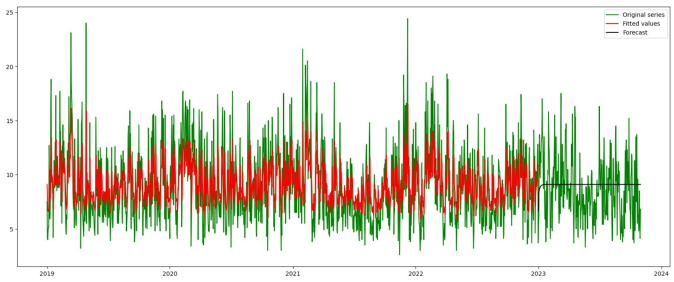

Fig. 17. Forecasting for ARIMA

	lb_stat	lb_pvalue
1	0.004102	0.948933
2	0.017079	0.991497
3	0.177485	0.981139
4	0.313805	0.988906
5	1.570584	0.904784
6	1.815391	0.935870
7	1.851646	0.967593
8	1.905178	0.983782
9	2.070185	0.990308
10	2.601669	0.989310

Fig. 18. Ljung-Box test for ARIMA model

better balance between model fit and complexity. SES leverages an exponentially decreasing weight for past observations, making it responsive to recent trends while adapting to changes in the data. Holt's Linear Trend, incorporating a linear trend component. Despite the higher RMSE, the inclusion of a trend allows for capturing more complex patterns in the data. However, the relatively high RMSE indicates that the linear trend might not be the most appropriate representation for the data. Triple Exponential Smoothing, an extension of SES with the addition of a seasonality component. This model is particularly effective for time series with both trend and seasonality. The observed performance indicated reasonable predictive capabilities, but the higher RMSE suggest some limitations. The ARIMA(1,0,2) model demonstrated competitive forecasting accuracy. However, the high AIC value of 7313.196 suggests a potential trade-off between fit and complexity which raises concerns about potential overfitting. ARIMA models are known for capturing autocorrelations and trends, but in this case, the high AIC raises concerns about overfitting. Ultimately, the particular requirements and time series characteristics determine which model is most appropriate. The SES model, with its comparatively lower RMSE and AIC, stands out as a reliable choice for its simplicity and effectiveness in capturing the underlying patterns. However, depending on the context and goals, a trade-off between complexity and accuracy may lead to the selection of a different model. Maintaining the balance between predictive accuracy and model simplicity while taking into account each model's advantages and disadvantages is essential. Further model refinement and exploration of alternative approaches may be warranted to enhance forecasting performance.

PART B: LOGISTIC REGRESSION

A. Introduction

In the pursuit of improving patient outcomes and enhancing the quality of life, the medical field is confronted with a formidable challenge—early disease diagnosis and prompt treatment. In order to improve patient outcomes and quality of life, the medical field faces a significant challenge in improving early disease diagnosis and prompt patient treatment. Every year, there are an increasing number of cases of cardiovascular disease. For an accurate diagnosis, it is critical to become aware of the cardiac condition as soon as possible. As a result, a machine learning model-based supportive system is required to predict the cardiac condition. The ability to predict one's cardiac condition depends on a number of relevant factors, including age, gender, weight, and fitness level. These factors not only contribute to the development of cardiac conditions but also influence the effectiveness of treatment strategies. The dataset has 6 features and 100 records with the dependent variable 'cardiac condition' being a dichotomous variable. It includes independent features that are dependent on predicting the presence or absence of a cardiac condition, such as age, gender, fitness score, and weight. A binary classification is likely to occur between the two categorical values for the dependent variable cardiac condition: present and absent. To perform prediction on binary classification data, the following dataset has undergone logistic regression analysis.

B. Descriptive statistics

To ensure the quality of the dataset, a comprehensive inspection has been performed as the first step in our data exploration process. The data must be analysed, transformed and interpreted in order to build the model. The lack of duplicates and missing values highlights the flawless quality of the data. Examining the dataset's descriptive statistics is the next logical step after determining its integrity shown in Fig. 19. The mean, median, and mode, measures of central tendency

	caseno	age	weight	fitness_score
count	100.000000	100.00000	100.000000	100.000000
mean	50.500000	41.10000	79.660300	43.629800
std	29.011492	9.14253	15.089842	8.571306
min	1.000000	30.00000	50.000000	27.350000
25%	25.750000	34.00000	69.732500	36.595000
50%	50.500000	39.00000	79.240000	42.730000
75%	75.250000	45.25000	89.912500	49.265000
max	100.000000	74.00000	115.420000	62.500000

Fig. 19. Descriptive Statistics of Cardiac dataset

offer information about the typical or central value of the dataset. The variability of the dataset is revealed by dispersion metrics such as range, variance, and standard deviation, while qualitative analyses provide in-depth understanding of the value distribution. A box plot was created to identify outliers in the dataset before applying the model, revealing a negligible outlier. Due to the small amount of data, eliminating outliers was not a reasonable solution as it was not harmful to the model's construction. Using a seaborn pairplot, Fig. illustrates

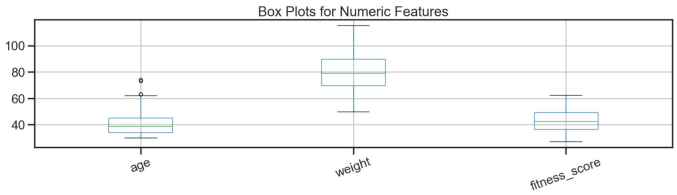

Fig. 20. Boxplot to check outliers

the relationship between several independent variables and the target variable.

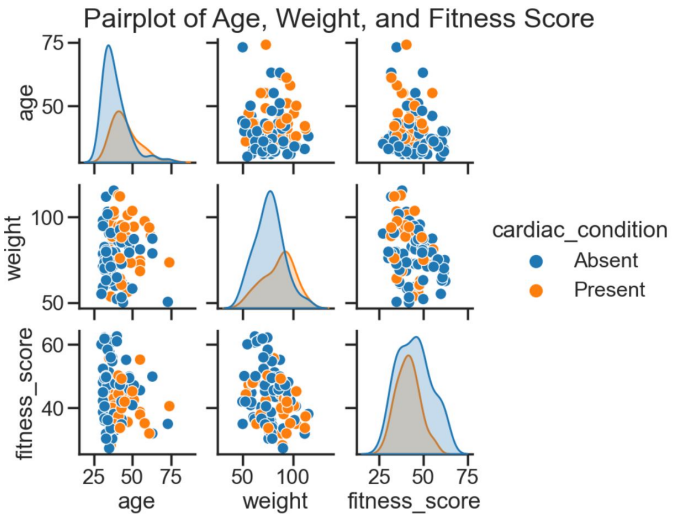

Fig. 21. Relationship of Independent variables with target variable

C. Methodology

The machine learning model called logistic regression (LR) is used to assess the correlation between the independent variable and the target variable, which is interval and ratio-scored categorical data. The formula for Logistic Regression is given by:

$$f(x) = \frac{1}{1 + e^{-x}}$$

We fit a logistic function with an "S" shape, referred to as the sigmoid function, that predicts two maximum values, 0 and 1, in logistic regression rather than a regression line. The ratio of something happening to nothing happening is called an odds ratio. Firstly, the data has been imported, and descriptive statistics have been performed to understand the data. Then, using Exploratory Data Analysis (EDA), the data is examined to ascertain whether a predictive model is a workable analytical tool for a particular task. Plotting visualizations in relation to the predictor variable has allowed researchers to examine the data distribution and the results of

Logit Regression Results

Dep. Variable:	cardiac_condition	No. Observations:	75
Model:	Logit	Df Residuals:	70
Method:	MLE	Df Model:	4
Date:	Tue, 02 Jan 2024	Pseudo R-squ.:	0.2110
Time:	13:04:10	Log-Likelihood:	-39.827
converged:	True	LL-Null:	-50.476
Covariance Type:	nonrobust	LLR p-value:	0.0002765

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9340	4.573	-0.642	0.521	-11.898	6.030
age	0.0787	0.034	2.315	0.021	0.012	0.145
weight	0.0351	0.027	1.288	0.198	-0.018	0.088
gender	-1.2579	0.946	-1.330	0.184	-3.112	0.596
fitness_score	-0.0745	0.053	-1.414	0.157	-0.178	0.029

Fig. 22. Summary for Logistic regression

both univariate and bivariate analysis. The crucial phase in data preprocessing is data mapping, which converts the categorical variable into numerical representations. The mapping dictionaries that are utilized are {'Absent': 0, 'Present': 1} for cardiac condition and {'Male': 0, 'Female': 1} for gender. The dataset has been divided into train and test sets using a random seed representing the student ID, 22210369. We have also employed cross-validation with GridSearchCV using a 5-fold approach. Two robust Python libraries, Statsmodels and Scikit-learn, were used to perform logistic regression data modeling. Furthermore, the model's performance is improved, and optimal predictive accuracy is ensured by integrating GridSearchCV for hyperparameter tuning.

D. Data Modelling and Evaluation

1) Model 1: LR using statsmodel: The LR model was fitted with the help of the Logit function and a model summary was generated, as shown in Fig. To maximise the log-likelihood and optimise the algorithm for changing the model parameter, the Maximum Likelihood Estimation (MLE) method has been used. Finding out how well the independent variable explains the variation in the target variable is done by measuring the pseudo R-squared, which comes out at 0.2110. We have observed log-likelihood as -39.827 which measures the degree to which the model correctly predicts the data. The coefficients representing each of the independent variable which changes per unit.

2) Model 2: Implemeting LR using GridSearchCV for Hyperparameter tuning: This study investigates the application of GridSearchCV, a method that iteratively searches a predefined hyperparameter grid to optimise model performance, in conjunction with logistic regression. It is used for cross-validated hyperparameter tuning, which involves assessing the model's performance for every combination of hyperparameters. The regularization strength, penalty type, and maximum iterations of the solver are all controlled by a range of values, ensuring a smooth and efficient process.

Evaluation

Odd ratio is referred as the association between predictor and outcome variable. For both the model, the odd ratio is same as shown in Fig. It indicates how a one-unit change in each predictor affects the odds of having a cardiac condition,

	odds_ratio	variable
0	1.082796	age
1	1.047437	weight
3	0.950071	fitness_score
2	0.512781	gender

Fig. 23. Odd Ratio

holding other variables constant. With respect to age, as one unit of it changes, the probability of having a cardiac condition increases by 8.19%. This suggests that older people may have higher odds of experiencing a cardiac condition. The odds of having a cardiac condition for females is 28.43%, which is approximately higher compared to males. While fitness levels reduce the odds by 7.82%, weight increases the likelihood of having a heart condition by 3.57%. According to the Wald test, there is statistically significant evidence to reject the null hypothesis, as indicated by the low p-value (0.0057). It shows a significant difference from zero in at least one of the coefficients. The data shows 16 true positives, 3 true negatives, 4 false positives, and 2 false negatives, indicating correct predictions of positive and negative instances.

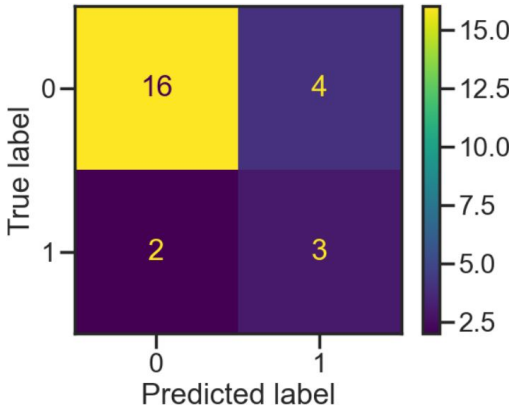

Fig. 24. Confusion matrix for Logistic regression

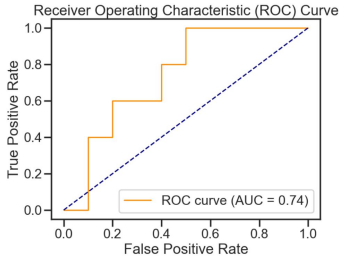

Fig. 25. ROC-AUC curve for Logistic Regression

The model's F1-score for class 0 is 0.84, with high precision (0.89) and moderate recall (0.80). Less favorable F1-score of 0.50 results from class 1 lower precision (0.43) and higher recall (0.60). The accuracy is 0.76. The macro average, highlighting a balanced representation of both classes, is 0.67. The weighted average, considering class imbalance, is 0.77. It appears from the data that the model performs better

when it comes to recognizing instances of class 0, but it still needs work to accurately classify instances of class 1. The scoring metric ”recall” focuses on eliminating false negatives, particularly in medical diagnosis scenarios where identifying positive instances is essential. The PCA was not performed as there was not much features present.

	precision	recall	f1-score	support
0	0.89	0.80	0.84	20
1	0.43	0.60	0.50	5
accuracy			0.76	25
macro avg	0.66	0.70	0.67	25
weighted avg	0.80	0.76	0.77	25

Fig. 26. Logistic classification Report

E. Conclusion

The logistic regression model, which was also optimized using GridSearchCV, shows good precision and recall for class 0, indicating that it can accurately detect negative cases. Still, there are problems with modest F1-scores and poor precision when classifying positive instances (class 1). The overall model accuracy stands at 76%. The evaluation underscores the need for further optimization, particularly in enhancing the model’s sensitivity to positive cases. The model may be improved for more evenly distributed performance in both classes through taking into account domain-specific complexities and possible feature engineering.