Statistics for Data Analytics (MSCDAD_C) Project Report

Name: Sushma Suresh Hebbalakar
Student ID:22224122

Master's in Data Analytics
National College of Ireland
03-01-2024

Time Series and Logistic Regression Analysis of Environmental and Health Data

Sushma Suresh Hebbalakar
Masters in Data Analytics
National College of Ireland
22224122

***Abstract*—This report combines two distinct yet complementary analyses Time Series Analysis of historical weather data and Binary Logistic Regression modelling for cardiac conditions. The time series analysis focuses on one of five weather variables (maxtp, mintp, gmin, cbl, or wdsp) determined by the last digit of the student numbers. For the chosen variable, we explore temporal patterns, trends, and seasonality, employing statistical methods to derive suitable models for forecasting and understanding long-term behaviour. In parallel, the report delves into a logistic regression model using the cardiac dataset, considering factors such as age, weight, gender, and fitness score to predict the likelihood of cardiac conditions. Potential dimensionality-reduction techniques are explored and justified for enhancing model interpretability.**

I. Introduction

This comprehensive report addresses both environmental and health aspects through Time Series Analysis and Binary Logistic Regression. The time series component utilizes extensive weather data from Dublin Airport, spanning from 1942 to 2023. Our aim is to unveil patterns and trends in one of the key weather variables, linking climate patterns to long-term environmental dynamics. Simultaneously, we explore the relationships between age, weight, gender, and fitness score with the presence or absence of cardiac conditions in a logistic regression model. Dimensionality-reduction techniques are introduced where deemed beneficial, offering insights into their impact on model interpretability. Data mining process on weather forecast in recent years is playing crucial role in human life. Start from normal person who'll check out weather before living house for any work to farmer who rely on nature for their work the weather forecast is important. We can use different model and smoothing techniques to find the desired data that can be rainfall, temperature humidity and many more [8] [5]. There are different techniques available for feature extraction and feature selection [1] which we can make use of to predict the desired result.

Amount of Rainfall prediction is a major issue for the weather department as it is associated with the human's life and the economy. Excess rainfall is the major cause of natural disasters such as drought and flood which are encountered by the people every year across the world. The time series machine learning model can developed to predict such disasters sooner and mke people aware [3]. here they have used ARIMA model to predict the rainfall.

Parallelly, making use of massive data available in the medical industry, we can build a models to predict disease which will act as medical decision support [7].Predicting and detecting cardiac disease has always been a difficult and time-consuming undertaking for doctors.For training and testing, a data collection containing diverse human health parameters is used. Many AI and ML algorithms are used to predict cardiac disorders [6].Missed diagnosis results in heart failure and cardiac arrest because the heart cannot pump oxygenated blood enough throughout the body.A cost-effective strategy to increase heart disease detection appropriateness is using computer-aided diagnostic categories [10], which could help i early detection and reduce the risk. We can make use of models available in ML like KNN and SVM [9]to predict the target but in our project we'll using logistic regression. There is also work done combining multiple models like Randomforest, Kstar and Bagging [2] and have concluded that ensemble model developed by combining the outputs of different models has performed better. Also a comparative study of five machine learning techniques, namely Logistic Regression, Naive Bayes, K-Nearest Neighbour, Decision Tree and Support Vector Machine, was conducted to compare the performance of the models for heart disease prediction [4]. And they have concluded that logistic regression has performed better comparatively. By intertwining these analyses, we aim to provide a holistic understanding of the interplay between environmental factors and health outcomes.

For time series will be using the diagnosis as shown in table1:

Diagnosis for Time Series		
Items	**Objective**	**Expected outcome**
Autocorrelation Function (ACF) Plot	Check for autocorrelation in the residuals.	If the residuals are independent, the ACF values should not show significant spikes at lags beyond the
Partial Autocorrelation Function (PACF) Plot:	Check for autocorrelation in the residuals after removing the effects of shorter lags.	Similar to ACF, the PACF values should not show significant spikes at lags beyond the first few if the residuals are independent.
Histogram or Q-Q Plot:	Check if the residuals are approximately normally distributed.	For a normal distribution, the histogram should resemble a bell curve, and the Q-Q plot points should fall approximately along a straight
Residuals Plot:	Check if the residuals exhibit constant variance over time.	The residuals plot should show no discernible pattern, indicating homoscedasticity.

Table 1: Diagnosis for Time series data

And for Logistic Regression diagnosis we'll use
Residual Analysis: to check for residuals in logistic regression using deviance residuals.
Confusion Matrix: By Evaluating the performance of the logistic regression model on the test set using a confusion matrix.
CRISP-DM The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is a widely used framework for developing data mining and machine learning projects that consists of the following key phases: **1. Business Understanding:**
Objective is to Understand the business problem and objectives.In Time series the objective is to predict the min temperature using timer-series analysis.And in cardiac condition the aim is to predict the cardiac condition using predicting variables.
2. Data Understanding:
Objective is to Gain insights into the available data.Here I am going to perform EDA involving descriptive statistics and visualization of different features.
3. Data Preparation:
Objective is Prepare the data for modeling which involves handling null values, applying transformations like min-max and scaling and handling outliers.
4. Modeling:
Objective is to Develop and train predictive models.In this phase I'll building models to predict the target variables. For time series simple models like Naive Forecast, Seasonal naive forecast, SMA, Mean model forecast are implemented. And for Cardiac_condition prediction logistic regression is implemented.
5. Evaluation: Objective is Assess model performance as discussed above the diagnoses will be done to evaluate and again modeling is performed.
6. Deployment: Objective is to Implement the model into the operational environment. Once the models have reached certain level, the model will be used to deploy for business. In here I am not implementing this step

CRISP-DM is an iterative process, and feedback from each phase can inform adjustments in earlier stages. It provides a structured approach to guide data-driven projects from problem understanding to model deployment.

II. Exploratory Data Analysis

First step is to understand the statistics of datasets the data file are read into data frames using python in jupyter. Below are the insights drawn from each of the datasets they were provided to perform analysis on:

1) Dataset 1: Below figure 1 shows that in weather data we have total 9 columns amongst which 6 of them are of float datatype and 3 of them are of object datatype.

```
Data columns (total 9 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   date                                          29889 non-null  object
 1   maxtp(Maximum Air Temperature - degrees C)    29889 non-null  float64
 2   mintp(Minimum Air Temperature - degrees C)    29889 non-null  float64
 3   gmin(Grass Minimum Temperature - degrees C)   29889 non-null  object
 4   rain(Precipitation Amount - mm)               29889 non-null  float64
 5   cbl (Mean CBL Pressure-hpa)                   29889 non-null  float64
 6   wdsp(Mean Wind Speed - knot)                  29889 non-null  float64
 7   pe(Potential Evapotranspiration - mm)         29889 non-null  float64
 8   evap(Evaporation -mm)                         29889 non-null  object
dtypes: float64(6), object(3)
```

Figure 1: Weather data information

And Figure 2 shows the descriptive statistics on numerical columns for weather data.

	count	mean	std	min	25%	50%	75%	max
maxtp(Maximum Air Temperature - degrees C)	29889.0	13.064900	4.908828	-4.7	9.4	13.0	16.9	29.1
mintp(Minimum Air Temperature - degrees C)	29889.0	6.157051	4.383088	-12.2	2.9	6.3	9.6	18.4
rain(Precipitation Amount - mm)	29889.0	2.074720	4.396479	0.0	0.0	0.2	2.2	92.6
cbl (Mean CBL Pressure-hpa)	29889.0	1003.520208	11.723154	949.6	996.2	1004.6	1011.7	1037.4
wdsp(Mean Wind Speed - knot)	29889.0	10.198658	4.609213	0.0	6.8	9.6	13.0	35.5
pe(Potential Evapotranspiration - mm)	29889.0	1.506986	1.001506	0.0	0.7	1.3	2.2	5.7

Figure 2: Descriptive statistics on weather_data

Since I'll be dealing with minimum temperature, it can seen that minimum temperature is -12C and maximum is 18.4C Distribution of data for categorical and numerical features will help in understanding the patterns in data distribution, presence of normal distribution, linearity with respect to target variable and so on. Below are the visualization obtained for weather dataset. Though we have 3 object data it is evident that they are of type date and numeric. Hence they were transformed to these data types and then Visualization was performed on this dataset.

Through the Boxplot in Figure 3, we see that the data for Minimum Temperature is slightly skewed to the left with the middle 50% of the data ranging between 4°C. All the data points range between -5°C and 17°C. Over three-fourths of the data points have reported a daily minimum temperature of below 11°C.

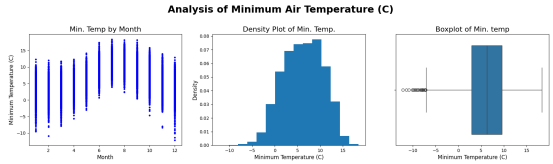

Figure 3: Analysis on minimum temperature_data

Additionally, from the month-wise scatter plot we see that most of higher minimum temperatures occur during the months of summer (June - August) with a steady fall in temperatures through Autumn and the lowest in the Winter months.

Similarly in Figure 4, we see that the data for Maximum Air Temperature is nearly evenly distributed with the middle 50% of the data ranging between 10°C and 17°C. We see an outlier at about 30°C spotted on plot. All the data points range between -1°C and 31°C. Over three-fourths of the data points have reported a daily maximum temperature of below 18°C.

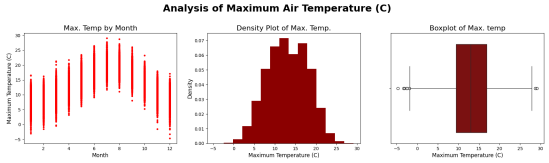

Figure 4: Analysis on maximum temperature_data

And from the month-wise scatter plot we see that most of higher maximum temperatures occur during the months of summer as indicated by the steady increase from march onwards and then a sharp spike from June to July. This is followed by a steady fall in temperatures through Autumn and the lowest in the Winter months. The outlier point seems to belong to the month of July.

In case of rainfall data, we notice that majority of the days have a reported rainfall of under 10mm. As opposed to the hypothesis drawn in the numerical summary analysis, we see here that the outlier values (days with very high rainfall) are not local to a specific season but are scattered throughout the year.

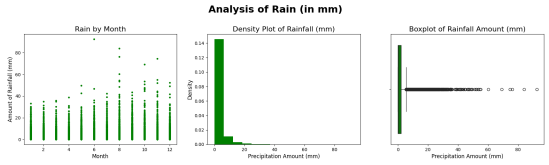

Figure 5: Analysis on rainfall_data

The density plot is extremely right skewed due to the presence of a significant number of outliers which is visible in the boxlot, as concluded through the descriptive statistics for Rainfall Amount.

In case of evaporation data, we notice that majority of the days evaporation is there which means, there is sunlight. As we see here also,we have the outlier values are not local to a specific season but are scattered throughout the year.

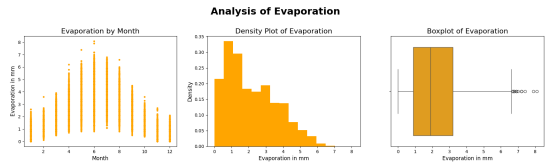

Figure 6: Analysis on evaporation_data

The density plot is right skewed due to the presence of a significant number of outliers which is visible in the boxlot, as concluded through the descriptive statistics for evaporation.

The average rainfall over year and month for all years i as shown in below Figure 7 and it can be seen that year 2023 had highest rainfall record i the give dataset.

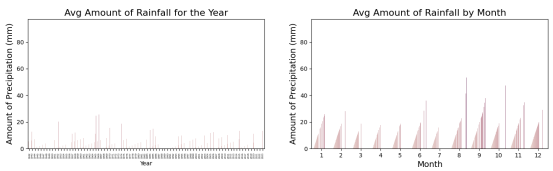

Figure 7: Analysis on average rainfall

The seasonal decomposition of min-temperature over period is as shown below.

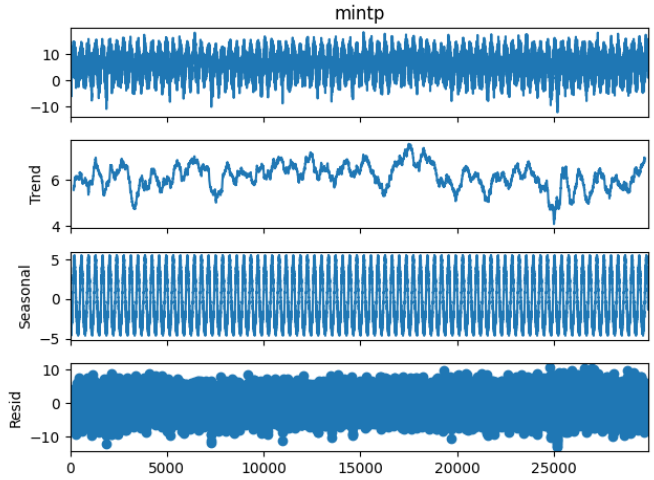

Figure 8: Seasonal decomposition of minimum temperature

Original Time Series: The original time series is shown in the upper part of the plot. This is the raw data that you are decomposing.

Trend Component: The trend component represents the long-term progression or underlying pattern in the data. It captures the overall direction in which the data is moving. Observe whether there are clear upward or downward trends.

Seasonal Component: The seasonal component represents repeating patterns or fluctuations that occur with a fixed frequency (e.g., daily, monthly, yearly). Look for regular patterns that repeat over time, indicating seasonality.

Residual Component: The residual component (also known as the remainder or error) represents the unexplained variation in the data after removing the trend and seasonal components. Check for any random or irregular patterns that might suggest noise factors.

2) Dataset 2: In Dataset 2 we have 6 columns as shown in Figure 9. In which gender column and target column are of datatype object and rest of them are float.

```
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   caseno             100 non-null    int64
 1   age                100 non-null    int64
 2   weight             100 non-null    float64
 3   gender             100 non-null    object
 4   fitness_score      100 non-null    float64
 5   cardiac_condition  100 non-null    object
```

Figure 9: Cardiac Condition Dataset Info

Figure 10 shows the descriptive statistics for dataset 2

	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

Figure 10: Cardiac Condition Dataset Info

The minimum and maximum values for different features have high variant values and we might need to perform scaling transformation. Let's try to explore the data with visualization. Figure 11 shows the maximum and minimum values with respect to age, weight and fitness score.

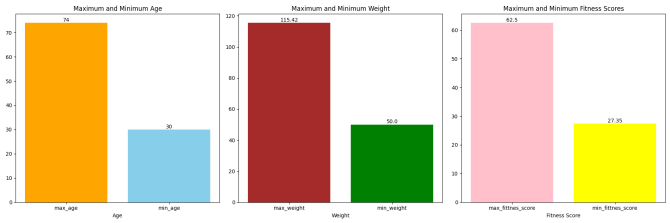

Figure 11:Max and min values for age,weight and fitness.

Figure 12 shows the distribution of age, weight and gender with respect to cardiac_condition. It's showing that age group 40-49 is more prone to cardiac condition, weight ranging from 90-99 has slightly higher chances of having cardiac condition and Male are prone to have cardiac disease with 77.1%

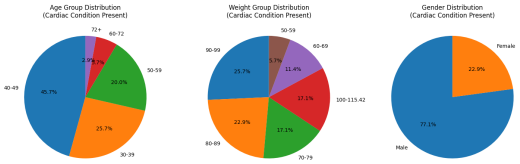

Figure 12: Distribution of different feature with respect to cardiac condition

Figure 13 shows the distribution of numeric variables and counts for chacter variables

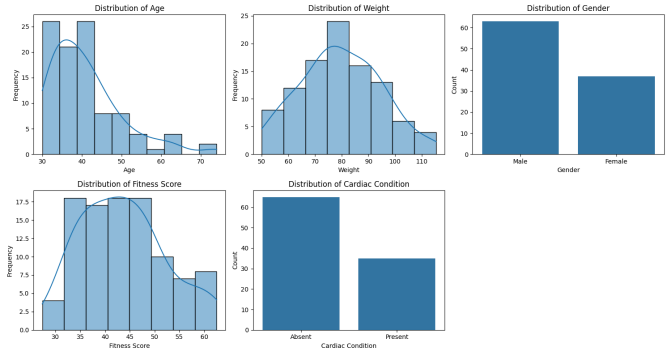

Figure 13: Data distribution in dataset

We can see that except for age column which is slightly right skewed, other numeric columns have almost normal distributed data. and Male count is dataset is higher than female count also the target column has highest instances for absent and less for present.

Figure 14 shows the correlation of all features with cardiac condition, I have converted the gender column values to numeric by label encoding 0 as Female and 1 as Male.

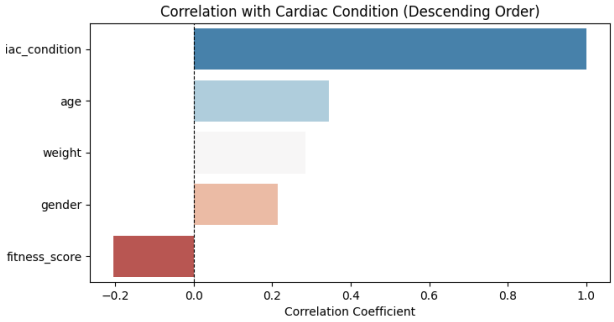

Figure 14: Correlation matrix with respect to target

All the features have significant correlation value with target variables.

Figure 15 shows the box plot for different features in dataset. The figure depicts that there are very less number of outliers.

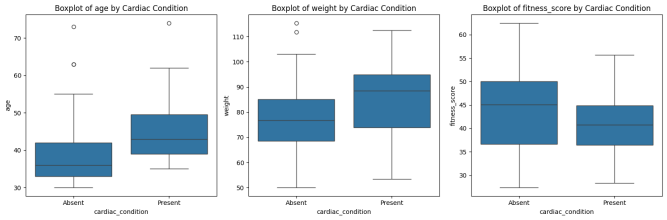

Figure 15: Cardiac Condition Dataset Box-plot

The box plot for age by cardiac shows that the cardiac condition might start to be present from the age of 35 with mean, median ranging from 40 to 50 age.Similarly with respect weight the cardiac condition can be seen from weight range of 50 with most cases seen for those weighing between 75 to 95. And those who have fitness score in the range of 36-44 are prone to cardiac condition.

Figure 16 shows the age, weight and fitness score distribution by gender.

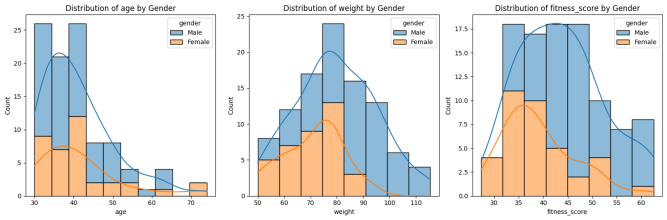

Figure 16: Gender wise age, weight and fitness score distribution

It can be seen that in all the cases the female count is less compared to male count.

Overall, pair plots are an effective and efficient way to gain a visual understanding of the interactions between variables in dataset. And Figure 17 shows the pair plot for our dataset

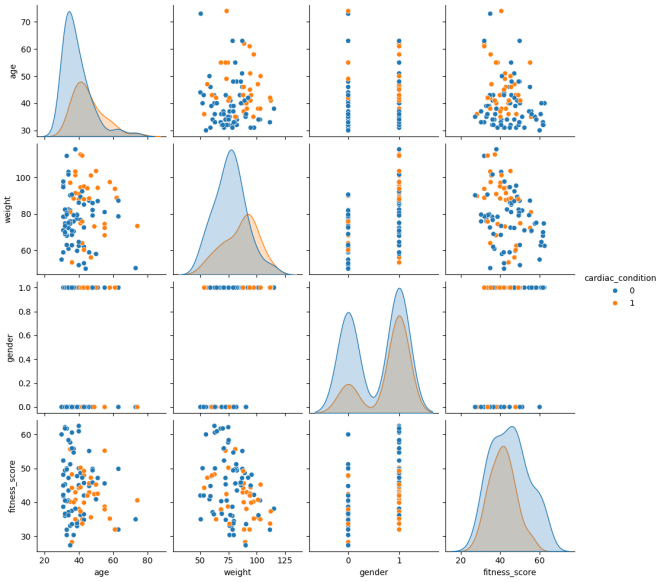

Figure 17: Pair plot for dataset2

III. Modeling

1) Dataset 1:: First we need to check f the data is stationary or not using methods like Augmented Dickey-Fuller test.If there is uncertainties found as part of formatting we might need to do difference.Below Figure 18 shows the result of ADF test.

```
ADF Statistic: -12.983652321952396
p-value: 2.9084994740401857e-24
formatted p-value: 0.000000000000000000
```

Figure 18: ADF test result.

ADF Statistic: The ADF statistic is -12.98. In the context of the test, a more negative ADF statistic provides stronger evidence against the null hypothesis of a unit root (non-stationarity). The more negative it is, the more likely it is that the time series is stationary.

p-value: The p-value associated with the ADF statistic is very close to zero (2.91e-24). The p-value is compared to a significance level (commonly 0.05). In this case, the p-value is significantly less than 0.05, leading to the rejection of the null hypothesis.

Formatted p-value: The formatted p-value is written as 0.00. This is essentially zero, indicating an extremely low probability of observing the data under the assumption that the series has a unit root.

In summary, based on the ADF test results,has statistical evidence to support the stationary of the time series.

We are good to train our model with datset what we have and no need to do any difference as the data is stationary. Simple Time series Model **Auto-regressive (AR) Model:** AR Predicts the next value based on linear combination of past values. AR(p) uses the past p values.Here I have given P value as 365 i.e an year data. And the below output was obtained as shown in Figure 19:

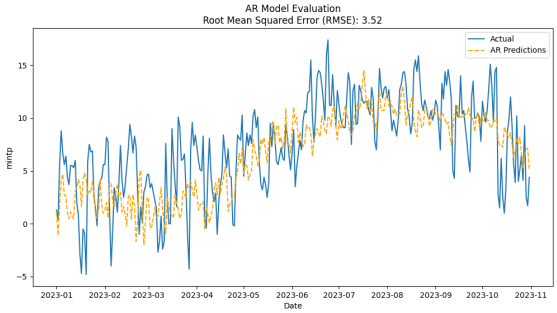

Figure 19: AR evaluation

Figure 20 shows that SMA forecast and seasonal nave forecast have performed better when compared to all 4 and season naive forecast is best among all.

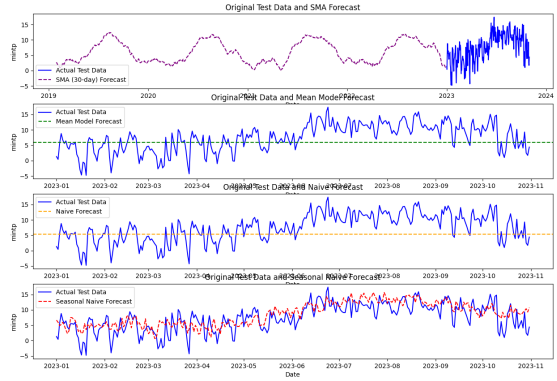

Figure 20: Different Simple models output.

Naive Forecasting: Assumes that future values will be the same as the most recent observed value. It's a simple and intuitive method.Here it has considered the last value of train_data

Mean Model: Assumes that future values will be the mean of the training set. It is a basic benchmark model.Mean value of train_data is considered

Seasonal Naive: Assumes that the next value will be the same as the value observed in the previous season. Useful for data with strong seasonal patterns.the p value here is 365

Simple Moving Average (SMA): Uses the average of the last 'window_size' values to forecast future values. It smoothens out fluctuations model is trained with window size 30.

Seasonal Naive Forecasting: Assumes that future values will follow the same seasonal pattern as the most recent observed season, making it suitable for time series with recurring seasonal trends.Here the period of 365 is considered

Figure 20 shows the wma predictions over actual data.

textbfWeighted Moving Average (WMA): Calculates a weighted average of the past 'window_size' values with weights increasing linearly. It provides a smoothed trend over time.I have given window size of 30 when window size was increased to 365 the prediction was not as accurate when it was 30.

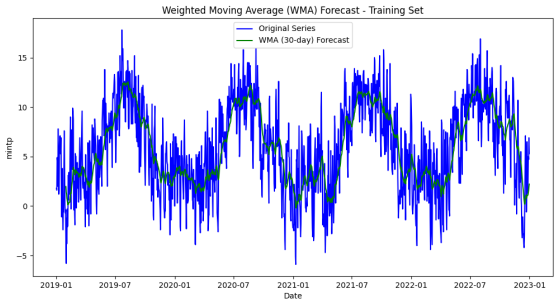

Figure 21: WMA model output.

Figure 21 shows Seasonal exponential smoothing model. **Simple Exponential Smoothing (SES):** Applies an exponential smoothing model to capture the underlying trend and make predictions based on exponentially weighted past observations. It's suitable for time series with a constant or slowly changing mean. Since our data doesn't contain constant or slowly changing mean, this model will perform low on our dataset.

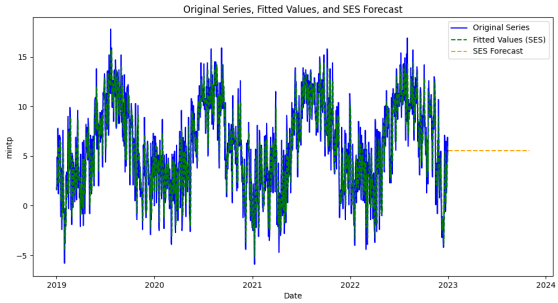

Figure 22: SES model output.

In simple words, the Holt forecast is a prediction made using the Holt-Winters method, a type of time series forecasting technique. The Holt-Winters method considers three components in the data: the average level, a trend indicating the direction of change, and seasonality representing repeating patterns.

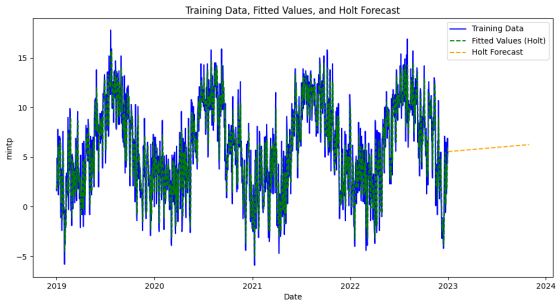

Figure 23: Holt model output.

The forecast, generated by the Holt-Winters method, incorporates these components to estimate future values in a time series.

ARIMA: To train the data for ARIMA model we first need to find p, d, q p values which we decide by seeing ACF and PACF

In the ACF plot, look for the point where the auto correlation values drop significantly. This may suggest the parameter q for the ARIMA model. In the PACF plot, look for the point where the partial autocorrelation values drop significantly. This may suggest the parameter p for the ARIMA model. Choose p and q based on where the ACF and PACF plots first cross the significance threshold (dotted lines). Figure 24

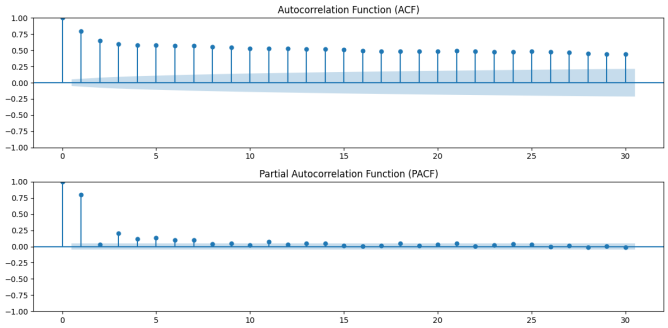

Figure 24: ACF and PACF

So as seen in 24 the q value will be 2 and p value will be 1 d is 0 as already explained. Figure 25 shows the test data ARIMA forecast result.

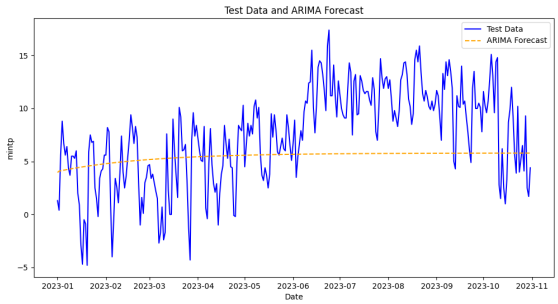

Figure 25: ARIMA forecast on test data.

Figure 26 shows the future forecast data for 5 years. The prediction made by all simple and ARIMA model.

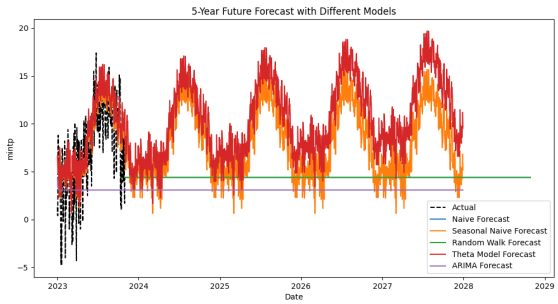

Figure 26: Future forecast data.

2) Datset 2: Once all the columns were encoded as explained in pre-processing, the the model training was done. Model 1 Without dropping any columns just after doing preprocessing trained the linear regression model to predict the cardiac condition. And the result was as shown in Figure 27:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.71      0.75        17
           1       0.50      0.62      0.56         8

    accuracy                           0.68        25
   macro avg       0.65      0.67      0.65        25
weighted avg       0.70      0.68      0.69        25

Coefficients: [[ 0.16070634 -0.00409535  1.31185807 -0.11286947]]
Intercept: [-2.71066459]
```

Figure 27: Model 1 result.

With accuracy of 68%

Model 2 Applied Min-max scaling on numerical features ignoring the categorical features which hold ordinal numerical values. I trained the model again and results were as shown in figure 28.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.82      0.78        17
           1       0.50      0.38      0.43         8

    accuracy                           0.68        25
   macro avg       0.62      0.60      0.60        25
weighted avg       0.66      0.68      0.67        25

Coefficients: [[ 1.74230071  0.55812895  0.87371374 -1.27879408]]
Intercept: [-1.24837324]
```

Figure 28: Model2 result.

With accuracy of 68% But this showed no much difference just there was increase in the recall value. Model 3 In model 3 I am trying to balance the training data with equal amount of classification values. That is right now the target variable count is
cardiac_condition 0 65 1 35 in which training data target column count is:
cardiac_condition 0 48 1 27 Using SMOTE I am trying to oversample the target column data and creating a balanced train data and tried to build model3. The result is as shown in figure 29.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.65      0.76        17
           1       0.54      0.88      0.67         8

    accuracy                           0.72        25
   macro avg       0.73      0.76      0.71        25
weighted avg       0.80      0.72      0.73        25

Coefficients: [[ 2.34169417  0.60000345  0.93393207 -1.47107546]]
Intercept: [-0.85186053]
```

Figure 29: Model 3 result.

The precision values has increased a lot but recall values has been reduced and overall accuracy has increased to 72%.

Model 4 In model 4 I am trying to remove fitness score and check if model would perform better.But, no major difference were found

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.65      0.76        17
           1       0.54      0.88      0.67         8

    accuracy                           0.72        25
   macro avg       0.73      0.76      0.71        25
weighted avg       0.80      0.72      0.73        25

Coefficients: [[2.44756675 1.04489967 0.61405532]]
Intercept: [-1.56595243]
```

Figure 30: Model 4 result.

with accuracy being 72%

IV. INTERPRETATION

Dataset 1:
When working with time series forecasting models, rather than having a separate intercept term, we interpret the constant or trend components of the model in the context of the time series being analyzed. Below table shows

Models	ARIMA Forecast Test Set Metrics:	Holt Forecast Test Set Metrics:	Simple Exponential Smoothing Test Set Metrics:
Mean Squared Error	21.92895	21.286079	23.494219
Root Mean Squared Error (RMSE):	4.682836	4.613684	4.847084
Mean Absolute	3.928666	3.84708	4.038975
Mean Absolute Percentage Error	1.57466E+14	1.6958E+14	1.64456E+14
Mean Error (ME):	1.951308	1.515616	1.873177
Mean Percentage Error (MPE):	26.2849	20.41595	25.232449
Mean Absolute Scaled Error	1.79194	1.754727	5.195155

Table 2: Evaluation Results of different models

Dataset 2:
Accuracy, Confusion Matrix, and Classification Report:
Please refer Table 3 for the accuracy , confusion matrix, and classification report to understand how the models are performing. As we can see Table 2 shows the performance for for all the 4 models that were trained with their interpretation metrices.

	MODEL 1(encoded data)			MODEL 3(balanced data)				MODEL 2(Scaled data)			MODEL 4(balanced and removed fitness)			
Accuracy	0.68			0.72				0.68			0.72			
Confusion matrix	[[12 5]			[[11 6]				[14 3]			[[11 6]			
	[3 5]]			[1 7]]				[5 3]]			[1 7]]			
Classification Report:	precision	recall	f1-score	precision	recall	f1-score	support	precision	recall	f1-score	precision	recall	f1-score	support
0	0.801	0.71	0.75	0.92	0.65	0.76	17	0.74	0.82	0.78	0.92	0.65	0.76	17
1	0.501	0.62	0.56	0.54	0.88	0.67	8	0.5	0.38	0.43	0.54	0.88	0.67	8
accuracy			0.68			0.72	25			0.68			0.72	25
macro avg	0.65	0.67	0.65	0.73	0.76	0.71	25	0.62	0.6	0.6	0.73	0.76	0.71	25
weighted avg	0.7	0.68	0.69	0.8	0.72	0.73	25	0.66	0.68	0.67	0.8	0.72	0.73	25
Coefficients:	[[0.16070634 -0.00409535 1.31185807 -0.11286947]]			[[2.34169417 0.60000345 0.93393207 -1.47107546]]				[[1.74230071 0.55812895 0.87371374 -1.27879408]]			[[2.44756675 1.04489967 0.61405532]]			
Intercept	[-2.71066459]			[-0.85186053]				[-1.24837324]			[-1.56595243]			

Table 3: Different models interpretation summary

V. Evaluation

In this section we are going to evaluate the models that we have built and see how they are performing based on different factors for each dataset.

1) Dataset 1:: As an evaluation step we are now going to check the if Residuals are Normally Distributed: The histogram of residuals should resemble a bell-shaped curve, indicating normal distribution.
A Q-Q plot to compare the distribution of residuals against a theoretical normal distribution.

There is significant change in the qq plot for exp_smoothing and also as part of evaluation the Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF) of residuals. There should be no significant spikes beyond the confidence intervals, indicating no remaining patterns. and figure 29 confirm the same.

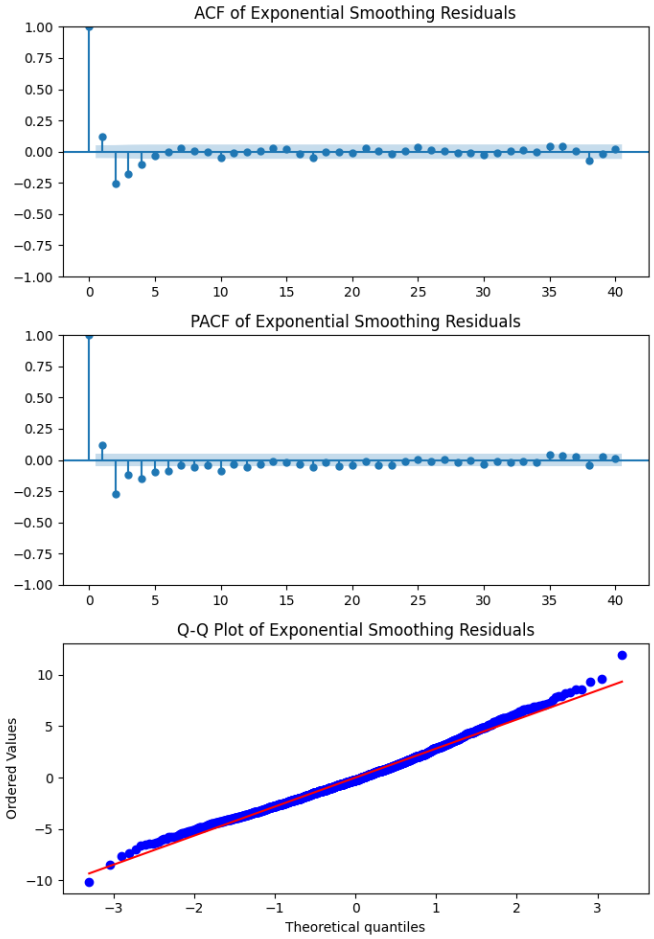

Figure 31: Evaluation of Exponential smoothing.

Figure 32 shows evaluation of Naive Forecast, even there we can see that the ACF and PACF have no significant spikes indicating no remaining patterns present.

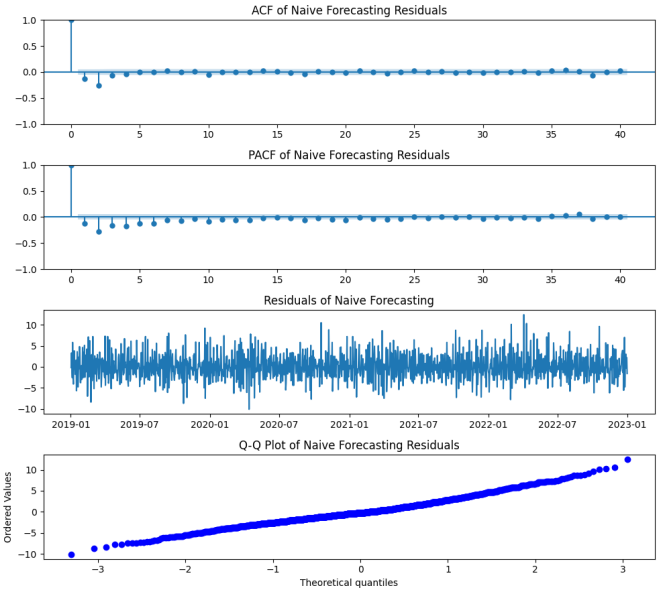

Figure 32: Evaluation of Naive forecast.

Similarly for ARIMA model AC and PACF has shown no significant spikes and also the residuals histogram is normally distributed resembling bell shaped curve. Figure 33 shows evaluation of ARIMA model

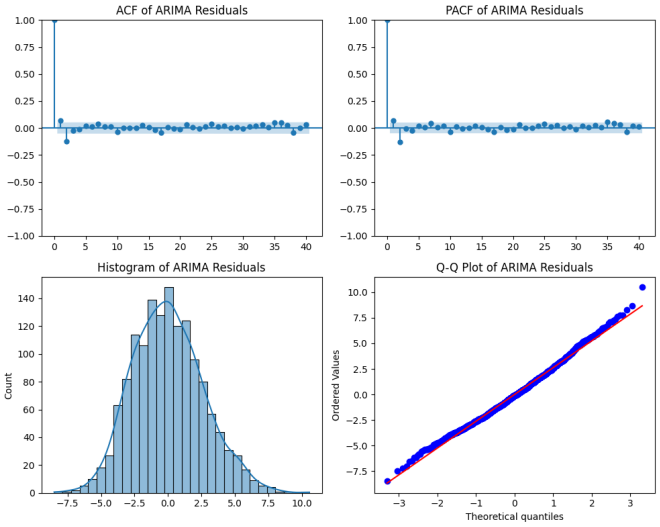

Figure 33: Diagnosis of ARIMA.

2) Dataset 2:: As part of evaluation we have already shown the confusion matrix along with other interpretation values. Now I would like to show the ROC curve as one more evaluating process. It's clearly evident that both Model 3 and Model 4 are performing better with good AUC value.

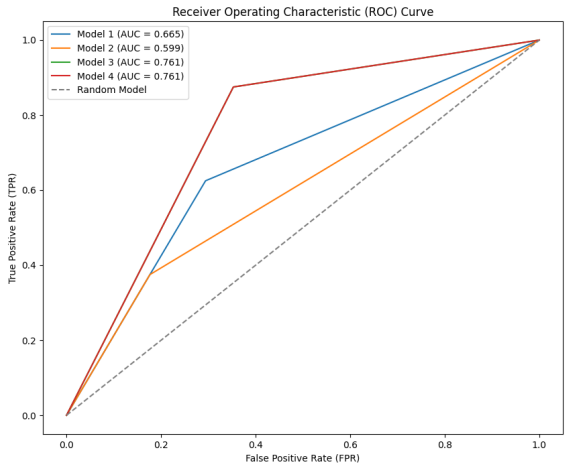

Figure 34: ROC curves for different models

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a model's ability to distinguish between two classes. It shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across different threshold values. In simple terms, the ROC curve helps visualize how well a model can separate positive and negative instances, allowing you to choose an optimal threshold that balances true positives and true negatives based on your specific needs. A curve closer to the top-left corner indicates better overall performance.

VI. Conclusion

1) Dataset 1: When examining weather data, I have used a thorough time series approach to check for seasonality and stationary. I tried different models like naive, mean, seasonal naive, drift, and linear without favoring any particular one. Additionally, we explored exponential smoothing methods such as SES to handle different data patterns, and I evaluated ARIMA as well. Using these various models improved our understanding of temporal complexities, ensuring accurate forecasts and a strong understanding of the underlying patterns.

2) Dataset 2: The results of Model 3 suggest that it has a good balance between precision and recall, with an accuracy of 72%.

Precision (Positive Predictive Value): The model correctly identified 54% of the positive cases (cardiac_condition = 1) out of all cases predicted as positive. In other words, when it predicts a positive case, it is correct about 54% of the time.

Recall (Sensitivity or True Positive Rate): The model captured 88% of the actual positive cases. In other words, it correctly identified 88% of the individuals with cardiac conditions out of all individuals with cardiac conditions.

F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. The F1-score here is around 0.67, indicating a reasonable balance.

Confusion Matrix: The confusion matrix shows the model's performance in terms of true positives, true negatives, false positives, and false negatives. It reveals that the model correctly predicted 11 instances of "No cardiac condition" and 7 instances of "Cardiac condition." However, it made 6 false positive predictions and 1 false negative prediction.

In summary, Model 3 seems to be a reasonable choice, achieving a good trade-off between identifying individuals with cardiac conditions and avoiding false positives. **Oversampling Impact (Model 3):** The model has performed better and also the AUC value is more when the data is trained using balanced dataset, with Accracy of 72% from 68%. Please refer Table 2 for models interpretation comparison

References

[1] Shobit Agrawal and Dilip Kumar Sharma. Feature extraction and selection techniques for time series data classification: A comparative analysis. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 860–865, 2022.

[2] Anurag, Shreya Kalta, Jatin Thakur, Himanshu Bhardwaj, and Yogesh Banyal. Optimized ensemble model for heart disease prediction using machine learning. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–5, 2023.

[3] U. Ashwini, K. Kalaivani, K. Ulagapriya, and A. Saritha. Time series analysis based tamilnadu monsoon rainfall prediction using seasonal arima. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1293–1297, 2021.

[4] S. Chua, V. Sia, and P. N. E. Nohuddin. Comparing machine learning models for heart disease prediction. In *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 1–5, 2022.

[5] KangWoon Hong and Taegyu Kang. A study on rainfall prediction based on meteorological time series. In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 302–304, 2021.

[6] Narendra Mohan, Vinod Jain, and Gauranshi Agrawal. Heart disease prediction using supervised machine learning algorithms. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–3, 2021.

[7] Shuge Ouyang. Research of heart disease prediction based on machine learning. In *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pages 315–319, 2022.

[8] Irfan Pratama, Putri Taqwa Prasetyaningrum, and Putry Wahyu Setyaningsih. Time-series data forecasting and approximation with smoothing technique. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 439–444, 2019.

[9] Dara DVVNS Saikumar, Uma Priyadarsini P. S, and I. Meignana Arumugam. Prediction of heart disease using k-nearest neighbour algorithm in comparison with support vector machine algorithm. In *2022 14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, pages 1–8, 2022.

[10] Shashi Kant Tiwari, Shweta Sinha, Karamjit Kaur, and Ram Sewak Singh. Prediction of heart diseases using nonlinear techniques and machine learning. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 530–535, 2023.