Statistics for data analytics

Akash pal student id x22211420
National College Of Ireland
MSc in Data Analytics Semester 1, 2023/2024

***Abstract*—This report is about two different studies where we have used computational methods to understand the data. In the first study we have gone through the information about people's health and try to find out the if they have heart problems or not .We have used Logistic regression and after that we have used other statical techniques to see how good our prediction model is , we have also check whether we are correct or incorrect or when can we do more better to get better results** .

***Keywords*—logistic regression, ARIMA/ SARIMA.**

I. Introduction

These reports will give two studies that was done for healthcare and the environment. The first study is about predicting heart problems. Nowadays young age group have affected for cardiac arrest worldwide, it's important to figure out what causes them. We have used data about people's age, weight, gender, how fit they are and whether they have any heart problems. We have used logistic regression to build a model. The goal was to see if these health things are connected to having heart problems. This can basically help us to make plans to prevent heart problems and make health rules and this can save life for many people.

II. Methodology

Logistic regression analysis of Cardiac conditions

A. Dataset description & understanding the dataset

The dataset aligns with our goals and data is in csv format and which includes 6 variables with a size of 100.

Variable	Type
Caseno	int
Age	int
Weight	float
Gender	object
Fitness_score	float
Cardiac_condition	Object

Table 1 : dataset variables

Each variable has is defined as follows.

- *Caseno : Case number ,likely an identifier for each record*
- *Age: Age of the individual*
- *Weight: weight of the individual*
- *Gender: Gender of the individual (Male or female)*
- *Fitness_score: A score presumably related to individual fitness level.*
- *Cardiac_condition: Indicate whether a cardiac condition is present or absent.*

In order to check the quality of data we try to check missing values using isnull() in python and here are the results

Variable	Missing values
Caseno	0
Age	0
Weight	0
Gender	0
Fitness_score	0
Cardiac_condition	0

Table 2 : missing values table

As we don't have any missing values which makes it easier for model building, if there were any missing values, we would have handled using mean, median or mode depending on our requirements.

B. Visualization of the dataset

As we proceed further for data preparation it's useful to get insights from data to identify any outliers and understand the distribution of these variables. We have created histograms for variables which are continuous as age, weight, fitness score & bar plots for categorical variables as gender, cardiac condition. This will basically help us to understand whether we are required for dimensionality reduction techniques as PCA are necessary and appropriate.

Distribution of age:

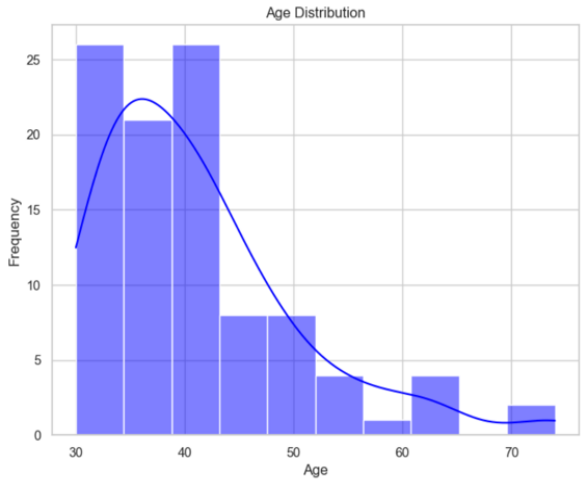


Fig 1

From fig 1 it depicts about age distribution

Age: Ranges from 30 to 75 years approximately

Fig1 clearly states that the distribution appears somewhat uniformly distributed with a little increase in frequency around 30-to-40-year age range.

Distribution of Weight:

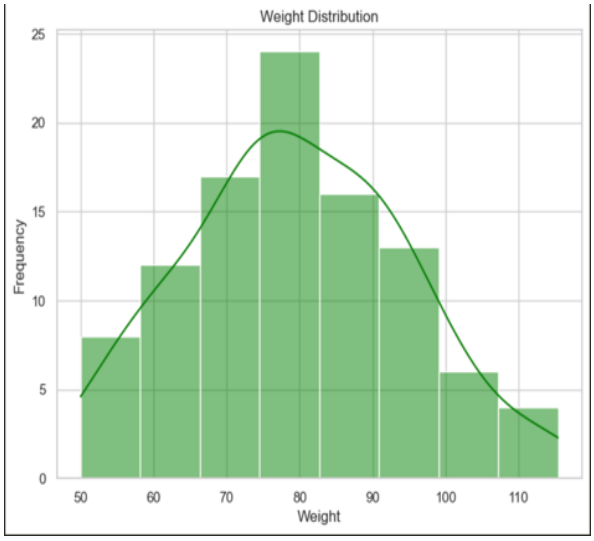


Fig 2

From fig 2 it depicts about weight distribution

Weight: it varies from 50 to 115 units

Fig2 clearly states that the distribution is roughly normal with a little skew to the right.

Distribution of fitness score:

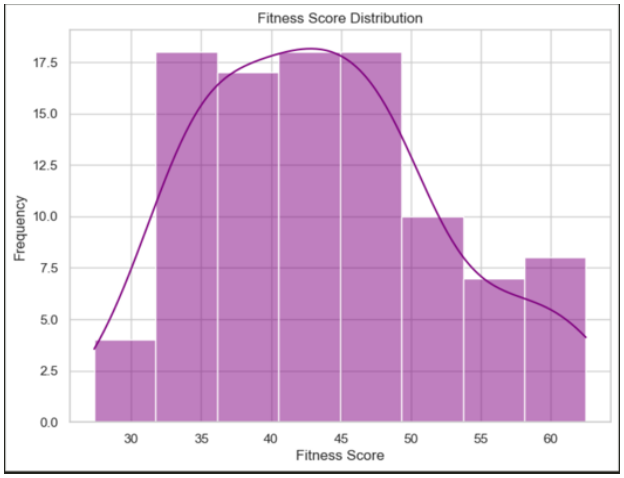


Fig3

From fig 3 it depicts about fitness score distribution

Fitness score : ranges from 28 to 65 approximately

Fig3 also appears that somewhat normally distributed, with a little skew to the left.

Distribution of Gender :

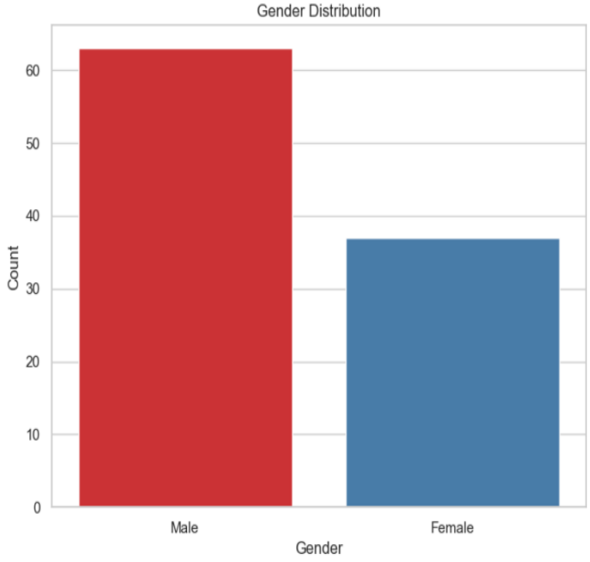


Fig 4

From fig 4 it depicts about gender distribution

Gender: Male & female

From fig 4 males are more frequent than females.

Distribution of cardiac condition:

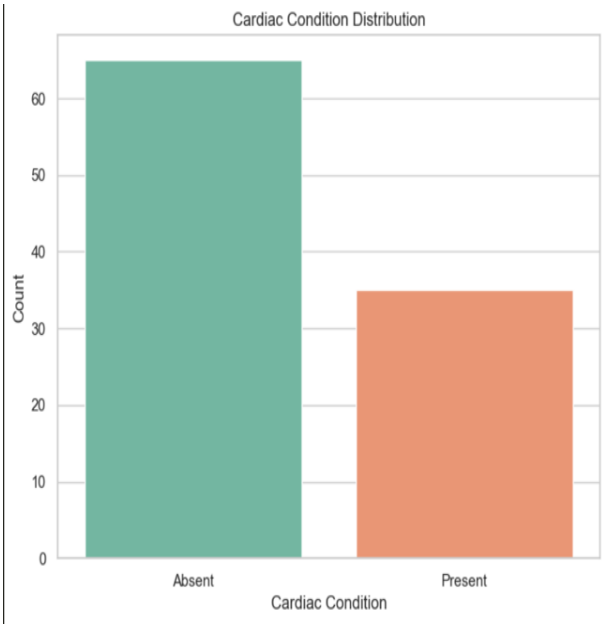


Fig5

From fig 5 it depicts about cardiac condition distribution

Cardiac condition: it's a binary variable with 'absent' & 'Present.'

From fig.5 a larger proportion of the dataset does not have a cardiac condition('Absent.')

To summarize , it doesn't appear to have any extreme outliers that would significantly skew our analysis. The visuals from fig 1 to fig 5 the distributions look reasonable for the variables involved. In addition to that we are not proceeding with dimensionality reduction as we have less data.

c. Preparation of data for modelling

Now since we have got our insights from above, we can proceed with preparation of data for modelling which includes encoding the categorical variables(gender & cardiac condition) & splitting of the dataset into training and test sets using random seed which is equal to our student number which is 22211420.

Since logistic regression requires numerical inputs , we will convert the categorical variables to numerical ones *(see table 1)*.The 'gender' variable will be converted into binary variable (0&1) and for 'cardiac_condition' we are representing 'present' as 1 and for 'absent' as 0 and for gender male is encoded as 1 female as 0 . For this encoding we are using label Encoder from sklearn. We are keeping our test size as 0.3 which means that 70% of our data will be training set and the remaining 30% will be test data.

Table 3 Training & test data table

Dataset	**Size**
Training set	70%
Test set	30%

d. Modelling Phase

Now we will further proceed with modelling phase. As we have converted our categorical variables into binary, we can fit our binary logistic regression model to the training data which is 70% *(see table 3).* This model will now help us to understand the relationship between the various factors (age, weight,gender,fitness_score) and the presence or absence of cardiac condition.

e. *Evaluation*

After the logistic regression is trained here are some evaluation

1) *Confusion matrix*
True negatives (Absent correctly identified):17.
False Positives (Absent incorrectly identified as present):4.
False negatives (Present incorrectly identified as absent):5.
True Positives(Present correctly identified):4

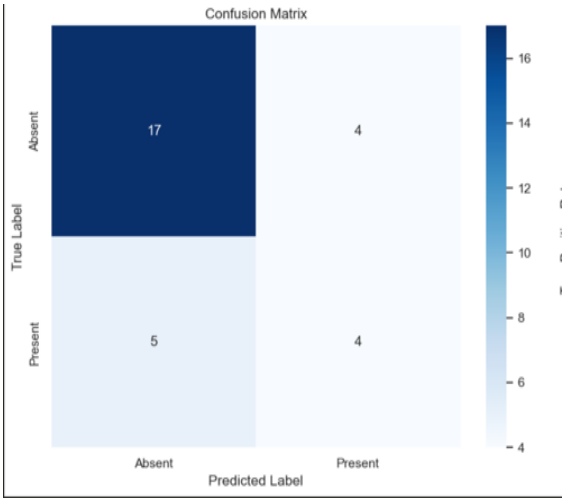


Fig 6

From fig 6 with the heat map of the confusion matrix model appears to be a better fit for identifying cases where the cardiac condition is absent (true negatives) compared to when it is present (true positives)

The heatmap makes it easier to see the number of observations for each combination of predicted and actual values.

2) *Roc-Auc Score* :
The Roc-AUC score is approximately 0.63, this basically represents the model's ability to distinguish between two classes i.e. absent & present for cardiac condition . A score of 1 represents a perfect classifier while a score of 0.5 represents a random guess.

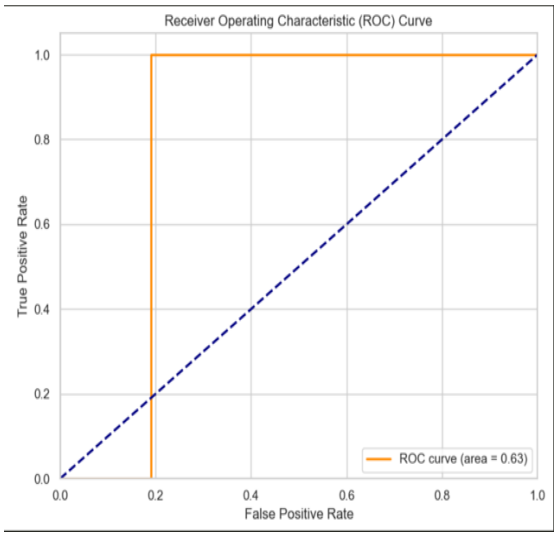


Fig7

From fig 7 the roc curves give an relationship of true positive rate(TPR) and false positive rate (FPR) at various thresholds.

The curve is above the diagonal line which represents random guess, but it can be improved.

3) *Coefficients:*

The coefficients of logistic regression model provide the change in the log odds of the outcome for a one unit increase in the predictor variable and the results can be taken from our trained model below is the table which shows the value of coefficients of the variable.

Variable	Coefficients
Age	0.115197
Weight	0.001450
Gender encoded	1.287771
Fitness score	-0.075669

Table 4 for coefficients

From table 4 here are the explanations:

Age :0.115 – when a person's age goes up by one year their chances of having a heart problem increases by about 0.115, assuming everything else stays the same.

Weight:0.001 –if someone gains weight it has very less effect on their chances of having heart problem.

Gender:1.288— Being a man which is encoded as 1 is linked to a higher chance of getting a heart problem compared to woman all other things being equal.

Fitness Score:-0.076— if a person has a higher fitness score their chances of having heart problem goes down.

4) Accuracy:

the model is right about 70% of the time when it predicts whether someone has any heart problem or not on a test set , it means that it got correct for 70 out of every 100 cases

f) Conclusion

a) Model performance:

Our model did okay –it got things right about 70% of the time when guessing if someone has heart problem or not , that's like 7 out of 10 questions correct . It's not perfect but not too bad either. We also looked at something called roc-auc (see fig7) which tells us how well our model can separate people with and without heart problems.

b) Influential factors :

- ***age***: the older you are the more likely you might have a heart problem ,if everything else stays the same.
- ***Weight***: if you gain weight , it might take you little towards heart problem.
- ***Gender***: Guys are more likely to have heart problems.
- ***Fitness score*** : if you are better in shape chances are less having a heart problem.

c) Data distribution insights:

from fig 1,fig2, fig3 ,fig 4& fig 5 gives a clear picture of how many guys or girls , how old people are, and how healthy they are . It helps us understand who's in the group and how healthy they are.

d)Interpretability vs model complexity:

we didn't use complicated methods to make our model easier to understand. We could have, but the data wasn't too messy, so we kept it simple.

e)Future considerations:

we can make our model better by adding more stuff or trying different ways to guess . Also, we should think about if there are more people with heart problems or without them because that can affect our predictions.

f) Limitations:

we're only looking at the data we have. There might be other things we don't know about that can also affect heart problems. This helps us to understand what things are crucial for health but remember we should learn more from the real world of medicine.

Time series forecasting on maximum air temperature.

Abstract— ***In this study we are focusing on "Maximum Air Temperature" over time. We basically want to see if we can predict what temperature will be like in future. Over here we used time series models as exponential smoothing and ARIMA/SARIMA techniques. We considered to have a view on the data from 2019 to 2023 to check which methods is best for predicting temperatures accurately***.

III. Introduction

The second study is about the weather. Our focus was to look at the hottest air temperature, which is crucial to understand the relationship of climate change. Basically, we wanted to understand and predict how cold the air will be in future. We have used methods smoothing &ARIMA/SARIMA. This will help us to know about the climate now and getting ready for hot or cold climates that can affect different things. This data was recorded by met Eireann which is one of the Ireland weathers stations

IV. Methodology

Time series analysis

A. Dataset preparation

We got a bunch of temperature data from Dublin airport ,starting from way back in 1942 until 2023. We made sure that date part looked nice ,like a proper date , so we could study how temperature change over time. We also picked out the main thing we care about , which is highest temperature each day , called 'maxtp'.

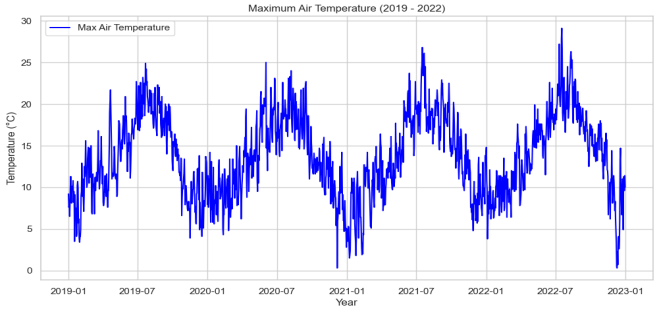
B. Preliminary Assessments

Dataset overview : the dataset that we had include various weather metrics including the maximum air temperature which is our variable interest. The data ranges over nearly 5 years and includes 1762 observations .Each value in the dataset corresponds to a specific day and includes the highest recorded air temperature for that day in degree Celsius.

Preliminary data analysis : when we explored the data to check the quality , missing values or anomalies there was nothing to find out because our data doesn't have any of these which is great positive sign for analysis.

Visual inspection: The fig8 shows distinct seasonal patterns. this reflects a typical temperature maritime climate of Ireland.

Fig8



Seasonal variation : the fig 8 shows the cyclical nature of the temperatures with higher temperature in the summer months and lower temperature in winter describing the seasonal changes over the years.

Trend analysis: fig 8 shows no distinct long-term upward or downward trend is visible over these four years which shows us stable temperature patterns within this frame.

Fig9

Fig 9 depicts time series decomposition of the maximum air temperature(maxtp):

1)**original** series : this is actual maxtp over time.

2)**trend**: appears to be a long-term fluctuation in the maxtp . some periods show an upward trend while others show downward trend. It might indicate different climatic patterns over the years.

3)**Seasonality** : a clear season pattern is evident which is expected for temperature data. This indicates regular fluctuations in temperature across different times of the year reflecting the changing seasons.

4)**residuals** : these are irregular components that cannot be attributed to trend or seasonality , this could either include random weather events or other anomalies.

Stationary check

We used a test called augmented dickey fuller (ADF) test to see if the temperature data is stable. Being stable is important for many temperature models. The pvalue is 3.420616863504843e-19. This is extremely small p-value which concludes that the time series of maximum air temperature(maxtp) is stationary.

Autocorrelation analysis

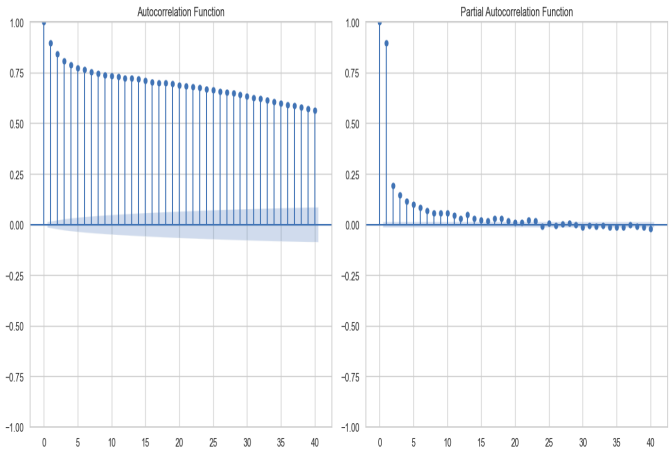


Fig 10

This helps us to understand If the maximum air temperature data is correlated with its own past values.

1) Auto correlation function (ACF): The ACF plot shows a gradual decline in the correlation as the lag increases. This indicates that the maximum air temperature on a given day is positively correlated with its past values, with the strength of the correlation decreasing over time.

2) Partial autocorrelation function(PACF): The PACF plot shows significant partial autocorrelations at the initial lags, and then these correlations largely fall within the confidence interval (suggesting they are not statistically significant) as the lag increases.

This pattern suggests that only the most recent past values have a significant linear influence on the current value of the maximum air temperature.

we can conclude that the maximum air temperature exhibits a significant amount of autocorrelation.

V. Modelling

To evaluate the performance of these models, we used the data from 2019 to 2022 for training and the data from 2023 for testing. The Mean Squared Error (MSE) was used as the metric for comparing the accuracy of the models. A lower MSE value indicates a better fit of the model to the data.

- Simple time series models
- Exponential smoothing
- ARIMA/SARIMA models

1) Simple Time series modelling :

For this we have used Naïve model and simple moving average (SMA) :
Naïve MSE = 48.07
SMA MSE = 84.58

2) Exponential Smoothing :MSE : 9.18

3). SARIMA(seasonal ARIMA): MSE :57.19

Discussion & Model selection

- The exponential smoothing model has the lowest MSE, indicating it performed the best in forecasting the maximum air temperature for 2023.
- The naïve model surprisingly outperforms the SMA and SARIMA models in terms of MSE. This might be due to the inherent characteristics of the time series data, where the last observed value is a reasonable predictor for the short-term future, especially in a stable and seasonal series like temperature.
- The SMA model has the highest MSE, suggesting it is the least suitable model among those tested. This could be due to the averaging window not capturing the seasonal fluctuations effectively.
- The SARIMA model while traditionally robust for seasonal data, did not perform as well as expected. This might be due to the choice of parameters or the specific characteristics of this dataset.

Optimum model for forecasting:

Based on the MSE values the exponential smoothing model is the best for forecasting maximum air temperature in our dataset. This model ability is amazing because it showed us both trend and seasonality.

Conclusion

The primary goal was to understand the underlying patterns in the data and to predict future temperatures using the most suitable model. After a detailed analysis, the key conclusions were drawn.

The Exponential Smoothing model, has handle both the trend and seasonality in the data, proved to be the most effective, as provided by its lowest Mean Squared Error (MSE) in our prediction for the year 2023. This finding shows the importance of choosing a model that aligns well with the specific variables of the dataset, particularly when dealing with seasonal data like temperature records.

On the other hand, the Simple Moving Average (SMA) and the SARIMA models did not perform as well as expected. The SMA model's high MSE showed its limitations in predicting time series data with strong seasonal patterns. Additionally, the SARIMA model's moderate performance suggested that while it is a powerful tool for handling seasonal data, the selection of its parameters can be challenging and crucial for its success.

The Naïve model, despite its simplicity, provided a good prediction, signifying that sometimes basic models can serve as effective importance or starting points, especially in time series data exhibiting less volatility.

Implications & future works

It basically demonstrates the necessity of a thorough understanding of the data's characteristics before model

selection and highlights the importance of Exponential Smoothing in handling seasonal temperature data.

For future work, we recommend exploring more techniques for parameter optimization, especially for models like SARIMA. Additionally, looking at external variables such as climate change indicators that can enhance the accuracy of the prediction.

REFERENCES

[1] Dodek, Peter M., and Barry R. Wiggs. "Logistic regression model to predict outcome after in-hospital cardiac arrest: validation, accuracy, sensitivity and specificity." *Resuscitation* 36.3 (1998): 201-208.

[2] Kabbilawsh, P., D. Sathish Kumar, and N. R. Chithra. "Trend analysis and SARIMA forecasting of mean maximum and mean minimum monthly temperature for the state of Kerala, India." Acta Geophysica 68.4 (2020): 1161-1174.