Study/Implementation of Time Series Analysis & Logistics Regression .

Sakshi Kishor Khanvilkar
Master of Science – Data Analytics National
College of Ireland Dublin, Ireland.
x22117776@student.ncirl.ie

Abstract*—Time series data are assessments of a parameter over time that are gathered at scheduled times. Stock prices, weather data, economic indicators, and many additional kinds of time series data are examples. Time series analysis is the procedure of deriving pertinent knowledge and making forecasts from data using methods based on statistics. Temperature time series data consisting of previous readings of temperatures over an extended period, frequently obtained at recurring times such as monthly, or yearly. We plan to use several time series models for evaluating previous measurements of temperature and make projections about temperature trends using this time series data. Such data can also be used to diagnose seasonal or annual patterns, as well as patterns that persist over time. Diabetes is a persistent metabolic process defined by high blood sugar levels (hyperglycemia) resulting from the release of insulin or action anomalies or both. Diabetes impacts millions of inhabitants throughout the globe and can lead to serious complications such as heart disease. Dealing with diabetes entails ongoing monitoring of blood glucose levels as well as alterations to medications, eating habits, and environment. In this context, this piece of content has two objectives. First, go over everything again. To accomplish this, this study describes prior significant articles that assessed time series data predictions using multiple time series models. Secondly, I would like to propose a diabetes prediction model that integrates a model of a few extrinsic diabetic factors as well as usual characteristics for improved diabetes classification. For time series projections, I employed and examined time series models such as exponential smoothing, ARIMA/SARIMA, and simple series models. For diabetic data, I adopted Logistic Regression to train the data, and the accuracy of Logistic Regression is 96%.

Keywords—Time series analysis, forecasting, modeling, logistic regression.

I. INTRODUCTION

Time series analysis, also referred to as a broad issue that aims to comprehend the highlighted structure of logical series of data to foresee and anticipate future observations, is of interest to a wide range of applications, which includes temperature data forecasting and modeling of dynamic nonlinear systems. Temperature time series analysis can aid in determining the presence of patterns, trends, and variations in seasons, which are vital for decision-making in a variety of sectors. For temperature information, methods for time series forecasting such as simple moving average, exponential smoothing, and ARIMA (autoregressive integrated moving average) can be employed. Algorithms like these can identify various trends and patterns in temperature data and provide outstanding predictions of future temperature readings. The features of the temperature data, such as its seasonality, trend, and randomness, influence the selection of an appropriate forecasting model. In this study, time series analysis and forecasting will be investigated in the context of temperature data. Time series models will be incorporated utilizing exponential smoothing, ARIMA, SARIMA, and simple series models. Secondly the diabetic analysis Diabetes Mellitus (DM) has the world's fourth highest condition death rate. According to the World Health Organization (WHO), 422 million people worldwide have diabetes, most of them reside in low- and middle-income countries, and diabetes causes 1.6 million deaths each year. Approximately 75% of diabetic patients have blood pressure (BP) results of around 130/80 mm Hg or demand medications that are antihypertensive. The proposed study's purpose is to improve the accuracy of detecting possible long-term damage in diabetic individuals. I decided on the logistics regression model for the data considering the diabetic data provided possessed the goal column as CLASS, which has categorical values along with other variables like gender, 'Gender', 'AGE', 'Urea', 'Cr', 'HbA1c', 'Chol', 'TG','HDL', 'LDL', 'VLDL', 'BMI'. This study is divided into four sections: section II by data description and descriptive statistics, section III methodology, section IV comparison, and section V conclusion. I employed several Python libraries, including a. pandas for reading cancer data, b. NumPy for numerical operations, c. Matplotlib and seaborn for data visualization, and d. Sci-kit Learn for utilizing a linear regression model.

II. Literature Review

[1] In this article, the author has proposed an illustration of a few outside diabetes variables along with normal factors for a better classification of diabetes. In the modern world, diabetes is a metabolic disorder. Diabetic patients are at an increasing rate every day. High blood glucose levels cause diabetic illnesses which in turn inevitably cause other illnesses like heart disease, kidney disease, etc. Increased hunger, thirst, fatigue, blurred vision, sores that do not heal, and unexplained weight loss are all signs of diabetes. Individuals with diabetes are at higher threat for disorders like retinal issues nerves impairment, etc. The dataset used consists of 250 variants containing 16 unique attributes. For this prediction, the author applied logistic regression, SVM and random tree and trained the data using 10-fold cross-validation, and the precision of logistic regression was 94.5%, support vector machine 96.5%, and random tree 97.5%.

[2] An Example from Nanjing. SARIMA (seasonal autoregressive integrated moving average) techniques are used in this paper to analyze the monthly average temperature in the city of Nanjing (China) between 1951 and 2017 using data from the years 1951 to 2014 as the training data set and 2015 to 2017 as the data set for the testing. Time series modelling and forecasting, which involves the analysis of past values to predict future values, has a significant impact on a variety of practical areas. The model selection and the forecasting accuracy are explained in detail. The proposed research approach achieves good forecasting accuracy.

[3] In this context, author has two objectives in this paper. The first objective is to provide a review of previous major papers that have studied time series forecasting in various application domains. The second objective is to propose a new approach to enhance the accuracy of ARIMA models by using an average of estimation error (AIGE) for forecasting time series. The experimental results suggest that this approach can enhance.

performance in the time series forecasting process In recent times, there has been a surge of interest in predicting time series databases in various fields of application. Forecasting has been one of the primary objectives of mining time series databases. Studies have demonstrated that time series forecasting can be used to inform appropriate decision-making in a variety of contexts. To date, a range of methods have been proposed to achieve the objective of predicting and analyzing literature in various directions.

[4] Diabetes is a huge issue in the healthcare industry, and it one of the biggest problems in Saudi Arabia. It is expected that the number of people with diabetes will keep going up, so predicting who is at risk is tough. This study compares two Machine Learning algorithms, Random Forest machine learning and Logistic regression, to predict diabetes. The author looked at 66,325 records from the MNGHA databases in Saudi Arabia from 2013 to 2015. The data set contains 64.47% male and 44.50% female prevalence of diabetes. The accuracy, recall, true positive rate, false positive rate, and f-measure for predicting diabetes for the Random Forest model were estimated at 0.883 (0.88), 0.88 (0.188), 0.82 (0.876) and 0.692 (0.703 (0.703)) respectively. The AUC (area under the curve) of the RF model was 0.94 (0.944) and the Logistic regression model was 0.708 (0.934). The RF algorithm demonstrated superior projection compared to the Logistic regression system for predicting diabetes across various solutions.

[5] The Indoor temperature time series analysis and forecasting is performed in two phases. The first phase involves the Fourier transformation and empirical mode decay of the time series to reveal temporal patterns within the data. The second phase involves the use of neural networks to predict future values. The results of this phase demonstrate the usefulness of the tools and encourage further development based on time frequency techniques for the design of the NN Fore-casting approach.

III. Data Description and Descriptive Statistic

Part A is time series analysis, and there are two data sets handed to me: monthly temperature data and yearly temperature data. The monthly dataset comprises one column and 1932 rows in a set whereas the yearly temp data set offers one column and 161 entries. Both data sets cover observations from the year 1844 to year 2004, and there are no null temperature records.

Part B is Diabetic prediction, there are 1000 instances and 14 attributes in the provided data file. Before continuing to the evaluation, it is crucial to check for duplication data and the absence of values or Nan values, to ensure that no data was duplicated. Following is a summary of the data record in tabular form. fig1.

Features	Data_Type
Gender	Categorical
Age	Numerical
Urea	Numerical
Cr	Numerical
HbA1c	Numerical
Chol	Numerical
TG	Numerical
HDL	Numerical
LDL	Numerical
VLDL	Numerical
BMI	Numerical
Daibetic	Categorical

Fig.01. Data Field Summary Part B

Fig.02. illustrates an overview of statistics for the variables in a 1000-observation dataset. Patient ID (ID), patient number (No_Pation), age (AGE), urea, creatinine (Cr), glycosylated hemoglobin (HbA1c), cholesterol (Chol), triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), very low-density lipoprotein (VLDL), and body mass index (BMI) constitute some of the variables.

	ID	No_Pation	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
count	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
mean	340.5	270551.41	53.53	5.12	68.94	8.28	4.86	2.35	1.2	2.61	1.85	29.58
std	240.4	3380757.82	8.8	2.94	59.98	2.53	1.3	1.4	0.66	1.12	3.66	4.96
min	1	123	20	0.5	6	0.9	0	0.3	0.2	0.3	0.1	19
25%	125.75	24063.75	51	3.7	48	6.5	4	1.5	0.9	1.8	0.7	26
50%	300.5	34395.5	55	4.6	60	8	4.8	2	1.1	2.5	0.9	30
75%	550.25	45384.25	59	5.7	73	10.2	5.6	2.9	1.3	3.3	1.5	33
max	800	75435657	79	38.9	800	16	10.3	13.8	9.9	9.9	35	47.75

Fig.02. Statistical Description Part B

The "count" row represents the total number of non-missing observations for all variables, which is 1000. Each variable's average value can be observed in the "mean" row. The "std" row displays the standard deviation of each variable, which indicates how dispersed the data is. The "min" row reveals the smallest observed value for each variable, while the "max" row represents the greatest observed value. The quartiles of each variable are shown in rows "25%", "50%", and "75%". For instance, the value in the "25%" row for the AGE variable is 51, indicating that 25% of the patients in the dataset are under the age of 51. The "50%" row displays the median, which is the figure that separates the lower 50% from the upper 50% of the data.

IV. METHODOLOGY PART A

Temperature time series models need to exist because temperature is a time-varying measure whose behavior over time can have a profound effect on a variety of categories such as agriculture, energy, public health, and others. By examining historical temperature data and acquiring time series models, we can recognize patterns and trends, forecast future temperature behavior, and make more educated judgments in areas such as crop planning, consumption of electricity forecasts, and public health response planning. I have two data sets of temperatures, one monthly and one yearly, encompassing the years 1844 to 2004. You may use time series models to assess and predict temperature behavior over time. Monthly temperature data can be utilized to identify fluctuations over time and seasonality, but yearly temperature data can provide a more comprehensive view of long-term patterns. I assessed both datasets using time series models such as ARIMA, SARIMA, and Prophet to seek out patterns, trends, and seasonality.

A. Trend Depiction And Data Interpretation

It is essential to ensure that time series data is properly organized and organized for analysis when employing it. I have two datasets case: monthly temperature data and yearly temperature data. To analyze the monthly temperature data, I start by reading the data into a Pandas Data Frame. The usual monthly temperature Additionally, there are two columns in the Data Frame: one for the date (with year and month) and one for the temperature. In a similar way the yearly temperature Data frame has two columns: the year column and the temperature column. For more analysis. I further create visuals for monthly temperature data and yearly temperature data. Monthly temperature frequency fig.03. describes the regularity with which temperature data is captured or measured, which in this case is monthly. On the x-axis, temperature values are presented against the year, providing insight into the temperature trend through time. The y-axis in this illustration represents temperature data, while the x-axis represents the year. I can observe the highest frequency in the temperature on 4.0 and 5.0.

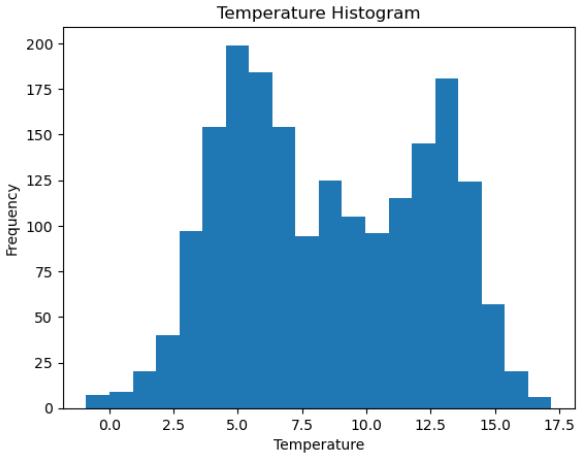


Fig.03. Temperature frequency monthly temperature data

Further I calculated the rolling mean and rolling standard deviation. Fig.04. Time-series statistical like rolling mean and rolling standard deviation are used to find trends and patterns in a collection. A rolling mean is a moving average with a defined window size that is used to identify long-term trends and smooth out data irregularities. On the other hand, the continuous standard deviation quantifies how far the data fluctuates from the rolling mean. The variability of the data within that window is determined, and it is calculated using the same size of window as the rolling mean. It is used to spot changes in the data's variability, which may signal changes in the underlying process responsible for producing the data.

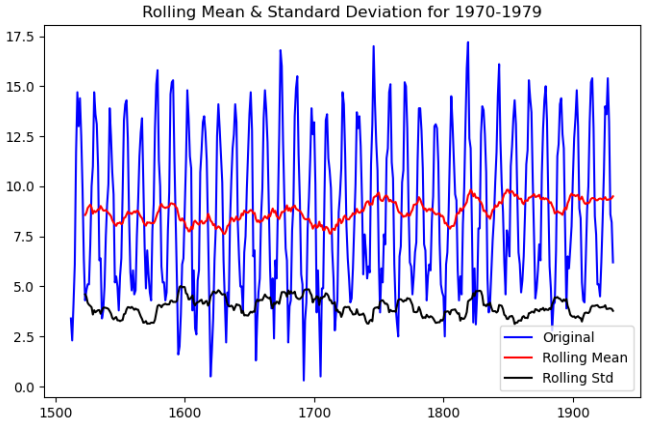


Fig.04. Rolling means standard deviation Monthly temperature data.

Next, I create the visuals for Yearly temperature Fig.05. illustration displays the temperature evolution from 1844 to 2004. The temperature is shown by the y-axis, while the year is portrayed by the x-axis. The visualization shows that

the temperature fluctuated over time and has been on an upward trend overall. The increasing tendency implies that the temperature has been increasing over time, which is essential to consider when researching climate urea data.

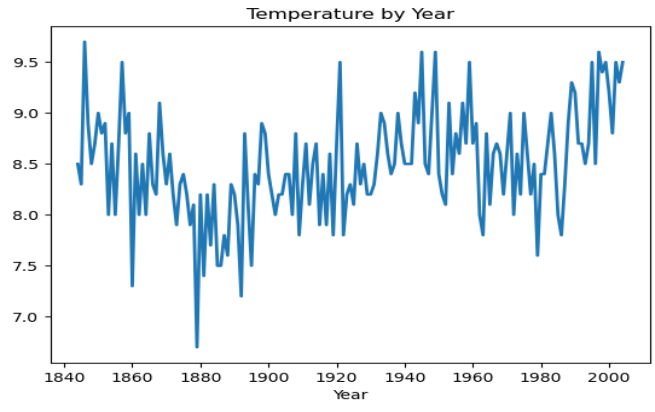


Fig.05.Yearly temperature trend.

The rolling mean and standard deviation for the decade 1970-1979 are being represented. The average of an exact amount of data points across a moving window is recognized as the rolling mean. The degree of variance in a group of data points over a moving window can be assessed by rolling standard deviation. The rolling mean and rolling standard deviation for yearly temperature data Fig.06.0

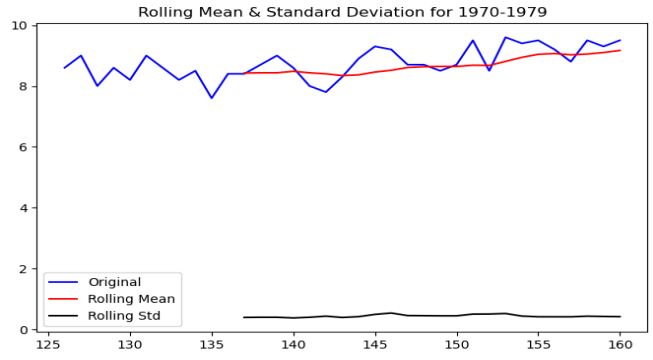


Fig.06. Rolling mean standard deviation Yearly temperature data.

B. Building Stationarity

When a time series is deemed to be stationary, it means that its statistical features remain unchanged throughout time. In other words, it demonstrates that there is no change in the series' mean, variance, or correlations structure over time. Since stationary time, series is predictable and exhibits a steady pattern throughout time, it is simpler to model and predict. The null hypothesis that a time series lacks stationary behavior due to the inclusion of a unit root is investigated using the Augmented Dickey-Fuller (ADF) test, which can be utilized to calculate the stationariness of monthly temperature data and annual data. We may disregard the null hypothesis and figure out that the data are stationary if the p-value of the ADF test is less than a threshold of significance (such as 0.05). For monthly temperature data, the p-value of the ADF test is 2.213013539022366e-05, which is substantially lower than 0.05. As a result, the null hypothesis may be declined, and it can be established that the monthly temperature data is static. This shows that the data's mean and variance have stayed stable throughout time and that there are no tendencies or variations in seasons that might have an impact on the series' statistical features. Given that the p value for the annual temperature data is more expansive than the widely accepted significance level of 0.05, this suggests that the data are non-stationary. Consequently, would be impossible to rule out the possibility that the data is not stationary. First order differencing, a typical technique for transforming non-stationary time series data into stationary time series data, is what I use when converting the non-stationary data in the yearly temperature data to stationary data.

C. Train Test Split

Establish training and test data frames for splitting the dataset into two sections for the creation of models and evaluation. The train data frame will be utilized for developing the time series models, and the test data frame will be used to validate them. The test data frame comprises temperature data from January 2004, while the train data frame covers temperature data up to 2003. We can compare the values that the model predicts for 2004 with the actual values we observed and evaluate the model's predictive power by using the actual temperature data for 2004 as the evaluation set and the actual temperature data for 2003 as the training set.

D. Implementation of Time Series model

Exponential Smoothing:

In time series analysis, the exponential smoothing tackle is used to eradicate noise or discrepancies from the data and generate forecasts. It depends on the premise that a time series' most recent observations are more pertinent than its earlier ones. Exponential smoothing yields older observations with exponentially decreasing weights, which get lower and smaller as the observations get older. Exponential smoothing's key benefit is that it is simple to comprehend and use, and it produces precise forecasts even for data with intricate patterns. It is utilized to the training set from 1844 to 2003 to apply an additive exponential smoothing model to the monthly temperature data, with a seasonal period of 12 (representing the 12 months in a year). The predicted values are then entered into the variable "predictions" to predict the average temperatures for the year 2004 using the model. Similarly, the same approach is done for the yearly temperature data by fitting an additive exponential smoothing model with a seasonal period of 1 year and using identical training and test sets. Fig.07. show the evaluation parameters of model for yearly and monthly temperature data frames.

ARIMA:

The statistical model known as ARIMA (Autoregressive Integrated Moving Average) is frequently utilized for prediction and time series analysis. It encompasses three vital components: moving average (MA), differencing (I), and autoregression (AR). The AR component mimics the relationship amongst the current value and the series' previous values, the MA component replicates the interaction between the current value and the series' prior errors, and the I component simulates the differencing process required to make the series stationary. The quantities of an ARIMA model's three parameters, p (order of the AR component), d (order of differencing), and q (order of the MA component), must be specified before the model can be fit. Model recognition or order selection is the process of determining the ideal values for these parameters. I acquired the order for ARIMA as p=4, d=1, q=3 for monthly temperature data and p=0, d=0, q=1 for yearly temperature data, after which I carried out the train dataset and evaluated on test dataset

ARIMA AIC:

The Akaike Information Criterion, or AIC, is a quantitative model selection metric. By evaluating the goodness of fit and the model's complexity, it indicates how good a model contrasts with other models. AIC is commonly employed in ARIMA modeling for determining the best model that strikes a compromise between complexity and goodness of fit. The log-likelihood of the model and the number of features included in the model are used to compute the AIC value. A better model will have a lower AIC score since it will better balance model complexity and quality of fit. The primary distinction between ARIMA and ARIMA AIC is that despite ARIMA is a model, ARIMA AIC is an algorithm for selecting models. A time series model is then fitted to the data using ARIMA, and ARIMA AIC is used to compare different periods ARIMA models and choose the best one based on a model quality criterion. ARIMA AIC aids in selecting a model that strikes a balance between model complexity and accuracy, eliminating overfitting.

SARIMA

A modified version of the ARIMA model called SARIMA (Seasonal Autoregressive Integrated Moving Average) serves to recognize seasonal patterns in time series data. Comparable to ARIMA, it entails differencing the data to make it stationary before fitting an amalgamation of autoregressive (AR), integrated (I), and moving average (MA) components to the resulting series.In SARIMA, seasonal search terms are also included to capture developments that repeat over set time intervals (such as monthly, quarterly, or yearly). The same ARIMA notation is used to characterize seasonal terms as non-seasonal terms, but with an extra set of parameters that compensate for the seasonal lag. In SARIMA, the pqd notation reflects the arrangement of the non-seasonal and seasonal ARIMA terms (p, d, and q) as well as the length of the seasonal cycle.
Specifically:
p: the model's total number of autoregressive terms.
d: the amount of difference used to make the series stationary, denoted by the letter.
q: How many moving average terms are there in the model?

Simple Time Series model – Seasonal Naïve:

A simple forecasting model used in time series analysis is the seasonal naive model. The estimates for forthcoming times are based on the actual values of the same periods in beforehand seasons, which is a specific case of the naïve technique. The seasonal naïve model makes a claim that the time series' seasonal pattern will remain constant all throughout time and that there are no patterns or different factors will influence it. While it might work for some time series data, it is typically not a highly reliable forecasting technique, especially for anticipates over a longer period.

V. METHODOLOGY PART B

The objective is to figure out an individual's probability to suffer from diabetes based on an assortment of diagnostic tests, such as blood glucose levels, blood pressure, body mass index (BMI), etc. Constructing a model that can accurately determine whether a patient possesses diabetes or not is the main goal of this binary classification challenge. By employing machine learning techniques, predictions may be established with greater accuracy and patients at risk for contracting diabetes can be found. The target column in the data, which encompasses 1000 instances and 14 attributes, is called "CLASS." The data frame comprises one categorical feature, "Gender," and the numerical aspects "Age," "UREA," "Cr," "HbA1c," "Chol," "TG," "HDL," "LD," "VLD," and "BMI."

A. Trend Depiction And Data Interpretation

The univariant analysis of the "gender" columns is shown in Fig. 8. The categories beneath each feature are represented by the count plot's x-axis, and the number of instances of each class is expressed by the y-axis. Looking at the distribution of each category inside the "Gender" characteristic, the majority of individuals fall into the male gender. The association between the gender and the target variable class, which corresponds to "yes" or "no," is depicted in Fig. 9. Contrary to the graph, men have a higher risk of acquiring diabetes.

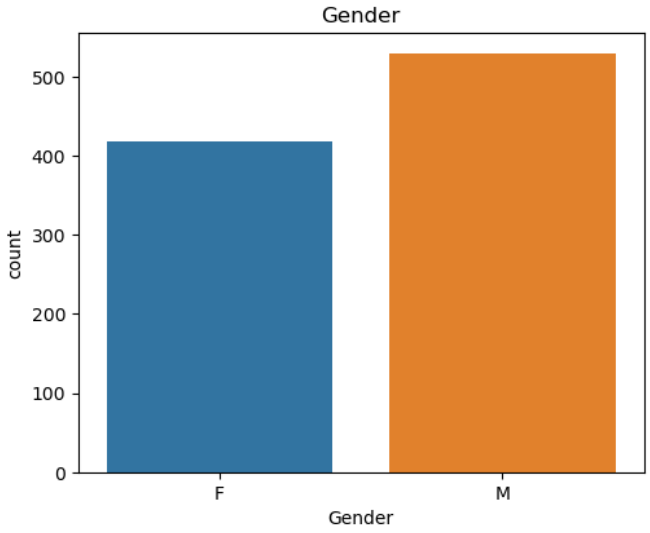


Fig .8. Univariant analysis of feature Gender

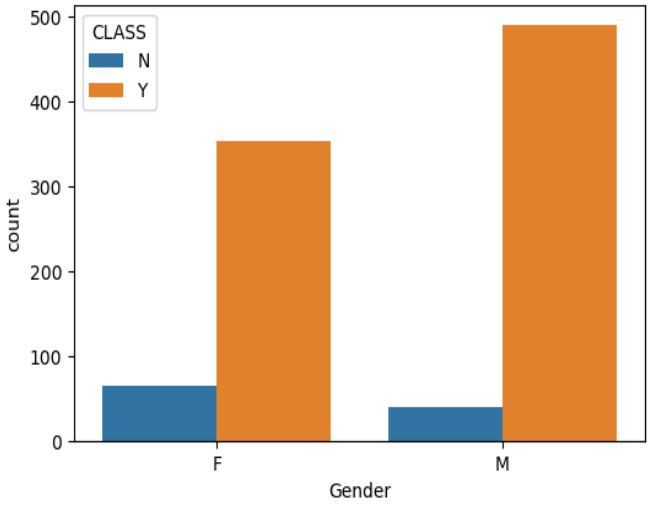
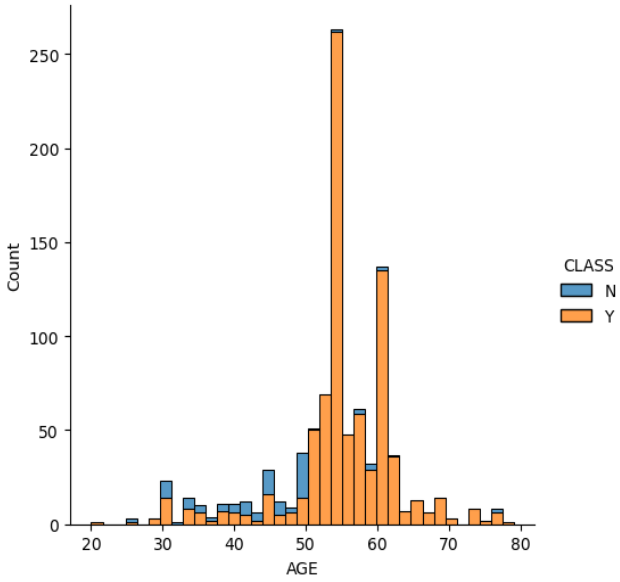
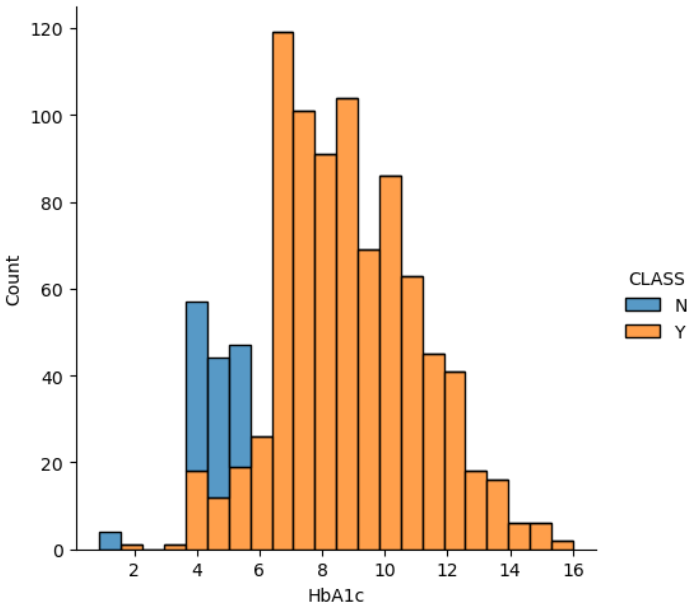
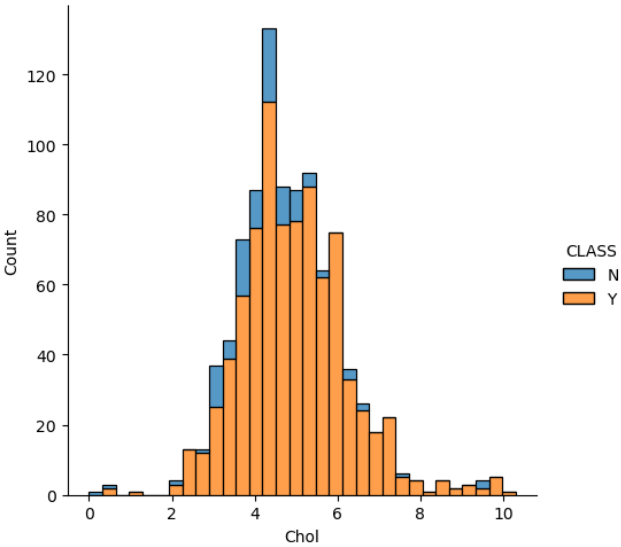


Fig.9. Relationship between Gender and Target Feature

The correspondence between a numerical variable and the target variable is shown in Fig. 10. When reviewing the relationship between age and the target, people in the 51–65 age group have an increased likelihood of developing diabetes. In furthermore, individuals with a blood sugar level exceeding 6.5 and 10 are at risk of becoming diabetic. The cholesterol and target column graph shows that people with a 4.5 to 5.0 chance of acquiring diabetes. Therefore, people with BMIs between 25 and 38 are at a higher probability of developing diabetes.







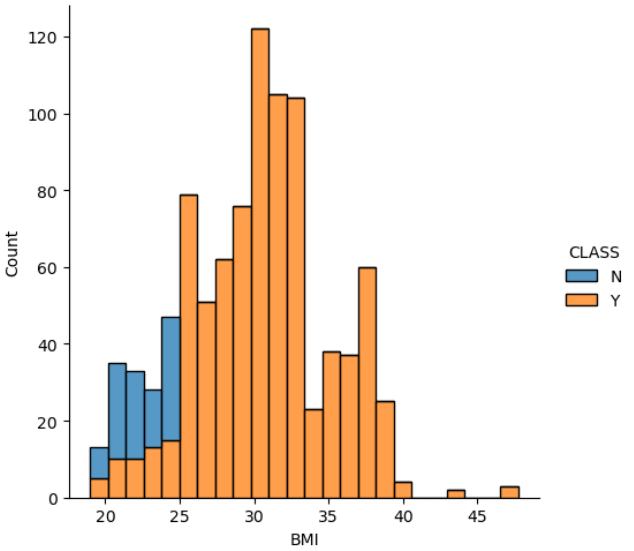


Fig 10. Relationship of Numerical Feature and target variable

B. Outlier Detection And Treating

Data points designated as outliers are those that are substantially distinct from many of the data points in a dataset. They have an enormous effect on the findings of statistical analyses and might have abnormally high or low values. Various factors, such as measurement errors, data entry problems, or inherent volatility in the data, can lead to outliers. To guarantee the accuracy and dependability of statistical analyses and models, it is crucial to recognize and effectively handle outliers. For detecting outliers, I have used IQR, Interquartile Range, or IQR, is a measure of variability based on categorizing a set of data into quartiles. It is frequently used to find outliers in data. The IQR, which shows the distribution of the middle 50% of the data, is determined as the difference between a dataset's third quartile (Q3) and first quartile (Q1) values. To find outliers using IQR, the data is deemed to have outliers if they fall outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR. An outlier is a data point that lies outside of the normal distribution, either below or above the upper or lower limits. Since IQR is unaffected by extreme values or data skewness, it provides a reliable tool for identifying outliers. After implementing IQR I found the outliers in 'Age','Urea','Cr','TG'. Have been treating those outliers using IQR.

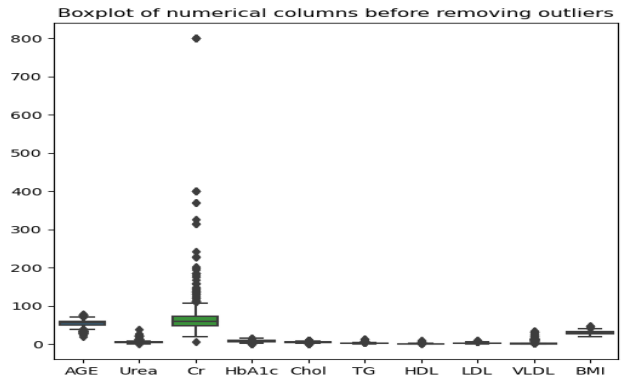


Fig 11. Outlier present

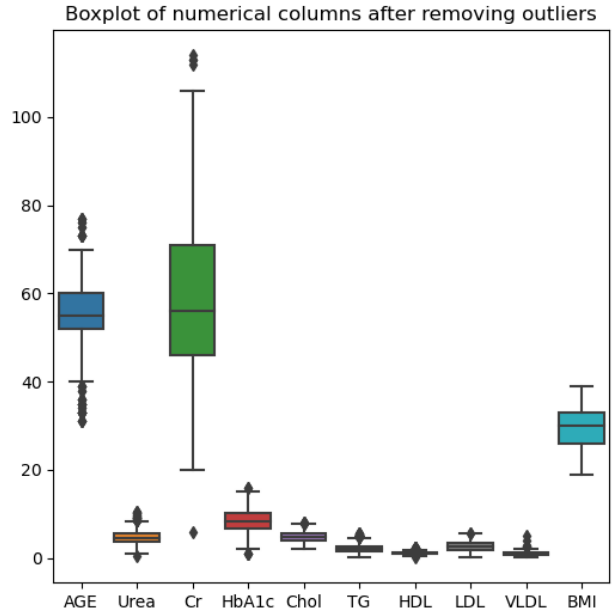


Fig. 12 After Treating the outliers.

C. Converting Categorical To Numeric

Use label encoding when converting the feature to a numeric notation before employing categorical variables in the model implementation. Categorical variables('Gender') undergo conversion into numerical values through the process of label encoding. Through this technique, a categorical variable's categories are each given a distinct integer value that may later be entered into machine learning models. Numerical values ultimately have no intrinsic significance because they are assigned based on the hierarchy of the categories. When there are only two categories or when the categories have a natural order (such as low, medium, and high), label encoding is frequently utilized.

D. Feature Selection

The process of determining a subset of pertinent features (variables) that are most effective in predicting the target variable is commonly referred to as feature selection in statistics and machine learning. Using p-values as an instrument of feature selection is one approach. The statistical significance of the association between the

dependent and independent variables can be examined quantitatively using the P-value. With this strategy, we estimate the p-value for each feature in relation to the target variable and only choose the traits that fall below a predetermined threshold (often 0.05 or 0.01). As a result, we only pick traits that have a statistically significant impact on predicting the target variable.
A Python module called Stats Models delivers an assortment of statistical models for data analysis. Ordinary least squares (OLS) regression analysis, which may be used for calculating the p-values for each feature in the model, can be performed by a module in it named statsmodels.formula.api. After feature selection I got five feature whos p-value is less than 0.05, those are 'LDL', 'TG', 'Chol', 'HbA1c', 'BMI'.

E. Feature Scaling.

An approach used in machine learning to standardize the range of independent variables or features of data is called feature scaling. Min-Max scaling is one of the feature scaling techniques. When using min-max scaling, the feature's values are rescaled to fall between 0 and 1. To achieve this, subtract the feature's minimum value and divide the result by the difference between its minimum and maximum values.
Certain machine learning algorithms perform better when all features are on the same scale, which is achieved using min-max scaling.

F. Implementation of losgitic model

In binary classification problems, where the response variable is a categorical variable that can have one of two possible values (for example, 0/1, yes/no, true/false), logistic regression is a statistical technique used to model and analyze the problems. Based on a set of chosen features, logistic regression can be used in the context of diabetic prediction to determine whether a person is likely to get diabetes or not. Given the input features, the logistic regression model calculates the likelihood that a given category would make up a response variable. The input features are summed up linearly to create an average likelihood score between 0 and 1, which is subsequently converted into a probability score using a logistic function. By deciding on a decision threshold (such as 0.5) above which the model predicts favorable results (such as diabetes) and below which it predicts undesirable results (such as no diabetes), the likelihood score is able to be transformed into a binary prediction. After evaluating the model on test data following are the scores I got

Parameters	Score
Recall	99%
Precision	95%
F1_Score	97%
Accuracy_Score	95%
ROC_AUC	94%

The correctness of a classification model can be evaluated using an assessment of performance statistics dubbed the confusion matrix. The number of instances of true positives, true negatives, false positives, and erroneous negatives for a specific categorization issue is displayed in a table. The confusion matrix fig. 13, defined is a 2x2 table with the four possible results in a binary classification problem, is true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The confusion matrix can be observed by employing a heatmap function from the Seaborn library. It uses different colors to represent the values in the matrix, making it simple to see how the model is functioning.

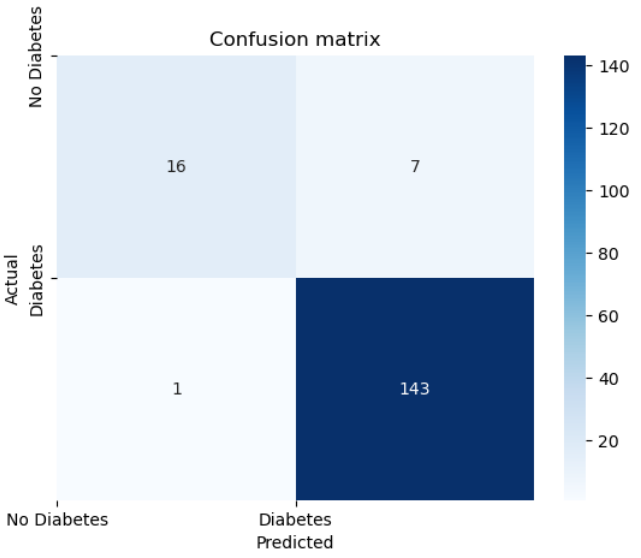


Fig 13. Confusion Matrix

IV . RESULT COMPARISON

Part A- Yearly and Monthly Temperature data

Exponential Smoothing

Parameters	Monthly_Temp_Data	Yearly_Temp_Data
MAE(Mean Absolute Error)	0.6954	0.7863
MSE(Mean Square Error)	0.6183	0.038
RMSE(root mean square error)	0.7863	0.1949

ARIMA

Parameters	Monthly_Temp_Data	Yearly_Temp_Data
MAE(Mean Absolute Error)	0.6492	0.3301
MSE(Mean Square Error)	0.7047	0.1089
RMSE(root mean square error)	0.8394	0.3301

ARIMA AIC

Parameters	Monthly_Temp_Data	Yearly_Temp_Data
MAE(Mean Absolute Error)	0.7047	0.8394
MSE(Mean Square Error)	0.6492	0.3301
RMSE(root mean square error)	0.8394	0.3301

SARIMA

Parameters	Monthly_Temp_Data	Yearly_Temp_Data
MAE(Mean Absolute Error)	0.7152	0.3083
MSE(Mean Square Error)	0.6531	0.095
RMSE(root mean square error)	0.8081	0.3083

Simple Series model -Seasonal Naïve

Parameters	Monthly_Temp_Data	Yearly_Temp_Data
MAE(Mean Absolute Error)	4.5	0.3999
MSE(Mean Square Error)	32.45	0.1599
RMSE(root mean square error)	5.696	0.3999

Part B

I determine the precision of our logistic regression model on the portion of the dataset where the CLASS is 'P'. To gauge the exactness of our predictions, we leverage the scikit-learn accuracy_score() tool. Since I assume that all cases will be positive and are aware that they are, we predict the accuracy to be 1.0 in this instance.

Accuracy on 'P' cases: 1.0

VI. CONCLUSION

Part A

It turned out that the ARIMA AIC model performed a good job of protecting the time series data' future values. This was demonstrated by the model's low AIC value, which demonstrates how the model can fit the data effectively while using a few parameters. The model's good precision score and its visual representation of the anticipated and actual values show that it was also able to capture the seasonal and trend patterns in the data. Consequently, I can say that the ARIMA AIC model is an effective way to predict future values of this time series dataset.

Part B

I can make the view that the model performed well with a 95% accuracy based on the evaluation of the logistic regression model for diabetic prediction. This demonstrates that the model can accurately forecast the development of diabetes. To enhance the model's performance, further analysis and improvement may be required.

VII. REFERENCES

[1] M. A. M. B. M. R. Z. M. H. S. Md. Mehedi Hassan, *Early Predictive Analytics in Healthcare for Diabetes Prediction Using Machine Learning*, p. 5, 2021.

[2] A. N. D. L. Peng Chen1, "Time Series Forecasting of Temperatures using SARIMA: An Example from Nanjing," *IOP Conference Series: Materials Science and Engineering*, p. 8, 2018.

[3] M. R. K. Soheila Mehrmolaei, "Time series forecasting using improved ARIMA," *978-1-5090-2169-7/16/$3l.00 ©20 16 IEEE*, p. 6, 2016.

[4] T. D. a. R. Alshammari, "Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes," *Journal of Advances in Information Technology Vol. 11, No. 2, May 2020*, vol. 11, p. 6, 2020.

[5] A. M. L.́. Manuel Romano Barbosa, "Temperature Time Series: Pattern Analysis and Forecasting," *2017 4th Experiment@ International Conference (exp.at'17)*, p. 6, 2017.