Time Series Analysis of Average Temperatures in Armagh: Modeling and Forecasting for Climate Trends

*Part A: Time Series Analysis

Saheli Dutta
x21246513student@ncirl.ie
National College of Ireland
Masters in Data Analytics

***Abstract*—This research study presents a time series analysis of average temperatures in Armagh, focusing on modeling and forecasting climate trends. Two datasets are used: monthly average temperatures and yearly average temperatures from 1844 to 2004. The monthly data shows stationarity, while the yearly data displays a trend. Three categories of models are considered: exponential smoothing, ARIMA/SARIMA, and simple time series models. The models are evaluated using diagnostic tests, and the data up to 2003 are used to forecast temperatures for 2004. The forecasts are compared with the actual 2004 data. Optimum models are selected based on performance. Different modeling approaches are required for monthly and yearly data. The selected models accurately captured patterns and provided reliable forecasts. This study contributes to understanding climate trends in Armagh, aiding informed decision-making and proactive measures to address climate risks.**

***Index Terms*—Climate trends, Time series analysis, Descriptive Statistics, Naive Method, Exponential smoothing, ARIMA models, SARIMA models, Stationarity, Forecasting, and Model evaluation.**

I. Introduction

This report presents a comprehensive analysis of average temperatures in Armagh using time series analysis techniques. The dataset consists of monthly average temperatures from January 1844 to December 2004 and yearly average temperatures from 1844 to 2004. These datasets provide insights into long-term trends and seasonal variations in Armagh's average temperatures.

Distinct characteristics are observed in the two datasets. The monthly data exhibits stationarity indicating consistent statistical properties over time and prominent seasonality. Conversely, the yearly data displays a noticeable trend.

Various time series models are employed to capture patterns and forecast future temperature trends. For the monthly data, methods such as seasonal naive, Holt-Winters' seasonal exponential smoothing, and SARIMA models are utilized. These models consider seasonality and provide short-term temperature variations. For the yearly data with a trend, models such as naive, exponential smoothing, and ARIMA models are applied. These models aim to capture underlying trends and forecast long-term temperature trends.

Diagnostic tests and visualizations are conducted to assess the models' performance and adequacy. These include the Augmented Dickey-Fuller (ADF) test, autocorrelation function (ACF), and partial autocorrelation function (PACF) plots, as well as the Ljung-Box test. Additionally, metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), Theil's U, Akaike information criterion (AIC), and Bayesian information criterion (BIC) are used to evaluate the models' performance.

The analysis is performed using the R programming language, leveraging its time series analysis libraries and functions. The objective of this study is to provide accurate and reliable forecasts for average temperatures in Armagh based on historical data.

II. Related Work

In the study "A Complete Tutorial on Time Series Modeling in R" by Tavish Srivastava, the author discusses the importance of time in business and the methods of prediction and forecasting. The study focuses on time series modeling as a powerful tool for analyzing time-based data and deriving hidden insights for informed decision-making.

Time series modeling is particularly useful when dealing with serially correlated data, commonly encountered in various business domains. The study serves as a guide, providing various levels of time series modeling and introducing the associated techniques. It aims to bridge the knowledge gap and equip analysts with the necessary skills to effectively utilize time series modeling in their analyses [1].

In the context of the present research on average temperatures in Armagh, time series modeling techniques discussed in the tutorial, such as exponential smoothing, ARIMA, and seasonal methods are applied to effectively model and forecast the temperature patterns. These techniques take into account the temporal nature of the data and can capture seasonality, trends, and other relevant components.

III. Methodology

A. *Data Sets and Characteristics*

The dataset for monthly data consists of 1932 observations of one variable. The data is stored in a data frame format. Each observation represents the average temperature for a specific month. The values range is in the sequential order of the months. For example, the first value corresponds to January, the second value to February, and so on (**Fig 1**).

The yearly dataset consists of 161 observations of one variable.

	x
1	8.5
2	8.3
3	9.7
4	8.9
5	8.5
6	8.7

Fig. 1: Monthly Time Series Dataset

Similar to the monthly data, the data is stored in a data frame format. Each observation represents the average temperature for a specific year. For example, the first value corresponds to the temperature for the first year, the second value to the temperature for the second year, and so on (**Fig 2**).

	x
1	4.5
2	2.4
3	4.8
4	9.1
5	10.9
6	12.9

Fig. 2: Yearly Time Series Dataset

Both datasets are valuable resources for analyzing and modeling average temperatures in Armagh. The monthly data provides a higher level of granularity, allowing for the examination of temperature variations within individual months. On the other hand, the yearly data provides a broader perspective, enabling the identification of long-term trends in average temperatures.

B. Data Exploration and Descriptive Statistics

This section aims to perform a descriptive analysis to gain insights into the variables present in the dataset.

1) The time series analysis is performed on the monthly temperature data using the time series object. The resulting time series object spans from January 1844 to December 2004, with a frequency of 12 (representing monthly observations). The class ”ts”, confirms its representation as a time series (**Fig 3**).
   Moving on to the summary statistics, the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values of the monthly temperature data is calculated (**Fig 4**).
   The summary reveals the following characteristics:
   - The median value of 8.2 represents the middle value in the sorted distribution of the monthly temperature data in Armagh. It indicates that approximately 50% of the observed temperatures fall below 8.2, while the other 50% fall above this value.
   - The mean temperature of 8.501 represents the average temperature calculated over the observed period.
2) The ‘start(ts_data)‘ function indicates that the yearly time series data starts in 1844 and ends in 2004, with a frequency of 1, indicating that the observations are collected on an annual basis (**Fig 5**).

```
+-------+--------+--------+--------+--------+--------+--------+
|       |  Jan   |  Feb   |  Mar   |  Apr   |  May   |  Jun   |
+-------+--------+--------+--------+--------+--------+--------+
| 1844  |  4.5   |  2.4   |  4.8   |  9.1   | 10.9   | 12.9   |
|       |        |        |        |        |        |        |
| Class |   ts   |   ts   |   ts   |   ts   |   ts   |   ts   |
|       |        |        |        |        |        |        |
| Freq  |   12   |   12   |   12   |   12   |   12   |   12   |
+-------+--------+--------+--------+--------+--------+--------+

+-------+--------+
| Start | 1844,1 |
| End   | 2004,12|
+-------+--------+
```

Fig. 3: Monthly Dataset Description

```
+---------+--------+
|         | Values |
+---------+--------+
|   Min.  | -0.900 |
| 1st Qu. |  5.300 |
|  Median |  8.200 |
|   Mean  |  8.501 |
| 3rd Qu. | 12.100 |
|   Max.  | 17.200 |
+---------+--------+
```

Fig. 4: Monthly Dataset Summary

	Start	End	Frequency	Class
	1844	2004	1	Time Series

Yearly Summary:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	6.700	8.200	8.500	8.489	8.800	9.700

Fig. 5: Yearly Dataset Summary

The summary statistics for the yearly temperature data in Armagh reveal the following characteristics:

- The minimum temperature recorded is 6.7, the median is 8.5, and the maximum temperature is 9.7.
- The 1st quartile (25th percentile) represents the temperature value below which 25% of the yearly temperature data in Armagh falls, while the 3rd quartile (75th percentile) represents the temperature value below which 75% of the yearly temperature data falls.

C. *Data Visualization*

1) In the analysis of monthly average temperature data in Armagh, several visualizations are created to explore the characteristics of the time series.
   - First, a time plot of the monthly average temperature is generated, providing a visual representation of the data over time. The plot allows for a visual assessment of any trends, patterns, or outliers in the data (**Fig 6**).

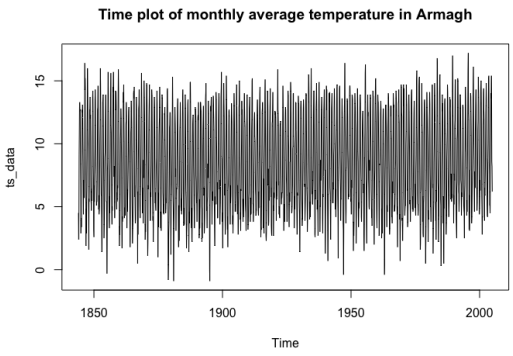


Fig. 6: Time plot of monthly average temperature in Armagh

The time plot of the monthly average temperature reveals a consistent range of values between 5 and 15 on the y-axis, suggesting no discernible trends or patterns in the data.

- Next, seasonal patterns are examined using two different visualizations [2]. The seasonal plot displayed the seasonal effects in the data, highlighting recurring patterns or fluctuations across the months (**Fig 7**).

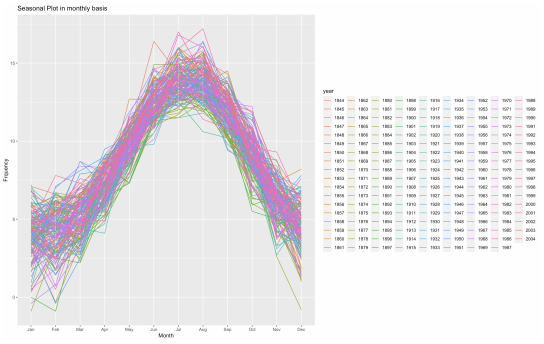


Fig. 7: Seasonal Plot

Additionally, the subseries plot provides insights into the distribution of observations within each season (**Fig 8**).
Both the seasonal plot and the subseries plot reveal consistent temperature rises from May to September, indicating a recurring seasonal pattern. Furthermore, a peak in temperatures is observed specifically during the months of July and August.

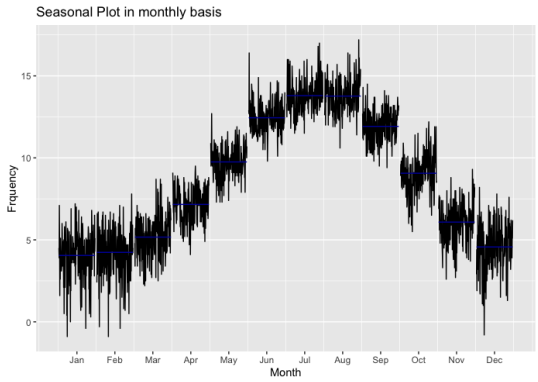


Fig. 8: Seasonal Subseries Plot

- To further investigate both the seasonal effects and potential outliers within the data, a boxplot is constructed to illustrate the variation in temperature across the 12 months.

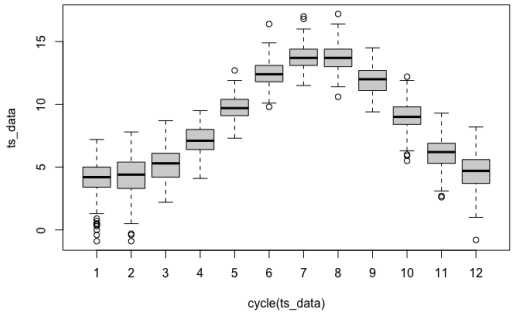


Fig. 9: Box Plot of Monthly Data

The boxplot reveals the same outcome as the season plot (**Fig 9**).
The dataset ts_data does not have any values that fall outside of the ”whiskers” of the boxplot. This means that there are no outliers detected based on the criteria used by the boxplot.stats() function [3].

- To decompose the time series into its underlying components, a decomposition plot is generated [4]. The plot displays the trend, seasonal, and remainder elements, allowing for a better understanding of the individual components’ contributions to the overall series (**Fig 10**).
  The decomposition plot reveals an additive seasonality pattern and a slight presence of trend in the time series. To further assess the trend’s stationarity, an Augmented Dickey-Fuller (ADF) test will be conducted.

- Lastly, the correlogram, consisting of the autocorrelation function (ACF) and partial autocorrelation

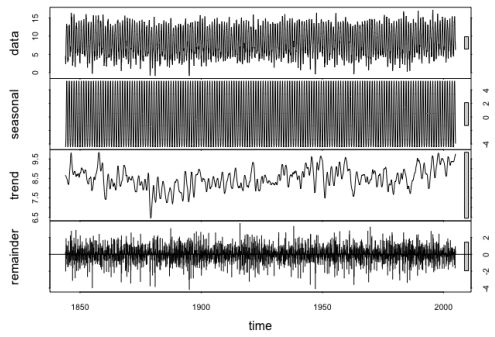


Fig. 10: Decomposed Data

function (PACF) plots [5], is utilized to assess the presence of autocorrelation in the time series (**Fig 11**).

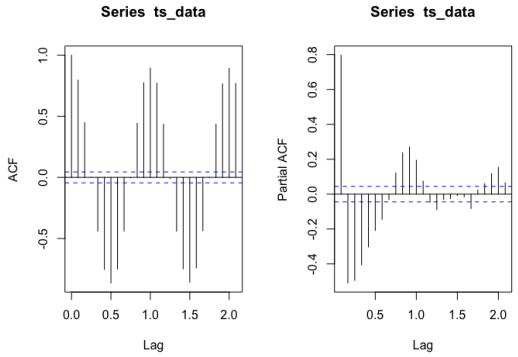


Fig. 11: Correlogram of Monthly Data

The autocorrelation plot shows strong positive correlations at lag 0 and lags 1 and 2, indicating a potential seasonal pattern in the data. The partial autocorrelation plot displays a significant positive partial autocorrelation at lag 0, implying a direct relationship between consecutive observations. The negative partial autocorrelation suggests the impact of the immediate previous observation.

2) In the analysis of yearly average temperature data in Armagh, various visualizations are generated to investigate the properties of the time series.
   - A time plot is created to visualize the yearly average temperature data over time (**Fig 12**). Clearly, the data is not stationary and exhibit some pattern in the plot.
     To make the data stationary, differencing is applied with the help of ndiffs() function. The differenced yearly average temperature time series is shown in a time plot, which helps in identifying any remaining patterns or trends (**Fig 13**).
     By differencing the data, the mean is adjusted to zero, helping to remove any persistent trends and

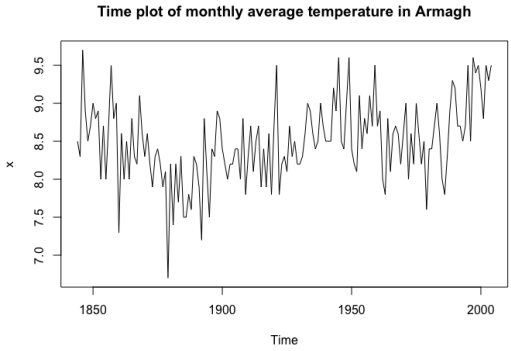


Fig. 12: Time plot of yearly average temperature in Armagh

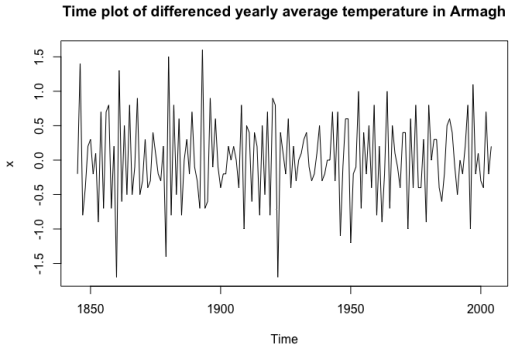


Fig. 13: Time plot of differenced yearly average temperature in Armagh

making the data trend stationary.
   - The correlogram, consisting of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, is utilized to examine the presence of autocorrelation in the differenced time series (**Fig 14**).

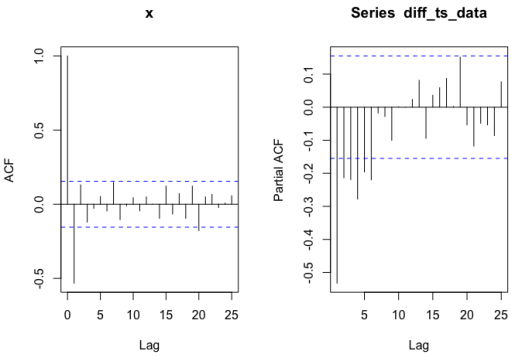


Fig. 14: Correlogram of Yearly Data

In the ACF, no prominent lags are evident, as most correlation values fall within the range of -0.2 to 0.2. Similarly, in the PACF, a negative correlation

is observed at lag 1, indicating a direct influence of the previous observation. However, there are no significant or prominent lags beyond lag 1 in the PACF, indicating no distinct pattern in the partial autocorrelations.

Overall, the correlogram does not show any significant or consistent autocorrelation patterns, suggesting the absence of strong dependencies or trends in the differenced time series data.

D. Time Series Properties

- For the yearly average temperature data in Armagh, the Augmented Dickey-Fuller (ADF) test is conducted to examine the stationarity of the data [6]. The test yielded a Dickey-Fuller statistic of -2.861 with a p-value of 0.2172. The p-value suggests that there is insufficient evidence to reject the null hypothesis of non-stationarity.
Similarly, for the differenced yearly average temperature data, the ADF test is performed. The test resulted in a very small p-value of 0.01. The p-value indicates strong evidence to reject the null hypothesis of non-stationarity in favor of the alternative hypothesis of stationarity (**Fig 15**). The mean seasonal cycle is 1, indicating a

Test	Dickey-Fuller	Lag Order	p-value	Stationary?
Original Data	-2.861	5	0.2172	No
Differenced	-9.8954	5	0.01	Yes

	Value
Mean seasonal cycle	1
Standard deviation of seasonal cycle	0

Fig. 15: Stationary Yearly Data

consistent pattern observed across the years. The lack of significant variability in the seasonal component suggests a consistent annual temperature pattern in the region.

For the monthly average temperature data in Armagh, the ADF test is conducted. The test produced a Dickey-Fuller statistic of -7.6152 with a p-value of 0.01. The p-value suggests significant evidence to reject the null hypothesis of non-stationarity in favor of the alternative hypothesis of stationarity (**Fig 16**).

Test	Dickey-Fuller	Lag Order	p-value	Stationary?
Original Data	-7.6152	12	0.01	Yes

	Mean	Standard Deviation
Monthly Data	6.5	3.452946

Fig. 16: Stationary Monthly Data

The mean seasonal cycle for the monthly data is computed as 6.5, indicating an average fluctuation around this value over the course of a year. The standard deviation of the seasonal cycle is calculated as 3.452946, suggesting the degree of variability in the monthly average temperature within each season.

IV. Evaluation Metrics

The evaluation metrics used to assess the accuracy of the generated forecasts against the actual data for the year 2004 include ME (Mean Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error), MASE (Mean Absolute Scaled Error), ACF1 (Autocorrelation of Residuals at Lag 1), and Theil's U statistic [7].

V. Model Performance - Monthly Data

1) The Seasonal Naive method is applied to forecast the monthly average temperature in Armagh. The dataset is split into a training set (until December 2003) and a test set (starting from January 2004). The Residual sd value of 1.705 represents the standard deviation of the forecast errors. A higher value indicates larger deviations from the actual values, suggesting a relatively higher level of uncertainty in the forecasts (**Fig 17**) [8].

```
Forecasts:
         Point Forecast     Lo 80     Hi 80      Lo 95     Hi 95
Jan 2004            4.3  2.114926  6.485074  0.9582173  7.641783
Feb 2004            4.2  2.014926  6.385074  0.8582173  7.541783
Mar 2004            6.4  4.214926  8.585074  3.0582173  9.741783
Apr 2004            8.8  6.614926 10.985074  5.4582173 12.141783
May 2004           10.0  7.814926 12.185074  6.6582173 13.341783
Jun 2004           13.2 11.014926 15.385074  9.8582173 16.541783
Jul 2004           15.2 13.014926 17.385074 11.8582173 18.541783
Aug 2004           15.4 13.214926 17.585074 12.0582173 18.741783
Sep 2004           12.9 10.714926 15.085074  9.5582173 16.241783
Oct 2004            8.5  6.314926 10.685074  5.1582173 11.841783
Nov 2004            7.6  5.414926  9.785074  4.2582173 10.941783
Dec 2004            5.1  2.914926  7.285074  1.7582173  8.441783
```

Fig. 17: Point Forcast Value for Seasonal Naive

The Ljung-Box test is conducted to check the residuals of the Seasonal Naive method, and the obtained p-value is less than 0.05, indicating significant evidence of autocorrelation in the residuals (**Fig 18**) [9].

```
        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 674.17, df = 24, p-value < 2.2e-16

Model df: 0.   Total lags used: 24
```

Fig. 18: Ljung Box Test for Seasonal Naive

The Seasonal Naive method performs reasonably well in forecasting the monthly average temperature in Armagh. The MPE (Mean Percentage Error) value of 3.507% indicates that, on average, the forecasts have a percentage error of approximately 3.507%. The MASE (Mean Absolute Scaled Error) value of 0.437 suggests that the Seasonal Naive method captures the variability of the data reasonably well when compared to a naive benchmark model.
The ACF1 (Autocorrelation of Residuals at Lag 1) value of -0.031 indicates minimal correlation in the residuals at Lag 1.

Theil's U statistic of 0.301 represents moderate forecasting performance. A value closer to 0 indicates better forecasting performance. (**Fig 19**).

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.005	1.705	1.334	-Inf	Inf	1	0.198	NA
Test set	0.200	0.729	0.583	3.507	7.01	0.437	-0.031	0.301

Fig. 19: Evaluation Metrics for Seasonal Naive

A plot of the forecasts and the actual data for the test period visually depicts the performance of the Seasonal Naive method (**Fig 20**).

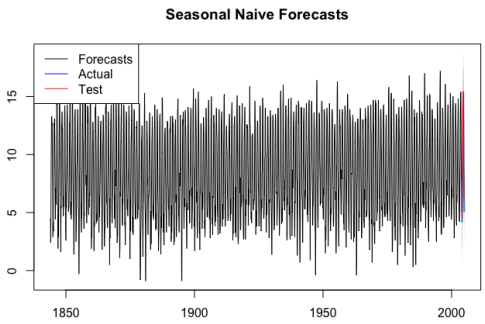


Fig. 20: Forecast Data

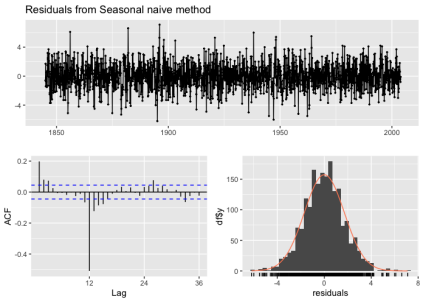


Fig. 21: Residuals of Seasonal Naive

Furthermore, the residuals of the Seasonal Naive method are checked for white noise, normal distribution, and autocorrelation (at Lag 12 autocorrelation present) (**Fig 21**).
However, there is room for improvement in terms of reducing the autocorrelation in the residuals and enhancing the accuracy of the forecasts.

2) Holt's method with additive seasonality is applied to forecast the monthly average temperature in Armagh. The Residual Standard Error (RSE) of 1.250518 indicates the average magnitude of the forecast errors.
The forecasts generated by the HoltWinters model for the test data along with 80% and 95% confidence intervals. The forecasts indicate the expected values for each month in 2004 (**Fig 22**).

```
Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Jan 2004       4.755004  3.152846  6.357161  2.304715  7.205292
Feb 2004       4.935544  3.330648  6.540441  2.481068  7.390021
Mar 2004       6.253573  4.645899  7.861246  3.794848  8.712297
Apr 2004       7.862041  6.251552  9.472531  5.399010 10.325073
May 2004      10.526687  8.913342 12.140032  8.059288 12.994085
Jun 2004      12.918052 11.301812 14.534292 10.446226 15.389877
Jul 2004      14.778224 13.159050 16.397398 12.301911 17.254537
Aug 2004      14.797802 13.175654 16.419949 12.316941 17.278662
Sep 2004      12.784633 11.159472 14.409794 10.299163 15.270102
Oct 2004      10.062583  8.434368 11.690797  7.572444 12.552722
Nov 2004       7.168671  5.537364  8.799979  4.673802  9.663541
Dec 2004       5.299075  3.664634  6.933516  2.799414  7.798736
```

Fig. 22: Point Forcast Data

The p-value obtained from the Ljung-Box test is highly significant, suggesting evidence of autocorrelation (**Fig 23**).

```
        Ljung-Box test

data:  Residuals from HoltWinters
Q* = 97.756, df = 24, p-value = 7.233e-11

Model df: 0.   Total lags used: 24
```

Fig. 23: Ljung Box Test

When evaluating the model's performance against the actual data for the year 2004, Lower MAE and RMSE value is achieved. The MASE value is indicating that the model performs reasonably well in capturing the data variability compared to the naive benchmark. The ACF1 value suggests a moderate level of correlation in the residuals at lag 1. Theil's U statistic of 0.326 indicates moderate forecasting performance (**Fig 24**).

```
                ME  RMSE   MAE  MPE  MAPE  MASE   ACF1 Theil's U
Training set 0.041 1.251 0.975 -Inf   Inf 0.731  0.162        NA
Test set     0.155 0.798 0.676 1.612 7.647 0.507 -0.398     0.326
```

Fig. 24: Evaluation Metrics of Holt's Method

The plot of the forecasts visually compares the HoltWinters forecasts (black line) with the actual test data (red line) (**Fig 25**).

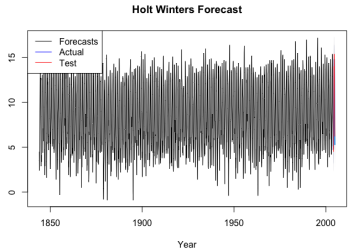


Fig. 25: Forecast Data

The residuals of Holt's method with the additive

seasonality model are evaluated for white noise, normal distribution, and autocorrelation (**Fig 26**).

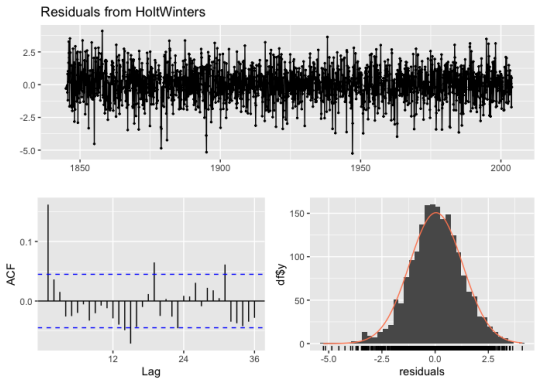


Fig. 26: Residuals of Holt's Method

Holt's method with additive seasonality provides reasonably accurate forecasts for the monthly average temperature in Armagh. However, there is room for improvement as the model should account for the significant autocorrelation present in the residuals.

3) The best model for forecasting the data is identified as SARIMA(2,0,2)(1,1,1)[12]. Additionally, alternative ARIMA models with different parameter configurations are tested, including ARIMA(2,1,0)(1,0,1)[12], ARIMA(1,1,1)(2,0,0), and ARIMA(1,1,1)(1,0,0). However, the SARIMA(2,0,2)(1,1,1)[12] model outperformed these alternatives based on the evaluation metrics and achieved a lower AIC value, indicating its superior fit to the data (**Fig 27**).

```
ARIMA(2,0,2)(1,1,1)[12]

Coefficients:
         ar1      ar2      ma1     ma2     sar1     sma1
      1.3267  -0.3287  -1.1169  0.1313  -0.0106  -0.9777
s.e.  0.1179   0.1175   0.1239  0.1210   0.0238   0.0074

sigma^2 = 1.428:  log likelihood = -3062.05
AIC=6138.11   AICc=6138.17   BIC=6176.98
```

Fig. 27: Best Sarima Model

The SARIMA(2,0,2)(1,1,1)[12] model also produces point forecasts with corresponding lower and upper bounds at the 80% and 95% confidence levels (**Fig 28**).

```
Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Jan 2004       4.560434  3.029053  6.091816  2.218388  6.902480
Feb 2004       4.702597  3.137880  6.267315  2.309568  7.095626
Mar 2004       5.938963  4.369354  7.508572  3.538453  8.339472
Apr 2004       7.723171  6.152466  9.293875  5.320985 10.125356
May 2004      10.354576  8.783433 11.925719  7.951720 12.757432
Jun 2004      12.917692 11.346263 14.489121 10.514399 15.320985
Jul 2004      14.502324 12.930654 16.073995 12.098662 16.905987
Aug 2004      14.443730 12.871833 16.015627 12.039721 16.847739
Sep 2004      12.542178 10.970060 14.114296 10.137831 14.946525
Oct 2004       9.875781  8.303444 11.448117  7.471099 12.280462
Nov 2004       6.764779  5.192225  8.337332  4.359766  9.169791
Dec 2004       5.185063  3.612294  6.757831  2.779721  7.590404
```

Fig. 28: Point Forecast Data

When evaluating the forecasts against the actual data for the year 2004, the model provides a mean error (ME) of 0.374, an RMSE of 0.861, and an MAE of 0.771. The MASE value is 0.578, indicating that the model performs reasonably well in capturing the data variability compared to the naive benchmark (**Fig 29**).

```
                ME  RMSE   MAE  MPE MAPE MASE  ACF1 Theil's U
Training set 0.012 1.189 0.927 -Inf  Inf 0.695 0.000        NA
Test set     0.374 0.861 0.771 4.316 8.645 0.578 -0.429     0.364
```

Fig. 29: Evaluation Matrics

The plot of the forecasts visually compares the SARIMA forecasts (black line) with the actual test data (red line) (**Fig 30**).

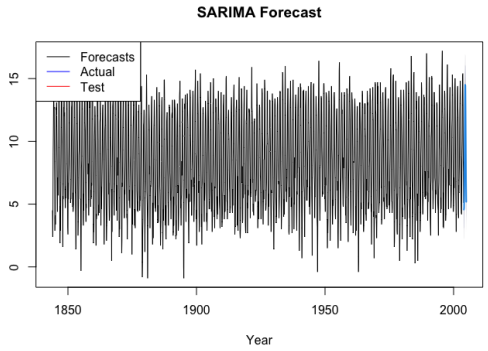


Fig. 30: SARIMA Forecast

The residuals of the SARIMA model are checked for autocorrelation using the Ljung-Box test, which yields a p-value of 0.1792, suggesting no significant autocorrelation in the residuals (**Fig 31**).

```
        Ljung-Box test

data:  Residuals from ARIMA(2,0,2)(1,1,1)[12]
Q* = 23.302, df = 18, p-value = 0.1792

Model df: 6.   Total lags used: 24
```

Fig. 31: Ljung Box Test

In addition, the residuals of the SARIMA model are checked for white noise and normal distribution. The residual standard error (RSE) for the SARIMA model is 1.189311 (**Fig 32**).

In conclusion, the SARIMA(2,0,2)(1,1,1)[12] model provides accurate and reliable forecasts for the specified time series data. The model exhibits satisfactory performance, as evidenced by low error measures. The residuals of the model demonstrate no significant autocorrelation, indicating that the model captures the underlying patterns in the data effectively.

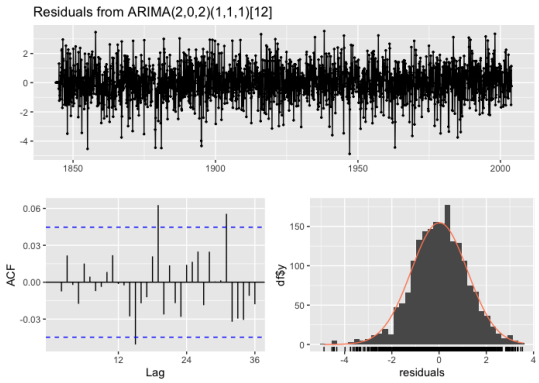


Fig. 32: Residuals of the Sarima Model

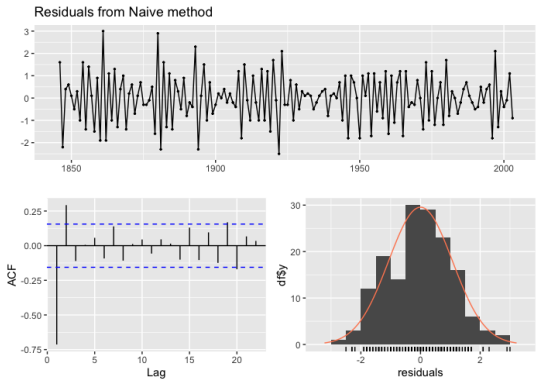


Fig. 36: Residuals From the Naive Model

VI. Model Performance - Yearly Data

1) The data is split into training and test sets, with the training set consisting of data up to the end of 2003 and the test set starting from 2004.
A Naive method is applied to forecast the data. The model yielded a residual standard deviation of 1.07. The forecasts for the test data showed a point forecast of -0.2 for 2004 (**Fig 33**).

```
Forecasts:
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2004           -0.2 -1.571206 1.171206 -2.297079 1.897079
```

Fig. 33: Forecast Data

The accuracy evaluation on the test set indicates a mean absolute error (MAE) of 0.400 and a mean absolute percentage error (MAPE) of 200% (**Fig 34**).

```
                ME RMSE   MAE MPE MAPE MASE   ACF1
Training set 0.0 1.07 0.852 NaN  Inf 1.00 -0.713
Test set     0.4 0.40 0.400 200  200 0.47     NA
```

Fig. 34: Evaluation Matrics

```
        Ljung-Box test

data:  Residuals from Naive method
Q* = 105.23, df = 10, p-value < 2.2e-16
```

Fig. 35: Ljung-Box Test

The residuals of the Naive forecast model don't appear to be white noise according to the Ljung-Box test (**Fig 35 and 36**).
Based on the evaluation metrics, the Naive method performed poorly in forecasting the yearly data, as indicated by the high mean absolute percentage error (MAPE) of 200%. Additionally, the residuals of the Naive forecast model do not appear to exhibit the characteristics of white noise, as indicated by the results of the Ljung-Box test. These findings suggest that the Naive method is not suitable for accurately predicting the yearly data.

2) The Exponential Smoothing (ETS) model, specifically the ETS(A, N, N) variant (Additive error, no trend, and no seasonality components) is applied to make forecasts for the test data.
The model produced a point forecast of 0.005 with a residual standard error (RSE) of 0.61 (**Fig 37**). The

```
Forecasts:
     Point Forecast      Lo 80     Hi 80     Lo 95    Hi 95
2004    0.005019504 -0.7810119 0.7910509 -1.197112 1.207151
```

Fig. 37: Point Forecast Data

Akaike Information Criterion (AIC) value for the ETS model is 654.4951. The accuracy evaluation on the test set revealed a mean absolute error (MAE) of 0.195 and a mean absolute percentage error (MAPE) of 97.49%. The negative autocorrelation function (ACF1) value of -0.532 suggests a moderate inverse relationship between consecutive residuals ETS model (**Fig 38**).

```
                ME  RMSE   MAE  MPE MAPE  MASE   ACF1
Training set 0.000 0.609 0.494 -Inf  Inf 0.579 -0.532
Test set     0.195 0.195 0.195 97.49 97.49 0.229     NA
```

Fig. 38: Evaluation Metrics

The plot displayed the forecasts along with the actual test data for comparison (**Fig 39**).

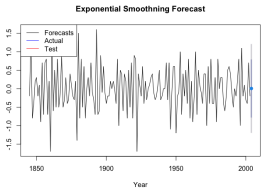


Fig. 39: Forecast Data

A residual plot is generated for the Exponential Smoothing (ETS) model, indicating whether the residuals exhibit white noise characteristics and follow a normal distribution (**Fig 40 and 41**).

```
        Ljung-Box test

data:  Residuals from ETS(A,N,N)
Q* = 57.314, df = 10, p-value = 1.164e-08

Model df: 0.   Total lags used: 10
```

Fig. 40: Ljung Box Test

Based on the Ljung-Box test results for the residuals of the ETS(A, N, N) model, the p-value is significantly low, indicating that the residuals are not consistent with white noise.

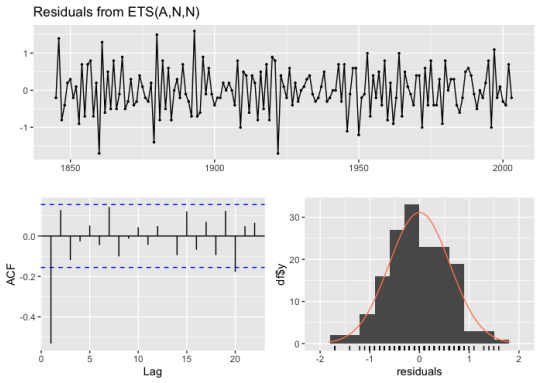


Fig. 41: Residual From Exponential Smoothing Model

This suggests that the ETS model may not adequately capture all the underlying patterns and information in the data.

3) Multiple models are evaluated for forecasting, including the ARIMA(1,1,1) and ARIMA(1,1,2) models. However, after considering the evaluation metrics and AIC value, the auto. arima function determined that the ARIMA(0,0,1) model yielded the best performance.
The model yielded a point forecast of -0.145 and a residual standard deviation (RSE) of 0.46 (**Fig 42**).

```
Forecasts:
     Point Forecast      Lo 80     Hi 80     Lo 95     Hi 95
2004     -0.1454509 -0.7395974 0.4486956 -1.05412 0.7632177
```

Fig. 42: Point Forcast Data

The Akaike Information Criterion (AIC) value for the ARIMA model was 210.93. The accuracy evaluation on the test set indicated a mean absolute error (MAE) of 0.345 and a mean absolute percentage error (MAPE) of 172.725% (**Fig 43**).

```
                    ME  RMSE   MAE     MPE    MAPE  MASE   ACF1
Training set 0.018 0.462 0.353     NaN     Inf 0.414 -0.032
Test set     0.345 0.345 0.345 172.725 172.725 0.406     NA
```

Fig. 43: Evaluation Matrics

The Ljung-Box test for the residuals showed a p-value of 0.3546, suggesting that the residuals are consistent with white noise (**Fig 44**).

```
        Ljung-Box test

data:  Residuals from ARIMA(0,0,1) with zero mean
Q* = 9.9498, df = 9, p-value = 0.3546

Model df: 1.   Total lags used: 10
```

Fig. 44: Ljung Box Test

The Ljung-Box test for the residuals showed a p-value of 0.3546, suggesting that the residuals are consistent with white noise (**Fig 45**).

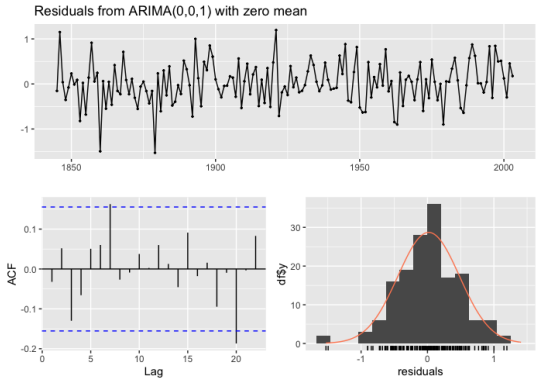


Fig. 45: Residual From ARIMA Model

The plot displayed the forecasts along with the actual test data for comparison (**Fig 46**).

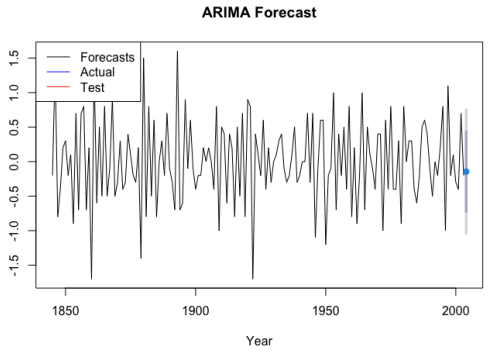


Fig. 46: ARIMA Forecast

Based on the evaluation of the test data for yearly forecasts, the ARIMA(0,0,1) model outperforms the

Exponential Smoothing (ETS) and Naive models. The ARIMA model exhibits the lowest mean absolute error (MAE) and root mean squared error (RMSE), indicating better accuracy in predicting the data. Furthermore, the residuals of the ARIMA model show lower autocorrelation (ACF1) compared to the ETS and Naive models. Therefore, the ARIMA(0,0,1) model is the preferred choice for yearly data forecasting based on these metrics.

VII. Conclusion

For Monthly data, Among the three forecasting methods, SARIMA, Holt-Winters, and Seasonal Naive, their performance is evaluated using various metrics on the test data. Holt-Winters demonstrated the best overall performance, with the lowest values for RMSE, MAE, MAPE, and MASE. It also had a reasonably good fit to the autocorrelation structure, as indicated by the near-zero negative autocorrelation function (ACF1) value. Theil's U value further supported the effectiveness of Holt-Winters in forecasting. However, SARIMA exhibits the lowest RSD (1.189311), indicating the smallest average forecast errors. To achieve accurate point forecasts Holt-Winters model is best and to capture the variability (complex patterns and autocorrelation) in the data SARIMA is the best option because of its lowest residual standard deviation.

In conclusion, when evaluating the yearly data, the Exponential Smoothing (ETS) model demonstrates better accuracy in terms of mean absolute error (MAE) and mean absolute percentage error (MAPE) for test sets. However, the ARIMA model exhibits superior overall forecast performance, considering the lower residual standard deviation (RSD) and significantly lower AIC value of 210.93. This indicates that the ARIMA model provides a better fit to the data and is more parsimonious. Therefore, based on the evaluation metrics, variability of residuals, and AIC value, the ARIMA model is the recommended choice for forecasting the yearly data.

References

[1] Tavish Srivastava, "A Complete Tutorial on Time Series Modeling in R", [Online]. Available: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling//

[2] "3 Ways to Visualize Time Series You May Not Know", [Online]. Available: https://towardsdatascience.com/3-ways-to-visualize-time-series-you-may-not-know-c8572952ea9c/

[3] "Stats and R" [Online]. Available: https://statsandr.com/blog/outliers-detection-in-r//

[4] "Time Series Analysis in R - Decomposing Time Series" [Online]. Available: https://rpubs.com/davoodastaraky/TSA1/

[5] "Interpreting ACF and PACF Plots for Time Series Forecasting" [Online]. Available: https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c: :text=Both

[6] "Augmented Dickey-Fuller Test in R" [Online]. Available: https://www.r-bloggers.com/2022/06/augmented-dickey-fuller-test-in-r//

[7] "Time Series Forecast Error Metrics You Should Know" [Online]. Available: https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27/

[8] Residual Standard Deviation: Definition, Formula, and Examples [Online]. Available: https://www.investopedia.com/terms/r/residual-standard-deviation.asp/

[9] "Ljung-Box Test: Definition + Example" [Online]. Available: https://www.statology.org/ljung-box-test//

Binary Logistic Regression Analysis for Diabetes Diagnosis based on Blood Results: A Statistical Report

*Part B: Binary Logistic Regression Analysis

Saheli Dutta
x21246513student@ncirl.ie
National College of Ireland
Masters in Data Analytics

***Abstract*—This report presents a binary logistic regression analysis conducted on a dataset of blood sample details from diabetic patients collected at an Iraqi University Hospital in 2020. The objective is to develop a predictive model for diabetes diagnosis based on blood results. Descriptive statistics and visualizations are utilized to gain insights into the dataset, while the binary logistic regression model is built and evaluated. The report discusses the model-building steps, including data preprocessing and variable selection. Assumptions of the model are verified, and model performance and fit are assessed. The results highlight the potential for accurate diabetes diagnosis based on blood test variables. In the context of this study, binary logistic regression is employed to estimate the probability of a patient being diagnosed with diabetes or not, leveraging blood test variables.**

***Index Terms*—Diabetes diagnosis, Blood Samples Variables, Binary Logistic Regression, Descriptive Statistics, Predictive Modeling, Model performance evaluations, and Statistical Study.**

I. Introduction

Diabetes is a chronic metabolic disorder characterized by high blood glucose levels. Timely and accurate diagnosis is crucial for effective management and prevention of complications. Predictive models based on blood test variables have emerged as valuable tools for diabetes diagnosis.

This study utilizes a dataset that consists of detailed blood sample information from diabetic patients. The dataset includes parameters such as urea, creatinine ratio, average blood glucose levels, cholesterol, triglycerides, HDL, LDL, VLDL, BMI, and patient demographics.

The motivation behind this research is to enhance diagnostic accuracy and efficiency in diabetes diagnosis by developing a reliable predictive model using binary logistic regression. By analyzing the relationship between the blood test variables and diabetes status, significant predictors are identified, leading to the establishment of a robust diagnostic model.

The subsequent sections of this report delve into data exploration, preprocessing steps, model-building processes, evaluation of model performance, and the final results. The report concludes with a discussion of the findings, their implications, and future research prospects.

The study highlights the potential of blood test variables as reliable indicators for diabetes diagnosis, leading to accurate and timely diagnoses and improved treatment strategies. Additionally, the developed model contributes to existing diabetes diagnosis research and encourages further investigations.

II. Related Work

The study titled "Prediction of Diabetes using Logistic Regression and Ensemble Techniques" by Priyanka Rajendra and Shahram Latifi focuses on the application of logistic regression, a commonly used classification model in machine learning, for predicting diabetes in patients. The authors explore the importance of early detection for effective diabetes management and prevention of complications. The study utilizes diagnostic measurements from two datasets: the PIMA Indians Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases, and a dataset based on a study of rural African Americans in Virginia. The authors employ feature selection techniques to enhance the model's performance and utilize ensemble methods to improve prediction accuracy. The results show the highest achieved accuracy of approximately 78% for Dataset 1 and 93% for Dataset 2. This study underscores the effectiveness of logistic regression and emphasizes the significance of data preprocessing, feature selection, and the utilization of ensemble techniques in advancing prediction models for diabetes [1].

These insights are valuable for my own research, as they provide a foundation for developing an accurate prediction model for diabetes using similar techniques. By incorporating pre-processing of data and feature selection into the model design, it is possible to enhance the accuracy and performance of the predictive system. This study serves as a vital reference, emphasizing the importance of data preprocessing and highlighting the potential benefits of utilizing logistic regression in advancing my own research on diabetes prediction and management.

III. Methodology

A. *Data Sets and Characteristics*

The dataset used in the study consists of 1000 entries with 12 columns. The dataset contains the following columns (**Fig 1**) The dataset includes a mix of numerical and categorical variables.

The independent variables (features) include Gender, AGE,

Column	Meaning	Data Type
Gender	Gender of the patient (Male or Female)	Categorical (String/Object)
Age	Age of the patient	Numeric (Integer)
Urea	Diamine, chief nitrogenous waste product in humans	Numeric (Float)
Cr	Creatinine Ratio, a parameter to assess kidney function	Numeric (Integer)
HbA1c	Average blood glucose (sugar) Levels	Numeric (Float)
Chol	Cholesterol, a parameter to assess liver function	Numeric (Float)
TG	Triglycerides, a type of fat in the blood used to transport energy	Numeric (Float)
HDL	High-density lipoprotein, the "good" cholesterol	Numeric (Float)
LDL	Low-density lipoprotein, the "bad" cholesterol	Numeric (Float)
VLDL	Very-low-density lipoprotein cholesterol	Numeric (Float)
BMI	Body Mass Index	Numeric (Float)
Diabetes	Presence of diabetes (N: No, Y: Yes, P: Pending)	Categorical (String/Object)

Fig. 1: Diabetes Dataset

Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, and BMI. The dependent variable is the CLASS column, which indicates the presence or absence of diabetes (**Fig 2**).

The ”ID” and ”No_Pation” columns in the dataset serve as

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
1	420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
2	680	87656	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
3	634	34224	F	45	2.3	24	4.0	2.9	1.0	1.0	1.5	0.4	21.0	N
4	721	34225	F	50	2.0	50	4.0	3.6	1.3	0.9	2.1	0.6	24.0	N

Fig. 2: Loaded Dataset

unique identifiers and do not hold any significant information for the prediction task of diabetes in logistic regression, therefore, they are typically excluded from the logistic regression analysis.

B. Data Exploration and Descriptive Statistics

This section focuses on conducting a descriptive analysis to enhance the understanding of the variables in the dataset. Summary statistics, such as mean, median, and standard deviation, are computed, and visualizations, including histograms and box plots, are utilized to explore the distributions of the variables and identify any outliers or patterns.

1) The unique values in the ”CLASS” column are checked, revealing three categories: ’N’ (No diabetes), ’P’ (Diabetes prediction unknown), and ’Y’ (Diabetes). For the purpose of analysis, the ’P’ cases are discarded, focusing only on the ’N’ and ’Y’ cases.

	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS	Gender_F	Gender_M
0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
2	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
3	45	2.3	24	4.0	2.9	1.0	1.0	1.5	0.4	21.0	0	1	0
4	50	2.0	50	4.0	3.6	1.3	0.9	2.1	0.6	24.0	0	1	0

Fig. 3: Modified Dataset

2) To represent gender, the ”Gender” column is one-hot encoded into two separate columns, ”Gender_F” and ”Gender_M,” signifying the female and male genders, respectively where the presence of a particular gender is indicated by a binary value (1) and the absence by (0). This encoding technique is employed to convert categorical data (gender) into a numerical format that can be easily understood and processed by machine learning algorithms [2]. The original ”Gender” column is subsequently dropped from the dataset (**Fig 3**).

3) Null values are also examined, and no missing values are found across any of the columns (**Fig 4**).

```
AGE         0
Urea        0
Cr          0
HbA1c       0
Chol        0
TG          0
HDL         0
LDL         0
VLDL        0
BMI         0
CLASS       0
Gender_F    0
Gender_M    0
```

Fig. 4: Absence of Null Values in The Dataset

4) The describe() function generates summary statistics for each variable in the dataset, including count, mean, standard deviation, minimum, maximum, and quartile values [3]. These statistics provide valuable insights into the central tendency, spread, and range of values within the dataset (**Fig 5**).

	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS	Gender_F
count	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000
mean	54.101373	5.159074	69.103485	8.408617	4.878691	2.362101	1.209081	2.616304	1.903485	29.893897	0.891235	0.441394
std	8.499612	2.975024	60.862961	2.544040	1.313356	1.417275	0.672423	1.127316	3.757012	4.869852	0.311508	0.496816
min	20.000000	0.500000	6.000000	0.900000	0.000000	0.300000	0.200000	0.300000	0.100000	19.000000	0.000000	0.000000
25%	51.000000	3.700000	48.000000	6.800000	4.000000	1.500000	0.900000	1.800000	0.700000	27.000000	1.000000	0.000000
50%	55.000000	4.600000	60.000000	8.100000	4.800000	2.000000	1.100000	2.500000	1.000000	30.000000	1.000000	0.000000
75%	59.000000	5.700000	73.000000	10.200000	5.600000	2.900000	1.300000	3.300000	1.500000	33.000000	1.000000	1.000000
max	79.000000	38.900000	800.000000	16.000000	10.300000	13.800000	9.900000	9.900000	35.000000	47.750000	1.000000	1.000000

Fig. 5: Dataset Description

- The mean value of ”HbA1c” indicates the average blood glucose level. The standard deviation suggests a moderate amount of variability in HbA1c values among the individuals. The minimum and maximum values represent the range of HbA1c values in the dataset. This range provides insight into the span of blood glucose levels observed.
- The quartile values give a sense of the distribution and central tendency of HbA1c levels among the patients. The median indicates that half of the patients have HbA1c levels below this value and the other half have levels above it.

5) In the data preprocessing step, outliers are removed from the dataset using two different approaches.
   - The first approach involves removing extreme outliers using the interquartile range (IQR) method,

Any value outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR is considered an outlier and removed from the dataset [4]. This method helps to remove extreme outliers that could significantly affect the analysis (**Fig 6**).

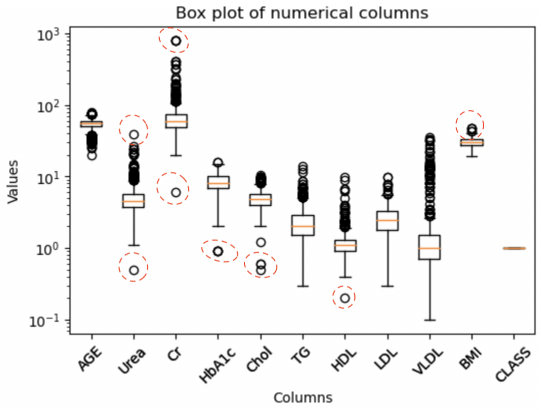


Fig. 6: After Removing Extreme Data Points Based on IQR

- Subsequently, a second approach is implemented to remove outliers that were more than three standard deviations away from the mean. This method relies on the assumption that values falling beyond three standard deviations from the mean are considered rare occurrences and likely to be outliers(**Fig 7**). The decision to use three standard deviations as

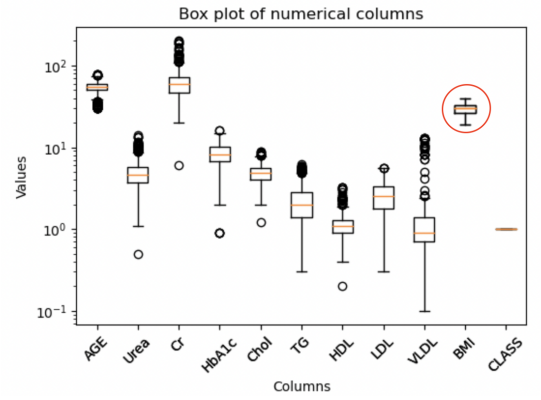


Fig. 7: Data Points Removed Based on Three Standard Deviation

the threshold to identify outliers is based on the empirical rule, also known as the 68-95-99.7 rule [5]. According to this rule, in a normal distribution, approximately 99.7% of the data falls within three standard deviations of the mean. Values beyond this range are considered rare and potential outliers.

By removing outliers, the study ensures that the dataset is more representative of the majority of observations and reduces the potential influence of extreme values on subsequent analysis and modeling.

6) **Data Visualization**

- A histogram of the variable 'AGE' stratified by the variable 'CLASS' is created to examine the distribution of age among different diabetes classes (**Fig 10**). The histogram displayed the frequency

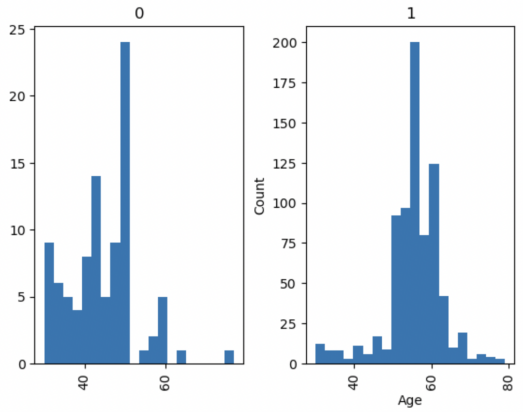


Fig. 8: Histogram of AGE by CLASS

count of age values for each class of diabetes. The histogram reveals a higher frequency of older individuals(55-60 years old) in the Positive class compared to the Negative class. This observation suggests a potential association between age and diabetes, with older individuals being more likely to have the disease.

- A scatterplot is created to visualize the relationship between the BMI (Body Mass Index) and Chol (Cholesterol) variables in the dataset (**Fig 9**). Each

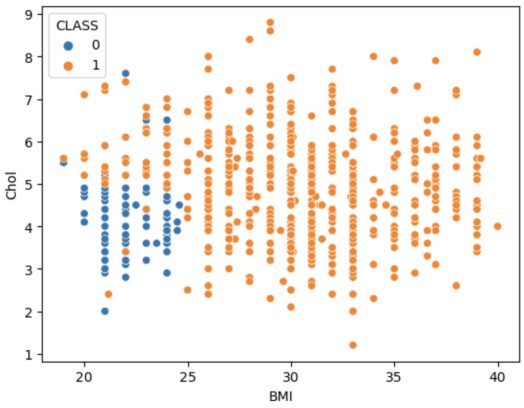


Fig. 9: Scatterplot of BMI vs Cholesterol

data point in the scatterplot represents a patient, and the color of the point represents the diabetes classification (CLASS).
From the scatterplot, it can be observed that there is a scattered distribution of data points, suggesting a potential lack of a strong linear relationship between BMI and cholesterol levels.

Specifically, it is noticeable that a significant number of patients within the age range of 20-25 exhibit lower cholesterol levels and BMI values. This observation suggests that individuals in this age group, with lower cholesterol and BMI, have a reduced likelihood of being diagnosed with diabetes.

- The target variable, "CLASS," represents the diabetes classification of the patients. The dataset consists of 754 instances classified as diabetes-positive (1) and 94 instances classified as diabetes-negative (0) (**Fig 10**).

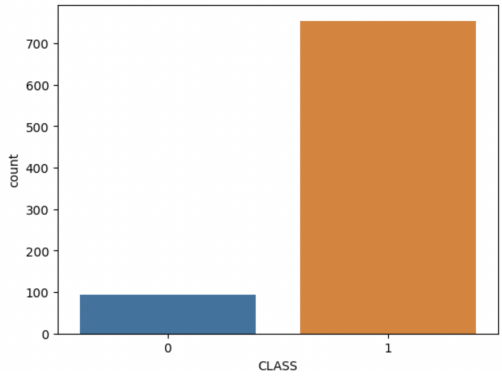


Fig. 10: Target Variable Count

7) Variance Inflation Factor (VIF) values are calculated to assess the presence of multicollinearity among the independent variables [6]. A VIF value of 5 or greater is typically considered indicative of significant multicollinearity (Fig 11). All variables have VIF values

const	0.000000
AGE	1.434043
Urea	1.716151
Cr	1.832988
HbA1c	1.331696
Chol	1.540038
TG	1.237210
HDL	1.200003
LDL	1.453272
VLDL	1.152515
BMI	1.366633
Gender_F	inf
Gender_M	inf

Fig. 11: VIF Values of Independent Variables

below the threshold of 5, suggesting that independent variables are not highly correlated with each other.
To further examine the relationship between variables, the correlation coefficients are computed. A correlation coefficient of 0.7 or higher suggests a strong linear relationship (**Fig 12**)[7].

Based on the correlation coefficients, no variables exhibit a correlation coefficient of 0.7 or higher, indicating that there are no strong linear relationships between the variables. The selected independent variables are suitable for inclusion in the logistic regression model, as they do not exhibit multicollinearity.

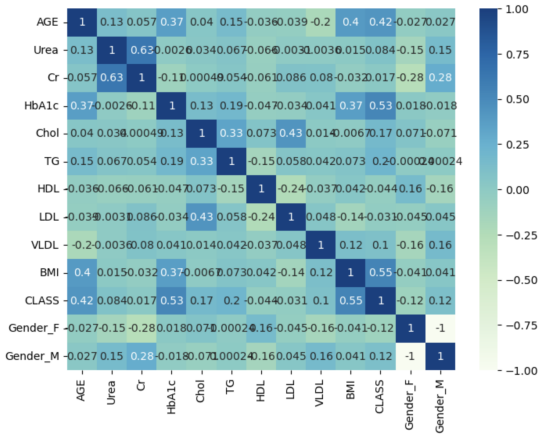


Fig. 12: Correlation Coefficients

C. *Model Building*

In the model-building phase, the dataset is split into independent features (X) and dependent feature (Y), representing the diabetes classification. The dataset is further divided into training and testing sets using the train_test_split function from the scikit-learn library [8], with a test size of 20% and a random state of 42 (for reproducibility) (**Fig 13**). The

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Fig. 13: Data Split Criteria

StandardScaler object from the scikit-learn library is used to scale the training and testing data to ensure standardized data. The independent features in the training set are transformed utilizing the scaler's fit_transform method.

The scaled independent features in the testing set have been transformed using the scaler previously fitted on the training set. The scaling process ensures consistency and allows the model to accurately predict unseen data.

The table presents a snapshot of the scaled independent features in the training set. The features, including AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, and BMI, have been standardized using the StandardScaler.
The scaled features exhibit a mean of 0 and a standard deviation of 1, ensuring all variables are on a similar scale (**Fig 14**). It is crucial to scale the data when using models

	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	Gender_F	Gender_M
708	0.073557	1.899872	0.464790	0.704430	-0.774652	-0.209577	0.853981	-1.065769	-0.231482	0.291011	0	1
769	-0.427667	-0.439775	1.412802	-0.491276	-0.515756	-0.295913	-0.629998	-1.456591	-0.289783	0.075219	0	1
354	1.451923	0.323153	-0.267764	0.588717	-0.343158	-0.209577	-0.382668	-0.088713	-0.231482	-0.140574	1	0
2	-0.552973	-0.083742	-0.698679	-1.339842	-0.515756	-1.159280	3.079950	-1.163474	-0.464684	-1.219535	1	0
887	-0.051749	2.815386	2.748637	0.627288	-0.343158	-0.813933	0.606651	-1.554297	0.409823	1.585764	0	1

Fig. 14: Scaled Independent Features

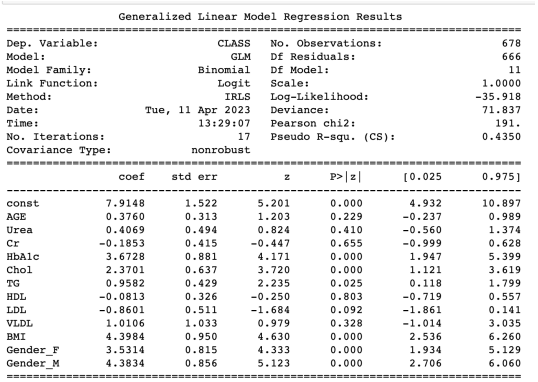
that are sensitive to the scale of the features, such as logistic regression.

D. *Final Model Parameters and Assumptions*

- The logistic regression model is constructed to predict diabetes classification based on the selected features

[12]. The model is initially fitted using all available independent features (**Fig 15**).

The model summary shows that the constant term



Generalized Linear Model Regression Results

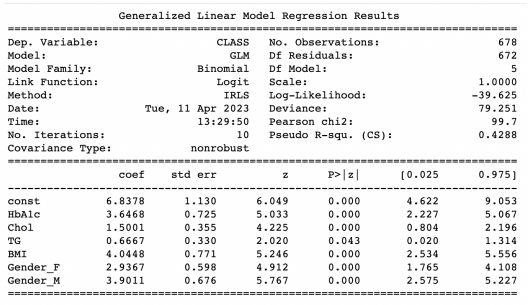
Dep. Variable:	CLASS	No. Observations:	678
Model:	GLM	Df Residuals:	666
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-35.918
Date:	Tue, 11 Apr 2023	Deviance:	71.837
Time:	13:29:07	Pearson chi2:	191.
No. Iterations:	17	Pseudo R-squ. (CS):	0.4350
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	7.9148	1.522	5.201	0.000	4.932	10.897
AGE	0.3760	0.313	1.203	0.229	-0.237	0.989
Urea	0.4069	0.494	0.824	0.410	-0.560	1.374
Cr	-0.1853	0.415	-0.447	0.655	-0.999	0.628
HbA1c	3.6728	0.881	4.171	0.000	1.947	5.399
Chol	2.3701	0.637	3.720	0.000	1.121	3.619
TG	0.9582	0.429	2.235	0.025	0.118	1.799
HDL	-0.0813	0.326	-0.250	0.803	-0.719	0.557
LDL	-0.8601	0.511	-1.684	0.092	-1.861	0.141
VLDL	1.0106	1.033	0.979	0.328	-1.014	3.035
BMI	4.3984	0.950	4.630	0.000	2.536	6.260
Gender_F	3.5314	0.815	4.333	0.000	1.934	5.129
Gender_M	4.3834	0.856	5.123	0.000	2.706	6.060

Fig. 15: The model with All Independent Features

(const) had a coefficient of 7.9148 and a standard error of 1.522. The coefficient indicates that when all other independent variables are zero, the log odds of being classified as diabetes increase by a factor of exp(7.9148), or approximately 2704.59. The p-value for the constant term was less than 0.001, indicating its statistical significance.

Among the initial independent features, several variables are found to have p-values greater than 0.05, suggesting that they are not significantly associated with the diabetes classification. Therefore, these variables are removed from the model. The updated logistic regression



Generalized Linear Model Regression Results

Dep. Variable:	CLASS	No. Observations:	678
Model:	GLM	Df Residuals:	672
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-39.625
Date:	Tue, 11 Apr 2023	Deviance:	79.251
Time:	13:29:50	Pearson chi2:	99.7
No. Iterations:	10	Pseudo R-squ. (CS):	0.4288
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.8378	1.130	6.049	0.000	4.622	9.053
HbA1c	3.6468	0.725	5.033	0.000	2.227	5.067
Chol	1.5001	0.355	4.225	0.000	0.804	2.196
TG	0.6667	0.330	2.020	0.043	0.020	1.314
BMI	4.0448	0.771	5.246	0.000	2.534	5.556
Gender_F	2.9367	0.598	4.912	0.000	1.765	4.108
Gender_M	3.9011	0.676	5.767	0.000	2.575	5.227

Fig. 16: The model with Selected Independent Features

model was then fitted using the remaining independent variables: HbA1c, Chol, TG, BMI, Gender_F, and Gender_M (**Fig 16**).

The likelihood ratio test compares the fit of the current model to a null model (a model with no predictors) to determine if the predictors in the current model significantly improve the fit. The test statistic is based on the difference in the log-likelihood values between the two models [9].

The difference in log-likelihood values of 3.707 indicates that the inclusion of all 11 predictors in the first model slightly improves the fit compared to the second model with only 5 predictors. However, this improvement is not statistically significant based on the likelihood ratio test (p-value of 1.0000), suggesting that the reduced model is sufficient for explaining the relationship between the predictors and the response variable.

The coefficients of the remaining independent variables provide insights into their relationships with the diabetes classification. For each one-unit increase in HbA1c, the log odds of being classified as diabetes increase by a factor of exp(3.6468), or approximately 38.35 holding all other variables constant (**Fig 17**).

Similarly, for each one-unit increase in Chol, TG,

```
Odds ratio for const: 932.455
Odds ratio for HbA1c: 38.353
Odds ratio for Chol: 4.482
Odds ratio for TG: 1.948
Odds ratio for BMI: 57.102
Odds ratio for Gender_F: 18.854
Odds ratio for Gender_M: 49.455
```

Fig. 17: Odds Ratio of Selected Features

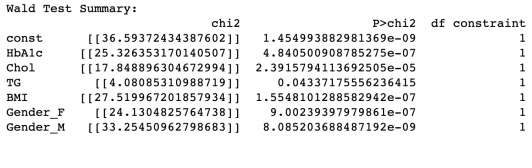
BMI, Gender_F, and Gender_M, the log odds of being classified as diabetes increase by factors of exp(1.5001), exp(0.6667), exp(4.0448), exp(2.9367), and exp(3.9011), respectively.

All remaining independent variables in the model are found to be statistically significant with p-values less than 0.05.

- **Wald Test Summary**: The Wald test is conducted to assess the significance of the coefficients in the logistic regression model. It is a statistical test that evaluates whether individual coefficients are significantly different from zero [10]. The test calculates a chi-square statistic and corresponding p-value for each coefficient.

The results suggest that HbA1c, Chol, BMI, and gender variables (Gender_F and Gender_M) are significant predictors of the response variable (**Fig 18**).

Based on the Wald test results, the interpretation of the



Wald Test Summary:

	chi2	P>chi2	df constraint
const	[[36.593724343876 02]]	1.454993882981369e-09	1
HbA1c	[[25.326353170140507]]	4.840500908785275e-07	1
Chol	[[17.848896304672994]]	2.3915794113692505e-05	1
TG	[[4.080853109887 19]]	0.043371755562364 15	1
BMI	[[27.519967201857934]]	1.5548101288582942e-07	1
Gender_F	[[24.1304825764738]]	9.00239397979861e-07	1
Gender_M	[[33.25450962798683]]	8.085203688487192e-09	1

Fig. 18: Wald Test

significant predictor is done like below: - **BMI**: The chi-square statistic is 27.52 with a p-value of approximately 1.55e-07, indicating that the BMI coefficient is significantly different from zero. This suggests that BMI is

a significant predictor of the response variable, and an increase in BMI is associated with higher odds of the outcome.

- **Chol**: The chi-square statistic is 17.85 with a p-value of approximately 2.39e-05, indicating that the Chol coefficient is significantly different from zero. This suggests that Chol is a significant predictor of the response variable, and an increase in Chol is associated with higher odds of the outcome.

IV. Model Performance

- The model is evaluated on the test dataset and the confusion matrix shows that out of 170 samples in the test dataset, 20 are correctly classified as negative (0), and 147 are correctly classified as positive (1), resulting in an overall accuracy of 98% (**Fig 19**).

The classification report provides additional performance

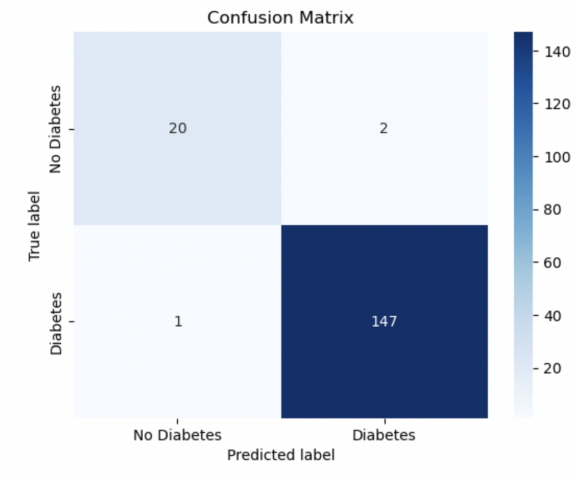


Fig. 19: Confusion Matrix

metrics [11]. The model achieved high precision (0.95 for class 0 and 0.99 for class 1), indicating a low rate of false positives. The recall values (0.91 for class 0 and 0.99 for class 1) indicate the model's ability to correctly identify positive cases (**Fig 20**).

The f1-score, which combines precision and recall, is

```
[[ 20   2]
 [  1 147]]
              precision    recall  f1-score   support

           0       0.95      0.91      0.93        22
           1       0.99      0.99      0.99       148

    accuracy                           0.98       170
   macro avg       0.97      0.95      0.96       170
weighted avg       0.98      0.98      0.98       170
```

Fig. 20: Evaluation Metrics

high for both classes (0.93 for class 0 and 0.99 for class 1), indicating overall good performance. These results suggest that the model performs well in classifying the test dataset.

- The ROC curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values and is commonly used to evaluate the performance of a binary classification model.

The AUC represents the overall performance of the model, with a higher value indicating better discrimination between the positive and negative classes.

In this study, the ROC curve and AUC are computed to assess the performance of the logistic regression model in predicting the outcome. The AUC value obtained is 0.998, indicating excellent discrimination between the classes (**Fig 21**). In the plotted ROC curve,

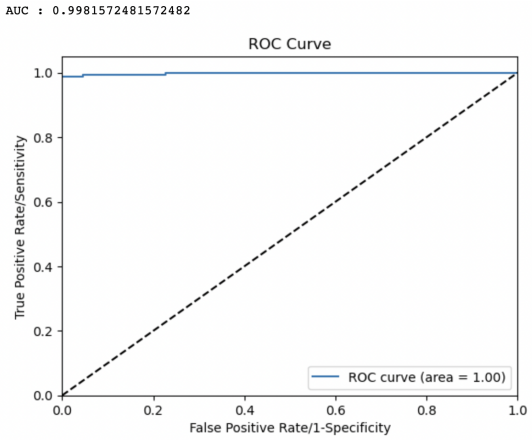


Fig. 21: Receiver operating characteristic (ROC) curve

the curve closely follows the top-left corner, further confirming the model's excellent performance. The diagonal line from 0 to 1 represents the performance of a random classifier, which has an AUC of 0.5. Overall, the high AUC value and the shape of the ROC curve indicate that the logistic regression model is capable of accurately predicting the outcome and exhibits strong discriminatory power between the classes.

- In order to further analyze the model's performance, the focus is on the "P" cases from the original data set. The logistic regression model is used to obtain predicted probabilities for these cases (**Fig 22**).

These values represent the predicted probabilities for

```
[0.96386537 0.90758035 0.99500197 0.99500197 0.99995138 0.9999976
 0.99995138 0.96680416 0.98843034 0.96680416 0.99962397 0.99984644
 0.04413298 0.99962397 0.9998815  0.99999769 0.99999959 0.99992529
 0.99992529 0.99992529 0.8562261  0.8562261  0.99346351 0.99978668
 0.99989873 0.9997655  0.99997203 0.9987308  0.99894504 0.99346351
 0.99989873 0.99985174 0.9981262  0.9997655  0.99170387 0.99949699
 0.19257292 0.99170387 0.99949699 0.19257292 0.96370156 0.99987407
 0.99971783 0.99971783 0.99987407 0.99871944 0.99999989 0.96254479
 1.         0.95864532 0.99963045 0.00871376 0.99927505]
```

Fig. 22: The array of predicted probabilities for the "P" cases

each "P" case as determined by the logistic regression model. The predicted probabilities range from very low values (e.g., 0.00871376) to high values close to 1 (e.g.,

0.99999989).

To classify these probabilities into binary predictions, a threshold of 0.5 is set. The accuracy of the predicted probabilities for the ”P” cases is calculated by comparing the binary predictions to the actual labels. In this case, the accuracy is determined to be 92.45%, which means that the model's predicted probabilities align with the actual outcomes.

The x-axis in the histogram represents the predicted probabilities of having diabetes, while the y-axis represents the frequency of ”P” cases falling within specific ranges of predicted probabilities.

The histogram indicates that a majority of ”P” cases have predicted probabilities between 0.07 and 0.10. This suggests that the model predicts with high confidence that most of these cases indeed have diabetes.

Additionally, a cluster below 0.02 implies that the model is highly confident in classifying these cases as non-diabetic. These individuals are likely to exhibit characteristics or features that strongly indicate the absence of diabetes according to the model's learned patterns. (**Fig 23**).

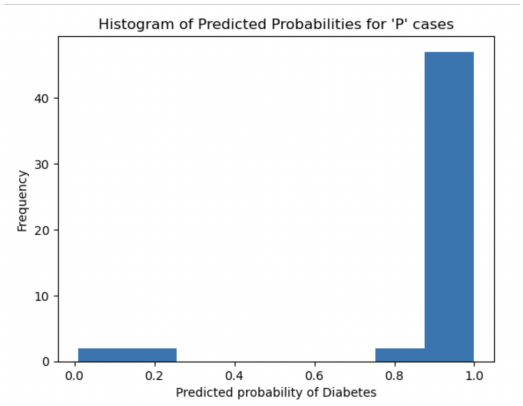


Fig. 23: Histogram of Predicted Probabilities in P Cases

The histogram demonstrates that the model can effectively distinguish between individuals with and without diabetes by assigning higher probabilities to cases known to have diabetes.

Hence, these findings suggest that the logistic regression model is a valuable tool for predicting the presence of diabetes in individuals, providing useful insights for diagnosis and treatment decisions.

V. Conclusion

In conclusion, logistic regression is applied to develop a predictive model for diabetes, considering various assumptions. These assumptions include having a binary or ordinal dependent variable, independent observations, minimal multicollinearity among independent variables, linearity of independent variables by removing outliers and transforming the data, and adequate sample size.

The model demonstrates good performance, as evidenced by high accuracy, recall, and a favorable confusion matrix. Additionally, odds ratios are calculated to interpret the effects of predictors on the odds of diabetes. The study's findings emphasize the significance of logistic regression in accurately predicting diabetes and highlight the potential for improved diagnosis, treatment, and intervention strategies in this domain. The results underscore the importance of robust statistical methods in tackling the challenges associated with diabetes prediction and management.

References

[1] Priyanka Rajendra, Shahram Latifi, ”Prediction of diabetes using logistic regression and ensemble techniques”, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990021000318/

[2] Amanda Fawcett, ”Data Science in 5 Minutes: What is One Hot Encoding?”, [Online]. Available: https://www.educative.io/blog/one-hot-encoding/

[3] pandas.DataFrame.describe [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html/

[4] ”Identifying Outliers: IQR Method ” [Online]. Available: https://online.stat.psu.edu/stat200/lesson/3/3.2: :text=Any

[5] ”Empirical Rule: Definition, Formula, Example, How It's Used” [Online]. Available: https://www.investopedia.com/terms/e/empirical-rule.asp/

[6] ”Variance Inflation Factor (VIF)” [Online]. Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asp: :text=A

[7] ”Scatter Plots and Linear Correlation” [Online]. Available: https://k12.libretexts.org/Bookshelves/Mathematics/Statistics/02

[8] Scikit-learn [Online]. Available: https://scikit-learn.org/stable/supervised$_learning.htmlsupervised-learning/$

[9] ”The Likelihood-Ratio Test” [Online]. Available: https://towardsdatascience.com/the-likelihood-ratio-test-463455b34de9/

[10] ”Wald Test: Definition, Examples, Running the Test” [Online]. Available: https://www.statisticshowto.com/wald-test//

[11] ”Classification Metrics Walkthrough: Logistic Regression with Accuracy, Precision, Recall, and ROC ” [Online]. Available: https://www.kdnuggets.com/2022/10/classification-metrics-walkthrough-logistic-regression-accuracy-precision-recall-roc.html: :text=in

[12] ”Simple Guide to Logistic Regression in R and Python ” [Online]. Available: https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r//