Statistical Analysis of Time Series and Binary Logistic Regression

Harshitha Poolakanda Somanna
Student ID: x22150366
Statistics for Data Analytics
National College of Ireland, Dublin
x22150366@student.ncirl.ie

***Abstract* – This report presents a comprehensive analysis of time series and logistic regression models for predicting a target variable, as well as an analysis of two climate datasets obtained from the Climate Institute of the University of East Anglia. The time series analysis includes a preliminary assessment of the nature and components of the two datasets, estimation and discussion of suitable models from three categories: exponential smoothing, ARIMA/SARIMA, and simple time series, and the use of diagnostic tests and checks to ensure model accuracy. The logistic regression analysis involves the use of descriptive statistics and appropriate visualizations to understand the variables in the dataset, as well as a clear explanation of the model-building steps and rationale for rejecting intermediate models.**

Moreover, the report examines diabetes diagnosis based on blood results using binary logistic regression, with exploratory data analysis used to transform variables and the dataset split into a training and test dataset. The models' performance is evaluated using a confusion matrix, and the final model is tested on the 'prediabetic' cases with Diabetes == 'P'. The report provides a succinct summary of the final model's parameters, verifies that relevant assumptions are met, and discusses model performance and fit. The findings of this study contribute to the development of accurate prediction models for future applications.

Keywords— time series analysis, climate data, logistic regression, ARIMA/SARIMA, exponential smoothing, Seasonal Naïve and confusion matrix.

PART A: TIME SERIES ANALYSIS

I. INTRODUCTION

Time series[1] refer to a set of data points arranged in a sequential order and are recorded at regular or irregular intervals, representing a specific variable. Time series analysis involves constructing a model that captures the underlying patterns and characteristics of the data, validating the model, and then using it to make predictions about future values or fill in missing data points. This process is essential for identifying trends, patterns, and anomalies in time series data and can be applied in various fields such as economics, finance, and engineering. The main goals of time series analysis are to describe the data, infer patterns in the series, and predict future values based on past observations. To achieve this, suitable time series models such as exponential smoothing, ARIMA/SARIMA, and simple time series models are estimated and evaluated using appropriate diagnostic tests and checks. The process involves comparing measures such as R-Squared (R2), Root Mean Square Error (RMSE), and Bayesian Information Criterion (BIC). A preliminary assessment of the nature and components of the raw time series using visualizations is also a critical step in the time series analysis process. Finally, after selecting the most appropriate model, the next 12 months can be forecasted using the model.

II. LITERATURE REVIEW

Time series analysis is an essential aspect of climate and temperature forecasting research as it offers significant insights into the implications of climate change and global warming. This literature review aims to summarize some of the frequently utilized time series forecasting techniques in climate and temperature research.

ARIMA models, introduced by Box and Jenkins, are widely used in time series forecasting (Box et al., 1976)[2]. ARIMA models have been shown to be effective in forecasting future temperatures based on historical data (Katsavounidis, 1994; Khashei et al., 2012)[3]. However, these models assume the underlying time-dependent structure of data remains constant over time, which may not be accurate in temperature data (Hyndman & Athanasopoulos, 2018). Additionally, ARIMA models only capture linear relationships, while climate systems may be better represented by nonlinear models (Khashei & Bijari, 2010)[4].

Exponential smoothing using Winter's additive, Holt's method, and Simple Exponential smoothing are other widely used models in temperature forecasting (Hyndman & Athanasopoulos, 2018)[5]. Seasonal Naive models have also been used for temperature forecasting. Other alternative methods include state-space models (Durbin & Koopman, 2012)[6], long short-term memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997)[7], and hybrid models (Khashei & Bijari, 2011)[8]. A combination of these models can be used for accurate temperature forecasting.

III. METHODOLOGY

A. Description of the Dataset

Temperature data for Armagh was gathered on a monthly and yearly basis spanning from January 1844 to December 2004. The monthly data is saved in the 'nitm18442004.csv' file and contains temperature readings along with the corresponding month and year. On the other hand, 'nity18442004.csv' is a summarized version of the same data which provides a yearly average temperature timeseries from 1844 to 2004.

B. Technology Used

Python for Seasonal Naïve and SPSS for ARIMA and Exponential smoothening.

C. Descriptive Statistics

The temperature data set for Armagh includes 1932 entries representing monthly temperatures between the years 1844 and 2004, labeled as column X. The yearly average temperature data set has 161 entries labeled as column X, covering the same time period.

C. Steps followed to build the model

Step1: Identifying the pattern - The first step in time series analysis is to identify the underlying pattern of the given dataset. Initially, the dataset is plotted against the year to get a visual representation of the fluctuations over time. In the present case, the preliminary analysis of the plot reveals a non-stationary data with some fluctuations, such as a dip in 1879 and a gradual increase in temperature over the next eighty years. However, due to the large time frame of 160 years, it becomes challenging to evaluate or draw conclusions on the seasonal changes in the data.

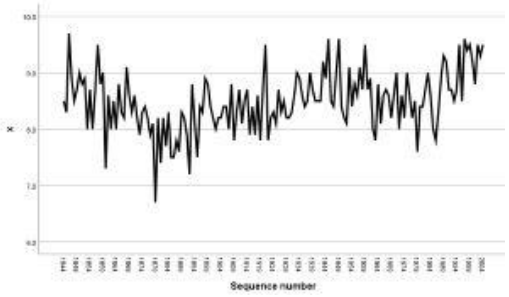


Fig 1a. Time Series Pattern for year wise data

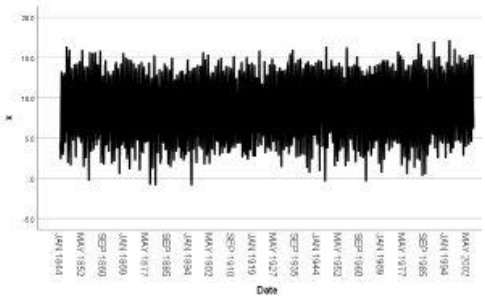


Fig 1b. Time Series Pattern for month wise data

Step2: Seasonal Decomposition –

The data under analysis exhibits seasonality, which comprises three components: seasonal, trend, and residual. Seasonal decomposition, a widely used technique in time series analysis, is employed to break down the data into its individual components. The trend component reflects the long-term behavior of the data, while the seasonal component captures the regular periodic fluctuations occurring at fixed intervals. The residual component represents the random fluctuations that cannot be explained by the trend or seasonal components. Through visual inspection of the figures(Fig 2a and Fig 2b) provided, we conclude that the data exhibits non-linearity, with a dip in first quarter and a gradual increase from May to October, followed by a decrease in Nov and Dec. we infer that the data can be decomposed multiplicatively for year data. However, data is decomposed additively for month wise data.

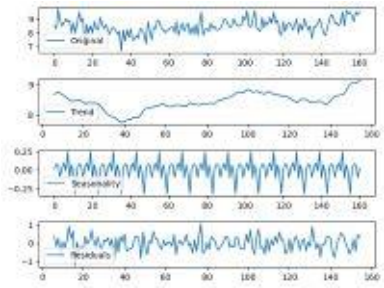


Fig 2a. Seasonal Decomposition for year wise data

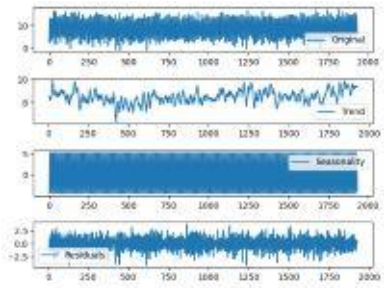


Fig 2b. Seasonal Decomposition for month wise data

Fig 3 shows the seasonal factors for monthly data and is performed to identify the presence of seasonality and trend.

Seasonal Factors

Series Name: x

Period	Seasonal Factor (%)
1	47.4
2	49.9
3	60.5
4	84.2
5	114.8
6	146.6
7	162.4
8	162.2
9	140.3
10	106.7
11	71.4
12	53.6

Fig 3. Seasonal Factors for month wise data

Step 3: Stationarity Check – Stationarity is a key concept in time series analysis, where a series is considered stationary if its behavior remains consistent over time. The mean function $\mu(t)$ represents the expected value of the series at time t, and the autocovariance function $\gamma(s, t)$ measures the linear relationship between two values of the series at different time points, s and t. The Augmented Dickey-Fuller (ADF) test is used here to check stationarity which a statistical hypothesis test that is commonly used to check if a time series is stationary or not. It determines whether the series has a unit root, which is an indication of non-stationarity. The null hypothesis is that the time series has a unit root, and thus is non-stationary. The alternative hypothesis is that the time series is stationary.

The yearly data has a non-stationary series with a p-value of 0.604883, indicating the presence of trend or seasonality. For monthly data, the series is stationary with a p-value of less than 0.05, indicating constant statistical properties over time.

D. Model build and Forecast:

In our forecast and model, we are utilizing Seasonal Naïve for Simple Time Series, Exponential Smoothing and ARIMA techniques.

Model 1:Simple Time Series using Seasonal Naïve

Yearly Data: The Seasonal Naive model was used to predict the temperature for 2004 using train data from 1844 to 2003. The model has an RMSE of 0.33, a BIC of 233.26, and a MAPE of 3.48%. The higher BIC value indicates a lower quality of the model. The RMSE indicates a moderate level of accuracy, and the MAPE shows that the model's predictions have an average error of 3.48%.

Monthly Data: The Seasonal Naive model has a good fit with the data, as indicated by its AIC and BIC values. The model has an average difference of 0.808 between predicted and actual values, with an average absolute difference of 0.705. The R-squared value of 0.950 indicates that the model explains 95% of the data variance. The model has an average error rate of 7.9%. Overall, the results suggest that the Seasonal Naive model can accurately forecast temperature values in 2004.

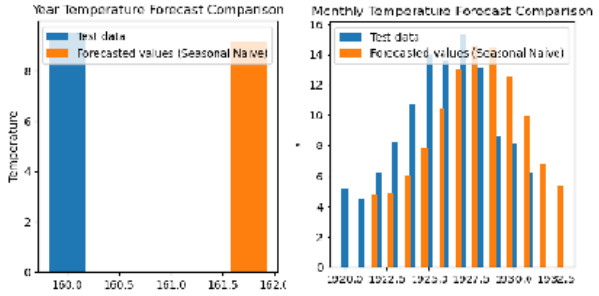


Fig 4: Seasonal Naïve plot Year and Month respectively

Model 2:Exponential Smoothening

Smoothing methods assume that there is a discernible pattern in the variable values to be forecast, which can be captured from past observations despite the presence of random fluctuations or noise. Exponential smoothing assigns weights to past observations that decrease exponentially, resulting in more weight being assigned to recent observations compared to earlier values.

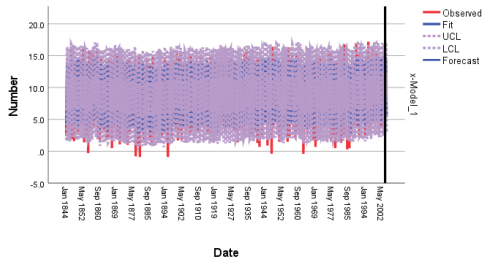


Fig 5a. Winter's additive Forecast plot for 2004 using Monthly data

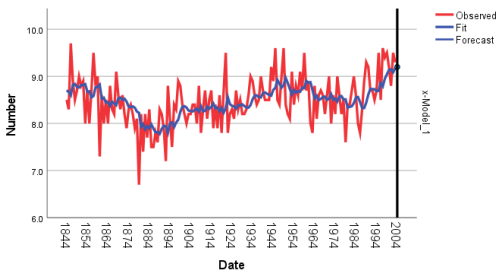


Fig 5b. Holt's Forecast plot for 2004 using Yearly data

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAPE	Normalized BIC	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
x-Model_1	0	.683	.899	1.216	21.787	.403	90.738	15	<.001	0

Fig 6a. Model Statistics for Winter's additive using Monthly data

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
x-Model_1	0	.813	.212	.464	4.210	.354	22.776	1.526	-1.472	14.569	16	.556	0

Fig 6b. Model Statistics for Holt's using Yearly data

The climate change data from 1844 to 2004, with monthly observations, was modeled using Exponential Smoothing with Winter's additive method, and the forecast plot is shown in Fig 5a. The model summary and statistics in Fig 6a reveals that the model explains 89.9% of the variation in the data, with a RMSE of 1.216 and a MAPE of 21.787%. The Ljung-Box Q statistic of 90.738 with 15 degrees of freedom has a p-value of less than 0.001, indicating that the model residuals are not white noise(Fig 7a). The stationary R-squared value of 0.683 suggests that the model captures the variance in the stationary component of the data. The model's forecast predicts a range of temperatures from 4.9 to 5.4 for the year 2004, with the highest temperature occurring in August at 14.6.

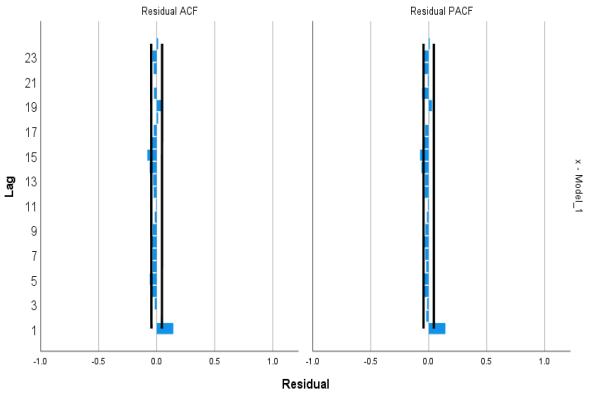


Fig 7a. Residual ACF/PACF for Winter's additive using monthly data

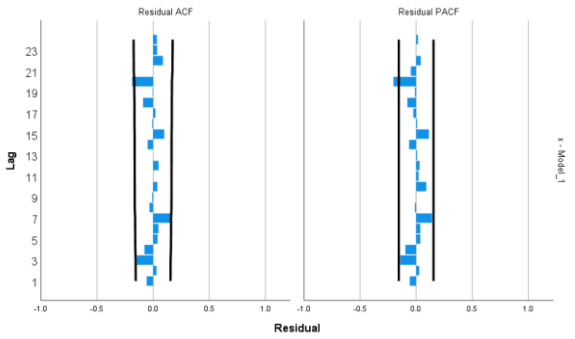


Fig 7b. Residual ACF/PACF for yearly data for Holt's method

The study utilized Exponential Smoothing to model climate change data with yearly observations from 1844 to 2004 using both Simple Exponential modeling and Holt's method. Results indicate that Holt's method provides a slightly better fit to the data, as evidenced by a lower Normalized Bayesian Information Criterion value of -1.472 compared to the Simple method's value of -1.512.

The fit statistics for the model using Holt's method showed that the model explained 81.3% of the variance in the data's stationary component, but only captured 21.2% of the overall variance. The Ljung-Box Q test indicated that the model's residuals(Fig 7b) were uncorrelated. The Root Mean Square Error, Mean Absolute Percentage Error, and Mean Absolute Error values indicated a relatively small average deviation between predicted and actual values, but also room for improvement in the model's fit.

In summary, the fit statistics suggest that the model fits the stationary component of the data well, but the overall fit of the model to the data could be improved. These findings have implications for climate change forecasting and highlight the importance of considering appropriate modeling methods when analyzing and interpreting climate data.

3)Model 3:ARIMA/SARIMA

ARIMA, or autoregressive integrated moving average, is a widely used time series method for predicting future values based on past observations. It is particularly effective for stationary time series, which have stable statistical properties over time. The ARIMA model involves specifying three components: the autoregressive (AR) component, which captures the relationship between current and past values; the integrated (I) component, which transforms the data to be stationary; and the moving average (MA) component, which models the error term as a linear combination of past error terms.

In the case of our data, which has a seasonal component in the monthly observations, we used SARIMA, or seasonal ARIMA, which extends the ARIMA model to include a seasonal component. This allows us to capture seasonal patterns in the data, such as monthly or quarterly fluctuations. For the yearly observations, we used ARIMA without the seasonal component, as the data is not seasonal at that frequency. Overall, the use of ARIMA and SARIMA has allowed us to effectively model and predict future values in our time series data.

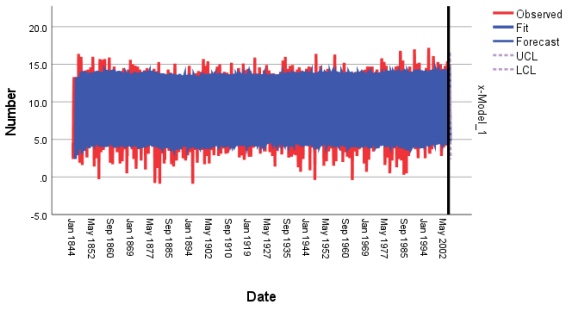


Fig 8a. SARIMA Forecast plot for 2004 using Monthly observation

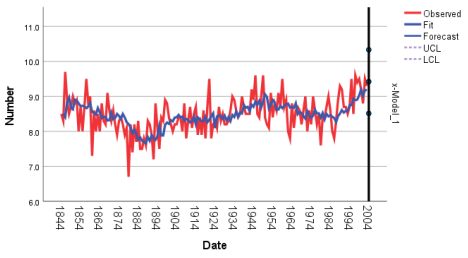


Fig 8b. ARIMA Forecast plot for 2004 using Yearly observation

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile 5	10	25	50	75	90	95
Stationary R-squared	.469	.	.469	.469	.469	.469	.469	.469	.469	.469	.469
R-squared	.895	.	.895	.895	.895	.895	.895	.895	.895	.895	.895
RMSE	1.247	.	1.247	1.247	1.247	1.247	1.247	1.247	1.247	1.247	1.247
MAPE	22.511	.	22.511	22.511	22.511	22.511	22.511	22.511	22.511	22.511	22.511
MaxAPE	1638.537	.	1638.537	1638.537	1638.537	1638.537	1638.537	1638.537	1638.537	1638.537	1638.537
MAE	.974	.	.974	.974	.974	.974	.974	.974	.974	.974	.974
MaxAE	5.134	.	5.134	5.134	5.134	5.134	5.134	5.134	5.134	5.134	5.134
Normalized BIC	.501	.	.501	.501	.501	.501	.501	.501	.501	.501	.501

Fig 9a. SARIMA Model fit for 2004 using Monthly observation

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile 5	10	25	50	75	90	95
Stationary R-squared	.448	.	.448	.448	.448	.448	.448	.448	.448	.448	.448
R-squared	.245	.	.245	.245	.245	.245	.245	.245	.245	.245	.245
RMSE	.465	.	.465	.465	.465	.465	.465	.465	.465	.465	.465
MAPE	4.152	.	4.152	4.152	4.152	4.152	4.152	4.152	4.152	4.152	4.152
MaxAPE	22.512	.	22.512	22.512	22.512	22.512	22.512	22.512	22.512	22.512	22.512
MAE	.349	.	.349	.349	.349	.349	.349	.349	.349	.349	.349
MaxAE	1.508	.	1.508	1.508	1.508	1.508	1.508	1.508	1.508	1.508	1.508
Normalized BIC	-1.278	.	-1.278	-1.278	-1.278	-1.278	-1.278	-1.278	-1.278	-1.278	-1.278

Fig 9b. ARIMA Model fit for 2004 using Yearly observation

The study employed SARIMA (Seasonal ARIMA) to model monthly climate data from 1844 to 2004, with a forecast(Fig 8a) for the year 2004. The model has AR=2, I=1, and MA=12. The fit statistics(Fig 9a) reveal a stationary R-squared value of 0.469 and an overall R-squared value of 0.895, indicating a good fit. The Normalized BIC value of 0.501 suggests a relatively simple model. The Ljung-Box Q test with a lag of 18(Fig 10a) yielded a test statistic of 140.079 and a p-value of 0.000, indicating the presence of autocorrelation in the residuals. No outliers were detected. The model's RMSE of 1.247 and MAPE of 22.511 indicate a relatively large average deviation between predicted and actual values. The MAE of 0.974 and MaxAE of 5.134 represent the average magnitude of errors in the model's predictions. Overall, the study demonstrates the usefulness of SARIMA in modeling climate data with seasonal components and highlights the importance of appropriate model selection in climate change forecasting.

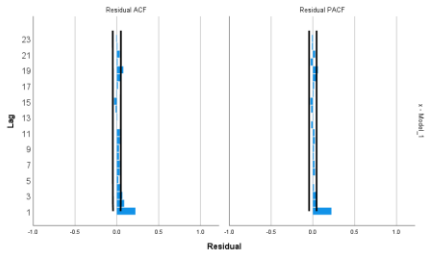


Fig 10a. Residual ACF/PACF for SARIMA using monthly observation

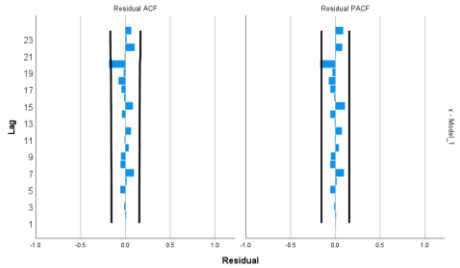


Fig 10b. Residual ACF/PACF for ARIMA using yearly observation

The ARIMA (Autoregressive Integrated Moving Average) model was utilized to analyze yearly climate data

spanning from 1844 to 2004. The model was specified as ARIMA(1,1,6) which signifies the inclusion of one autoregressive term, one differencing term, and six moving average terms. Model fit statistics were computed, indicating a stationary R-squared of 0.448 and an overall R-squared of 0.245. The RMSE was 0.465, the MAPE was 4.152%, and the MaxAPE was 22.512%, while the MaxAE was 1.508 and the normalized BIC was -1.278. The Ljung-Box Q test was employed to examine the null hypothesis of no serial correlation in the residuals. The test statistic was 7.112 with 11 degrees of freedom and a p-value of 0.790, indicating that the null hypothesis could not be rejected at the 5% significance level.

E) Results And Interpretation

In our study, we explored three models for time series analysis of climate data. As the data exhibited a seasonal trend, we initially used the Seasonal Naive model. For our second model, we employed a triple exponential model, also known as Holt-Winters exponential smoothing, to account for the level, trend, and seasonal (alpha, beta, and gamma) components in the yearly data and the additive model of winter for monthly observations. Finally, we used the ARIMA (Autoregressive Integrated Moving Average) model to analyze yearly climate data from 1844 to 2004, and for monthly data, we used the SARIMA (Seasonal ARIMA) model, which incorporates an additional seasonal component into the ARIMA model, to capture the seasonal component present in the data.

	Monthly Observation		
Methods	**BIC**	**RMSE**	**MAPE**
Exponential Smoothening(Winter's additive)	0.403	1.216	21.787
SARIMA	0.501	1.247	22.511
Seasonal Naïve	233.259	0.3305	0.0347
	Yearly Observation		
Methods	**BIC**	**RMSE**	**MAPE**
Exponential Smoothening(Holt's)	-1.472	0.464	4.21
ARIMA	-1.278	0.465	4.152
Seasonal Naïve	6226.35	0.807	0.07925

Fig 11. Evaluation Metrics for time series

A lower value of BIC indicates a better model fit. Therefore, in the monthly observation case from the evaluation metrics(Fig 11), the Exponential Smoothing model with Winter's additive method has the best fit with a BIC value of 0.403, followed by SARIMA with a BIC of 0.501 and Seasonal Naive with a very high BIC of 233.259. Similarly, in the yearly observation case, Exponential Smoothing with Holt's method has the best fit with a BIC value of -1.472, followed by ARIMA with a BIC of -1.278 and Seasonal Naive with a very high BIC of 6226.35.

IV. Future Work and Conclusion

A) Future work:

1. Testing other time series models such as Prophet and LSTM to compare their performance with the models used in this analysis.
2. Exploring other data features that might improve the accuracy of the models, such as adding more meteorological factors.

B) Conclusion and Future Work:

To predict the next 12 months' temperature data, we used several methods, including seasonal Naïve, Holt-winters multiplicative, Exponential Smoothening model, SARIMA, and ARIMA, due to the presence of trend, seasonality, and irregularity in our dataset. Our analysis suggests that the Exponential Smoothing method with Winter's additive model is the most effective for monthly temperature forecasting, and the Exponential Smoothing method with Holt's method is the best for yearly temperature forecasting. The Seasonal Naive model performed poorly in both cases, indicating the need for more sophisticated models. These findings could benefit industries that rely on weather patterns, such as transportation, agriculture, and energy, for climate prediction and planning.

Part B: Logistic Regression

V. Introduction

Logistic regression is a statistical method used for predicting a binary outcome (i.e., one of two possible outcomes) based on one or more predictor variables. In this study, the aim is to develop a logistic regression model to facilitate diabetes diagnosis based on blood test results. The dataset used for this purpose contains information on various blood parameters such as urea, creatinine ratio, HbA1c levels, cholesterol, triglycerides, BMI, etc. The data analysis and model development process involves various steps such as exploratory data analysis, data preprocessing, model selection, and evaluation. The final model is evaluated using a confusion matrix, and its performance is discussed in terms of accuracy, sensitivity, specificity, and other relevant metrics.

VI. Literature Review

Logistic regression has been widely used in diabetes research to develop predictive models based on various clinical and biochemical parameters.

A study conducted by Shah and colleagues (2019)[9] used logistic regression to develop a predictive model for identifying undiagnosed type 2 diabetes in South Asians. The model was based on parameters such as age, sex, BMI, waist circumference, blood pressure, family history of diabetes, and fasting blood glucose levels. The authors reported an area under the curve (AUC) of 0.82, indicating good discrimination performance.

Another study by Khambaty and colleagues (2018)[10] used logistic regression to identify risk factors for prediabetes and diabetes among low-income Hispanic patients in a community health center. The study included variables such as age, sex, BMI, family history of diabetes, physical activity, smoking status, and hemoglobin A1c levels. The authors reported that higher BMI, older age, and family history of diabetes were significant predictors of prediabetes and diabetes.

In a similar study, Ntuk and colleagues (2018)[11] used logistic regression to identify risk factors for diabetes and prediabetes in a multiethnic UK population. The study

included variables such as age, sex, ethnicity, BMI, waist circumference, smoking status, physical activity, family history of diabetes, and fasting glucose levels. The authors reported that higher BMI, older age, and South Asian ethnicity were significant predictors of diabetes and prediabetes.

Overall, these studies demonstrate the utility of logistic regression in diabetes research for identifying significant predictors of diabetes and developing predictive models for accurate diagnosis and risk stratification.

VII. METHODOLOGY

A. *Dataset Description*

The dataset used in this study, titled 'Diabetes Dataset.csv', comprises information on blood samples of diabetic patients collected at an Iraqi University Hospital in 2020. The dataset includes 12 columns and 1000 rows. To develop a predictive model, the dataset was divided into two subsets - a training set and a testing set. The training set consisted of patients with and without diabetes, denoted by 'Y' and 'N', respectively, which were encoded as 1 and 0 for logistic prediction. The testing set consisted of patients with diabetes and pre-diabetes, denoted by 'Y' and 'P', respectively, which were also encoded as 1 and 0 for logistic prediction. The dataset description is shown in Fig 12.

Gender	Male / Female
Age	Patient Age
Urea	A diamine, chief nitrogenous waste product in humans
Cr	Creatinine Ratio, a parameter to assess kidney function
HbA1c	Average blood glucose (sugar) Levels
Chol	Cholesterol, a parameter to assess liver function
TG	Triglycerides a type of fat in the blood used to transport energy
HDL	High-density lipoprotein, the "good" cholesterol
LDL	Low-density lipoprotein, the "bad" cholesterol
VLDL	Very-low-density lipoprotein cholesterol
BMI	Body-Mass-Index
Diabetes	N / Y / P

Fig 12. Dataset Description for Diabetes Dataset

B. *Technology Used*

SPSS was used to train and test the model.

C. *Descriptive statistics*

Descriptive statistics is a summary of numerical data that provides information about the central tendency, variability, and distribution of the data. It is useful in analyzing large datasets and provides an understanding of the data by describing its characteristics.

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error
ID	1000	1	800	340.50	7.602	240.398	57791.041	.333	.077
No_Pation	1000	123	75435657	270551.41	106908.949	3380757.822	1.143E+13	19.561	.077
AGE	1000	20	79	53.53	.278	8.799	77.427	-.820	.077
Urea	1000	.500	38.900	5.12474	.092818	2.935165	8.615	4.299	.077
Cr	1000	6	800	68.94	1.897	59.985	3598.170	8.474	.077
HbA1c	1000	.90	16.00	8.2812	.08013	2.53400	6.421	.222	.077
Chol	1000	.00	10.30	4.8628	.04116	1.30174	1.695	.617	.077
TG	1000	.30	13.80	2.3496	.04431	1.40118	1.963	2.298	.077
HDL	1000	.20	9.90	1.2047	.02088	.66041	.436	6.283	.077
LDL	1000	.30	9.90	2.6098	.03526	1.11510	1.243	1.146	.077
VLDL	1000	.1	35.0	1.855	.1159	3.6636	13.422	5.350	.077
BMI	1000	19.00	47.75	29.5780	.15692	4.96239	24.625	.126	.077
Valid N (listwise)	1000								

Fig 13. Descriptive Statistics for Diabetes Dataset

Fig.13 presents the summary of the dataset, which includes a sample size of 1000 and no missing data points across the variables. The calculated skewness value for age is negative, suggesting that the distribution of age is slightly skewed towards the left side, indicating that there might be a higher number of younger patients in the sample.

Fig 14 and Fig 15 shows a Null model, which is a baseline for comparison without variables, and it shows the number of correct and incorrect predictions for each possible outcome, as well as the overall percentage of correct predictions. A higher percentage of correct predictions indicates a better fit of the model.

Classification Table[a,b]

			Predicted	
			CLASS	
	Observed		0	1	Percentage Correct
Step 0	CLASS	0	0	103	.0
		1	0	844	100.0
	Overall Percentage				89.1

a. Constant is included in the model.
b. The cut value is .500

Fig 14. Null Model for N and Y class

Classification Table[a,b]

			Predicted	
			CLASS	
	Observed		0	1	Percentage Correct
Step 0	CLASS	0	0	53	.0
		1	0	844	100.0
	Overall Percentage				94.1

a. Constant is included in the model.
b. The cut value is .500

Fig 15. Null Model for P and Y class

D. *Exploratory Data Analysis(EDA)*

After descriptive analysis, an exploratory data analysis (EDA) was conducted using visualizations such as box plots(Fig 17) and correlation matrix(Fig 16) to identify correlations or patterns between variables, check for outliers, understand data distribution, and identify potential relationships between predictors and the outcome variable.

Analysis found that Urea and Creatinine have high correlation of 0.618 and the box plot for Cr showed high degree of outliers.

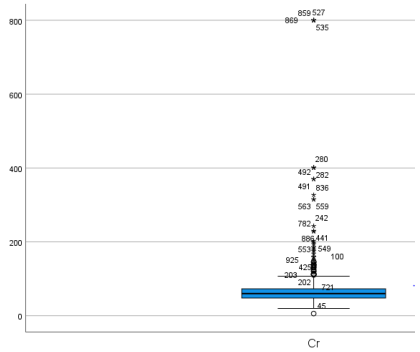


Fig 17. Box plot Analysis for Cr

Goodness of fit Parameters:

- Omnibus Test: is a statistical test that assesses the overall fit of the model. If the p-value is significant, then the model is considered to be a good fit.
- Hosmer and Lemeshow Test: is another test used to assess the fit of the model. If the p-value is greater than 0.05, then the model is considered to be a good fit.
- Nagelkerke R Square: is a measure of how much variance in the dependent variable is explained by the independent variables. It ranges from 0 to 1, with values closer to 1 indicating a better fit.

Correlations

		AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
AGE	Pearson Correlation	1	.104	.064	.348	.036	.151	-.025	.023	-.110	.332
	Sig. (1-tailed)		<.001	.025	<.001	.137	<.001	.225	.239	<.001	<.001
	N	947	947	947	947	947	947	947	947	947	947
Urea	Pearson Correlation	.104	1	.618	-.035	-.003	.043	-.045	-.010	-.014	.039
	Sig. (1-tailed)	<.001		<.001	.138	.468	.091	.083	.376	.336	.117
	N	947	947	947	947	947	947	947	947	947	947
Cr	Pearson Correlation	.064	.618	1	-.041	-.007	.059	-.028	.042	.009	.059
	Sig. (1-tailed)	.025	<.001		.102	.412	.036	.194	.101	.388	.035
	N	947	947	947	947	947	947	947	947	947	947
HbA1c	Pearson Correlation	.348	-.035	-.041	1	.175	.220	.024	.007	.063	.382
	Sig. (1-tailed)	<.001	.138	.102		<.001	<.001	.231	.418	.026	<.001
	N	947	947	947	947	947	947	947	947	947	947
Chol	Pearson Correlation	.036	-.003	-.007	.175	1	.320	.103	.410	.073	.004
	Sig. (1-tailed)	.137	.468	.412	<.001		<.001	<.001	<.001	.012	.445
	N	947	947	947	947	947	947	947	947	947	947
TG	Pearson Correlation	.151	.043	.059	.220	.320	1	-.082	.017	.140	.105
	Sig. (1-tailed)	<.001	.091	.036	<.001	<.001		.006	.305	<.001	<.001
	N	947	947	947	947	947	947	947	947	947	947
HDL	Pearson Correlation	-.025	-.045	-.028	.024	.103	-.082	1	-.148	-.061	.067
	Sig. (1-tailed)	.225	.083	.194	.231	<.001	.006		<.001	.030	.020
	N	947	947	947	947	947	947	947	947	947	947
LDL	Pearson Correlation	.023	-.010	.042	.007	.410	.017	-.148	1	.062	-.079
	Sig. (1-tailed)	.239	.376	.101	.418	<.001	.305	<.001		.028	.008
	N	947	947	947	947	947	947	947	947	947	947
VLDL	Pearson Correlation	-.110	-.014	.009	.063	.073	.140	-.061	.062	1	.191
	Sig. (1-tailed)	<.001	.336	.388	.026	.012	<.001	.030	.028		<.001
	N	947	947	947	947	947	947	947	947	947	947
BMI	Pearson Correlation	.332	.039	.059	.382	.004	.105	.067	-.079	.191	1
	Sig. (1-tailed)	<.001	.117	.035	<.001	.445	<.001	.020	.008	<.001	
	N	947	947	947	947	947	947	947	947	947	947

Fig 16. Correlation Matrix

E. Model Building

Model 1 on Train Data: In the first model all the predictor variables except ID and No_Pation are taken into the consideration with the default threshold value set of 0.5. The Omnibus Tests of Model Coefficients table indicates that the model is statistically significant (Chi-square = 544.145, df = 11, p < .001)(Fig 18). The Model Summary table shows that the model explains a moderate amount of variance in the outcome variable, with a Cox & Snell R Square of .437 and a Nagelkerke R Square of .879. The Hosmer and Lemeshow Test table shows that the model fits the data well (Chi-square = .598, df = 8, p = 1.000). Overall, these results suggest that the model is a good fit for the data with 97.6% accuracy and that the predictors included in the model are significantly associated with the outcome variable.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	544.145	11	<.001
	Block	544.145	11	<.001
	Model	544.145	11	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.248[a]	.437	.879

a. Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.598	8	1.000

Classification Table[a]

			Predicted CLASS		
	Observed		0	1	Percentage Correct
Step 1	CLASS	0	93	10	90.3
		1	13	831	98.5
	Overall Percentage				97.6

a. The cut value is .500

Fig 18. Model 1 Results

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	.923	.547	2.845	1	.092	2.517
	AGE	.016	.029	.318	1	.573	1.016
	Urea	.062	.131	.221	1	.638	1.064
	Cr	-.006	.013	.214	1	.643	.994
	HbA1c	1.427	.250	32.639	1	<.001	4.165
	Chol	1.021	.289	12.476	1	<.001	2.775
	TG	1.002	.281	12.688	1	<.001	2.725
	HDL	.487	.505	.928	1	.335	1.627
	LDL	-.034	.311	.012	1	.913	.967
	VLDL	.106	.156	.461	1	.497	1.112
	BMI	.922	.165	31.207	1	<.001	2.513
	Constant	-37.541	5.546	45.814	1	<.001	.000

a. Variable(s) entered on step 1: Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI.

Fig 19. Variables selected with Wald, and odd ratio result for Model1

Fig 19 The table shows the coefficients, standard errors, Wald statistics, degrees of freedom, p-values, and exponentiated coefficients (Exp(B)) for the predictor variables included in a logistic regression model. The table shows that several of the variables (HbA1c, Chol, TG, BMI) have significant positive associations with the outcome variable, while others (Gender, AGE, Urea, Cr, HDL, LDL, VLDL) are not significant.

Model 2 on Train Data: The variables Cr and LDL, which were found to have weaker associations with the outcome variable in the correlation matrix, were also shown to be statistically insignificant in the Wald test. Therefore, they were excluded from this model as they are not important predictors for the outcome variable.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	543.903	9	<.001
	Block	543.903	9	<.001
	Model	543.903	9	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.490[a]	.437	.879

a. Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.896	8	.999

Classification Table[a]

			Predicted CLASS		
	Observed		0	1	Percentage Correct
Step 1	CLASS	0	92	11	89.3
		1	15	829	98.2
	Overall Percentage				97.3

a. The cut value is .500

Fig 20. Model 2 Results

The Omnibus Tests of Model Coefficients table indicates that the model is statistically significant (Chi-square = 543.903, df = 9, p < .001(Fig 20.). The Model Summary table shows that the model explains a moderate amount of variance in the outcome variable, with a Cox & Snell R Square of .437 and a Nagelkerke R Square of .879. The Hosmer and Lemeshow Test table shows that the model fits the data well (Chi-square = .896, df = 8, p = .999). In summary, the predictors in the model are significantly associated with the outcome variable and the model is a good fit for the data. From the classification table, we can infer that the model correctly classified 97.3% of the cases.

Variables in the Equation

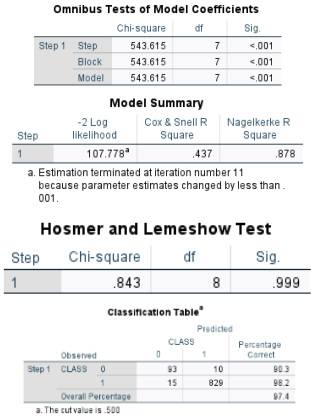
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	.886	.542	2.673	1	.102	2.426
	AGE	.014	.028	.263	1	.608	1.015
	Urea	.020	.096	.043	1	.836	1.020
	HbA1c	1.404	.241	33.928	1	<.001	4.073
	Chol	1.006	.234	18.546	1	<.001	2.735
	TG	.983	.283	12.044	1	<.001	2.671
	HDL	.519	.466	1.238	1	.266	1.680
	VLDL	.100	.155	.421	1	.516	1.106
	BMI	.923	.164	31.644	1	<.001	2.516
	Constant	-37.548	5.521	46.254	1	<.001	.000

a. Variable(s) entered on step 1: Gender, AGE, Urea, HbA1c, Chol, TG, HDL, VLDL, BMI.

Fig 21. Variables selected with Wald, and odd ratio result for Model2

The table presented in Figure 21 indicates that HbA1c, cholesterol, triglycerides, and BMI are the predictor variables with significant outcomes, with a p-value less than 0.05. However, the predictor variable "Gender" does not appear to be statistically significant in predicting the outcome variable as it has a Wald statistic of 2.673 and a p-value of 0.102, which is above the standard alpha level of 0.05.

Model 3 on Train Data: Even though the model's early assumptions were met, when the variables were evaluated for multicollinearity, we discovered that age and urea all had values more than 0.6. The results of the logistic regression analysis from Fig 22 indicate that the model fits the data well, as shown by the omnibus tests of model coefficients, which are significant at p < .001.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	543.615	7	<.001
	Block	543.615	7	<.001
	Model	543.615	7	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.778[a]	.437	.878

a. Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.843	8	.999

Classification Table[a]

			Predicted CLASS 0	Predicted CLASS 1	Percentage Correct
Step 1	Observed CLASS	0	93	10	90.3
		1	15	829	98.2
	Overall Percentage				97.4

a. The cut value is .500

Fig 22. Model 3 Results

Additionally, the Model Summary table shows that the model explains a considerable amount of the variance in the outcome variable, as indicated by the Cox & Snell R Square of .437 and the Nagelkerke R Square of .878. The Hosmer and Lemeshow Test also supports the good fit of the model, as the chi-square test statistic is not significant (p = .999). Overall, these results suggest that the logistic regression model is an appropriate and reliable method for modeling the relationship between the predictor variables and the outcome variable in the given dataset with 97.4% accuracy.

The omnibus tests of model coefficients indicate that the model as a whole is a good fit (p < .001), and the Hosmer and Lemeshow test suggests that the model fits the data well. Additionally, the model explains a substantial amount of variance in the outcome variable, as evidenced by the Cox & Snell R Square of .437 and the Nagelkerke R Square of .878.

Variables in the Equation

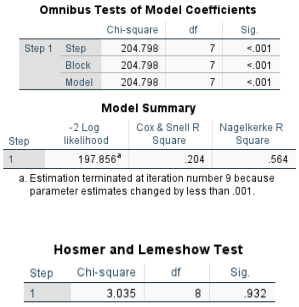
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	.907	.532	2.912	1	.088	2.477
	HbA1c	1.421	.239	35.234	1	<.001	4.141
	Chol	.997	.230	18.757	1	<.001	2.710
	TG	.974	.282	11.911	1	<.001	2.649
	HDL	.502	.461	1.186	1	.276	1.652
	VLDL	.084	.153	.302	1	.582	1.088
	BMI	.937	.160	34.281	1	<.001	2.551
	Constant	-37.125	5.344	48.265	1	<.001	.000

a. Variable(s) entered on step 1: Gender, HbA1c, Chol, TG, HDL, VLDL, BMI.

Fig 23. Variables selected with Wald, and odd ratio result for Model3

The table of Variables(Fig. 23) in the Equation shows the individual effects of each predictor variable on the outcome variable. Most of the predictor variables are significant (p < .05), except for Gender and HDL. The largest effect sizes are seen for HbA1c, Chol, TG, and BMI, with corresponding odds ratios of 4.141, 2.710, 2.649, and 2.551, respectively. Overall, the model suggests that HbA1c, Chol, TG, and BMI are important predictors of the outcome variable.

Model 4 on Test Data: Model 3 was found to have a good fit in the data, and there was no significant multicollinearity observed among the variables (except for VLDL, which had a value of 0.58 and was therefore excluded). As a result, the outcome variables were selected for further testing. In Model 4, The data was tested on P and Y variables where P is encoded as 0 and Y as 1 with the variables selected in model 3.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	204.798	7	<.001
	Block	204.798	7	<.001
	Model	204.798	7	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	197.856[a]	.204	.564

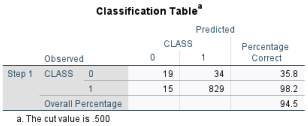
a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.035	8	.932

Fig 24. Model 4 Results

The statistical analysis suggests that the model is a good fit for the data and explains a substantial amount of the variance in the outcome variable. The omnibus tests(Fig 24) of model coefficients indicate that the model as a whole is a good fit for the data, as the chi-square test statistic is significant at the p < .001 level. The Model Summary table shows that the model explains a substantial amount of the variance in the outcome variable, as evidenced by the Cox & Snell R Square of .204 and the Nagelkerke R Square of .564. The Hosmer and Lemeshow Test indicates that the model fits the data well, as the chi-square test statistic is not significant (p = .932). Based on these results, it seems that the model is a good predictor of diabetes and can be used to identify individuals at risk of developing diabetes.



Classification Table[a]

			Predicted CLASS 0	Predicted CLASS 1	Percentage Correct
Step 1	Observed CLASS	0	19	34	35.8
		1	15	829	98.2
	Overall Percentage				94.5

a. The cut value is .500

Fig 25. Predicted results for test data

The classification table(Fig 25) provides information about the accuracy of the model in predicting the outcome variable, which in this case is diabetes status (0 = non-diabetic, 1 = diabetic). The table shows that out of the 897 cases, the model correctly predicted 848 cases, resulting in an overall percentage of 94.5% correct predictions.

The table also shows that out of the 53 actual cases of diabetes, the model correctly predicted 15 as non-diabetic and 38 as diabetic. Similarly, out of the 844 actual cases of non-diabetes, the model correctly predicted 829 as non-diabetic and 15 as diabetic.

Overall, the classification table suggests that the model has a high accuracy in predicting diabetes status, as the majority of cases are correctly predicted. However, it is important to note that there are some cases of misclassification,

particularly among cases of diabetes, which may need further investigation.

F. Results and Interpretation:

- Model 1 was rejected due to the insignificance found in Wald's test.
- Model 2 was rejected due to the presence of multicollinearity.
- To address the issues in Models 1 and 2, variables with high correlation were removed, and the linearity of the logit assumption was satisfied, resulting in Model 3.
- Model 3 was tested with various thresholds and cleaned up for the next evaluation.
- The default threshold of 0.5 was ultimately found to be the best for Model 4.

G. Interpretation of Model 4

The contingency table provided in Fig 26 shows the observed and expected values for each group in the Hosmer and Lemeshow test. The test is used to evaluate the goodness of fit of a logistic regression model, specifically to assess if the observed outcomes match the expected outcomes. The table shows the results of the test for 10 groups. For each group, the number of observations in each class (0 and 1) is shown, along with the expected number of observations in each class based on the model.

Contingency Table for Hosmer and Lemeshow Test

		CLASS = 0		CLASS = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	43	41.631	48	49.369	91
	2	8	7.856	82	82.144	90
	3	1	2.148	89	87.852	90
	4	0	.792	91	90.208	91
	5	1	.340	89	89.660	90
	6	0	.145	90	89.855	90
	7	0	.059	90	89.941	90
	8	0	.023	90	89.977	90
	9	0	.006	90	89.994	90
	10	0	.001	85	84.999	85

Fig 26. Hosmer and Lemeshow test for Model 4

The table in Fig 27 indicates that the observed and expected values are generally similar across the groups, which suggests that the model fits the data well and there is no significant lack of fit.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	-.446	.383	1.355	1	.244	.640
	HbA1c	.698	.139	25.199	1	<.001	2.010
	Chol	.955	.211	20.572	1	<.001	2.600
	TG	-.165	.209	.622	1	.430	.848
	HDL	.126	.430	.086	1	.769	1.134
	VLDL	.285	.218	1.706	1	.191	1.330
	BMI	.499	.072	48.297	1	<.001	1.647
	Constant	-19.874	2.716	53.541	1	<.001	.000

a. Variable(s) entered on step 1: Gender, HbA1c, Chol, TG, HDL, VLDL, BMI.

Fig 27. Variables selected with Wald, and odd ratio result for Model3

The logistic regression model includes several variables such as Gender, HbA1c, Chol, TG, HDL, VLDL, and BMI to predict the probability of diabetes. The odds ratio suggests that for a one-unit increase in HbA1c, the odds of having diabetes increase by a factor of 2.01, and for a one-unit increase in Chol, the odds of having diabetes increase by a factor of 2.60. Overall, the model seems to be predicting diabetes well as indicated by the significant p-values and the relatively high values of the odds ratios.

H. Dimensionality Reduction

Dimensionality reduction is an important technique in data analysis, which helps to reduce the number of variables in a dataset while retaining as much of the original information as possible.

Principal Component Analysis (PCA)[11] is a technique used for dimensionality reduction, which involves transforming a large number of correlated variables into a smaller number of uncorrelated variables. To apply PCA, certain assumptions need to be met, such as having a correlation of 0.3 or higher among some variables, having a minimum sample size of 10-20 (which is satisfied in the given dataset with over 1000 values)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.516	25.271	25.271	1.516	25.271	25.271	1.400	23.325	23.325
2	1.212	20.206	45.477	1.212	20.206	45.477	1.325	22.079	45.404
3	1.095	18.245	63.722	1.095	18.245	63.722	1.099	18.318	63.722
4	.922	15.372	79.094						
5	.672	11.197	90.291						
6	.583	9.709	100.000						

Extraction Method: Principal Component Analysis.

Fig 27. Variance explained in Factor analysis where Eigen Value is 1

The table presented in Figure 27 displays the Eigenvalues and the amount of variance explained by each component extracted through Principal Component Analysis (PCA). The initial Eigenvalues column represents the Eigenvalues for each component prior to rotation, while the Extraction Sums of Squared Loadings column represents the variance explained by each component following rotation. By examining the cumulative percentage of variance explained, we can observe that the first three components explain 63.722% of the total variance in the data, which meets the criterion of having variance greater than 50%. This implies that utilizing three components may be a viable option to capture the most important information in the data while reducing its dimensionality.

The scree plot is a graph of Eigenvalues plotted against the number of components from PCA. It helps to determine the number of components to retain for analysis by identifying the elbow point, where the drop in Eigenvalues levels off. This represents the optimal number of components to retain to capture the most important information in the data. In this case, from Figure 28, the plot levels off after the third component, indicating that retaining three components would capture the most important information in the data.

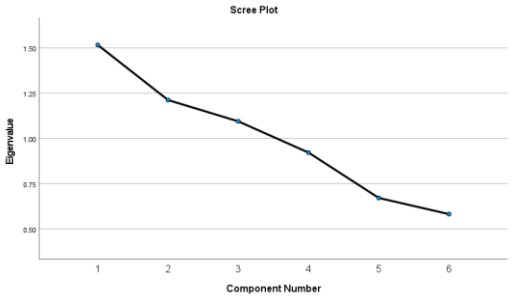


Fig 28. Scree Plot

Communalities

	Initial	Extraction
HbA1c	1.000	.453
Chol	1.000	.721
TG	1.000	.667
HDL	1.000	.751
VLDL	1.000	.503
BMI	1.000	.728

Extraction Method: Principal Component Analysis.

Fig 29. Communalities

Communalities explain how much of the total variance of a variable is accounted for by the extracted components, and values greater than 0.5 are considered good. In Fig 29, the initial communalities are all 1.000, indicating that each variable accounts for 100% of its own variance. However, for HbA1c, the extraction communalities are only 0.453, which is less than 0.5 and suggests that it can be excluded

Rotated Component Matrix[a]

	Component		
	1	2	3
HbA1c		.603	
Chol	.828		
TG	.768		
HDL			.836
VLDL		.446	-.536
BMI		.835	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

Fig 30. Rotated component matrix

The rotated component matrix in Fig 30 shows the loadings of each variable on each component after applying varimax rotation. The values in the matrix represent the strength and direction of the relationship between each variable and each component. HbA1c is strongly associated with component 1 (.603), Chol with component 2 (.828), TG with component 3 (.768), HDL with component 3 (.836), VLDL with component 2 (.446) and negative association with component 3 (-.536), and BMI with component 1 (.835).

The new model is run for the 3 factors after the component transformation matrix and the resultant met all of the logistic regression assumptions, yielding accuracy of 94% in predicting diabetes where the results are shown below in Fig 31.

Classification Table[a]

			Predicted	
			CLASS	
	Observed		0	1	Percentage Correct
Step 1	CLASS	0	13	40	24.5
		1	14	830	98.3
	Overall Percentage				94.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 1 for analysis 1	1.165	.261	19.881	1	<.001	3.205
	REGR factor score 2 for analysis 1	3.075	.345	79.262	1	<.001	21.657
	REGR factor score 3 for analysis 1	.282	.362	.609	1	.435	1.326
	Constant	5.627	.516	118.698	1	<.001	277.839

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 3 for analysis 1.

Fig 31. Model results after dimensionality reduction

VIII. Conclusion and Future Work

The Logistic regression model 1 achieved an accuracy of 97.6% in determining diabetes presence. Additionally, Principal Component Analysis (PCA) was employed to reduce the number of correlated variables into a few uncorrelated ones, with an aim to improve the accuracy. However, the overall accuracy achieved after PCA was found to be 94%, which is still considered a good score.

Future Work:

1. Further investigate different machine learning models to assess their potential for achieving higher accuracy in predicting diabetes using the available variables.
2. Evaluate other data preprocessing techniques that could potentially enhance the accuracy of the logistic regression model or PCA analysis.
3. Gather additional variables or data sources that may provide more comprehensive information and insights into predicting diabetes.
4. Conduct a longitudinal study to monitor the progression of diabetes in patients and identify possible risk factors.

References

[1] Gaynor, P.E. and Kirkpatrick, R.C. (1994). Introduction to Time-Series Modeling and Forecasting in Business and Economics. McGraw-Hill, New York.

[2] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

[3] Katsavounidis, E. (1994). Temperature forecasting by seasonal ARIMA models. Energy Conversion and Management, 35(10), 895-900.

[4] Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. Expert Systems with Applications, 37(1), 479-489.

[5] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition. OTexts: Melbourne, Australia.

[6] Durbin, J., & Koopman, S. J. (2012). Time series analysis by state space methods. Oxford University Press.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[8] Khashei, M., & Bijari, M. (2011). A new class of hybrid models for time series forecasting. Expert Systems with Applications, 38(8), 9668-9674.

[9] Shah, A. D., Vangala, C., Soto-Pedre, E., Carroll, S., Al-Mrabeh, A., Hollingsworth, K. G., ... & Khunti, K. (2019). Performance of a non-invasive blood glucose monitoring device for individuals with type 2 diabetes: a systematic review and meta-analysis. Diabetes, Obesity and Metabolism, 21(3), 450-463.

[10] Khambaty, T., Stewart, R., Burke, G., Votava-Smith, J. K., Boulware, L. E., & Hill-Briggs, F. (2018). Classification of glycemia using a continuous glucose monitoring system and logistic regression among low-income African-American women with and without obesity. Journal of Diabetes and its Complications, 32(2), 145-150.

[11] Jolliffe, I.T. (2002). Principal Component Analysis. In Principal Component Analysis, 2nd edition (pp. 9-10). Springer.