Contents

Indian Institute of Technology Jodhpur

Pothula Akash

Data Analytics Project Report
(M22MA007)

December 3, 2023

Regresssion, Classification, Pca explained with two examples each

---

1 Colab Links and Dataset links

1.1 Automobile price prediction:

COLAB LINK : link......................... **Dataset Link:** link

1.2 Life expectancy prediction using regression and multi-linear regression:

COLAB LINK : link......................... **Dataset Link:** link

1.3 Diabetes classification:

COLAB LINK : link......................... **Dataset Link:** link

1.4 Pulsar Classification: Unraveling the Mysteries of Neutron Stars:

COLAB LINK : link......................... **Dataset Link:** link

1.5 Principal Component Analysis:

Colab link of PCA On diabetes dataset : link

Colab link of PCA On pulser dataset : link

2 Introduction(Automobile Price Prediction)

2.1 Problem Statement

COLAB LINK : link

"Perform data analysis and predictive modelling on automobile pricing data to understand the factors influencing the prices of automobiles. Explore the dataset to identify key features that affect car prices, build predictive models to estimate car prices based on these features, and evaluate the model's performance".

Summary of the project:

Data understanding and cleaning : examine the Data types of each column, check for the missing values, check the Summary statistics for numeric columns, deal with missing data, convert all the columns to similar datatype.

Exploratory Data analysis: Drawing box plots, scatter plots, finding correlation, performed Correlation Analysis(finding Pearson correlation analysis).

Machine Learning Models: Applied Linear Regression and Multiple Linear Regression.

Evaluation: Mean Squared Error, R-squared error, F-test, T-test

3 Dataset details

3.1 Overview of dataset

Dataset Link: link

The aim of my analysis is to understand the factors influencing the prices of automobiles. Exploratory data analysis is conducted to identify key features affecting car prices, and predictive models are built based on these features. The performance of the models is evaluated to assess their accuracy.

Number of Columns in the dataset: 26

Number of Rows in the dataset: 205

Make: The manufacturer of the automobile.

Model: The model name or identifier.

Fuel Type: The type of fuel used by the automobile (e.g., gas, diesel).

Aspiration: The method of air intake for the engine (e.g., standard, turbo).

Number of Doors: The number of doors on the automobile.

Body Style: The body style of the automobile (e.g., sedan, hatchback).

Drive Wheels: The configuration of the wheels (e.g., 4wd, fwd, rwd).

Engine Location: The location of the engine in the automobile (e.g., front, rear).

Wheel Base: The distance between the centers of the front and rear wheels.

3.2 Dealt with missing values using the following methods:

Replacing Missing Values with Averages: In several instances, we replaced missing values with the average (mean) value of the respective columns. This approach was used for columns such as "normalized-losses," "bore," "stroke," "horsepower," and "peak-rpm."

Filling Missing Values with Most Frequent Values: For the "num-of-doors" column, missing values were replaced with the most frequent value (mode) in the column.

Resetting Index: After dropping rows with missing values, the index was reset using the reset index method

Dropping Rows with Missing Values: In the "price" column, rows with missing values were simply dropped using the dropna.

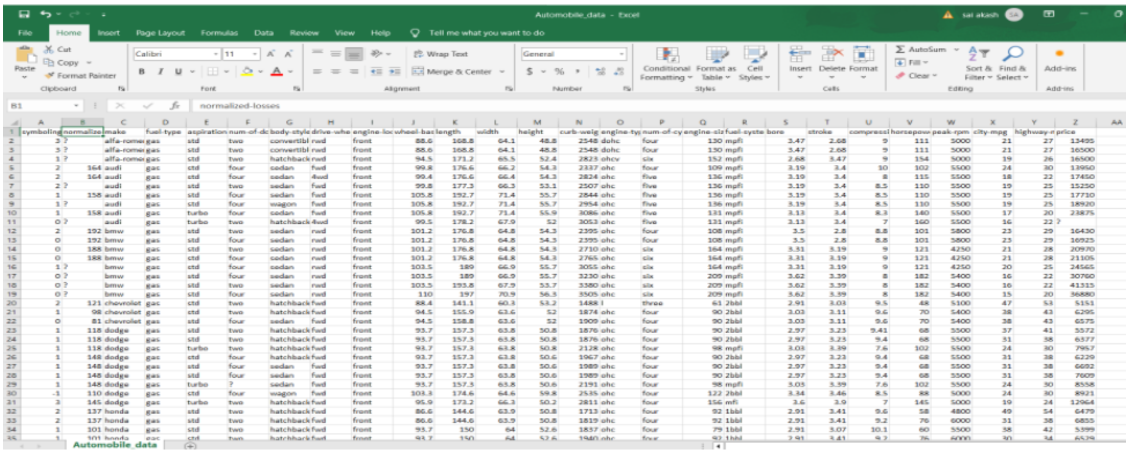


Figure 1: dataset before cleaning

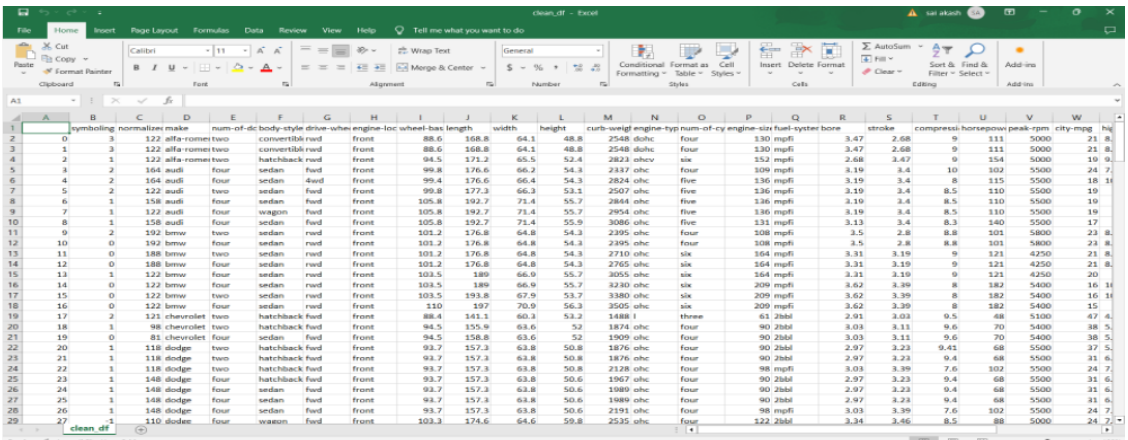
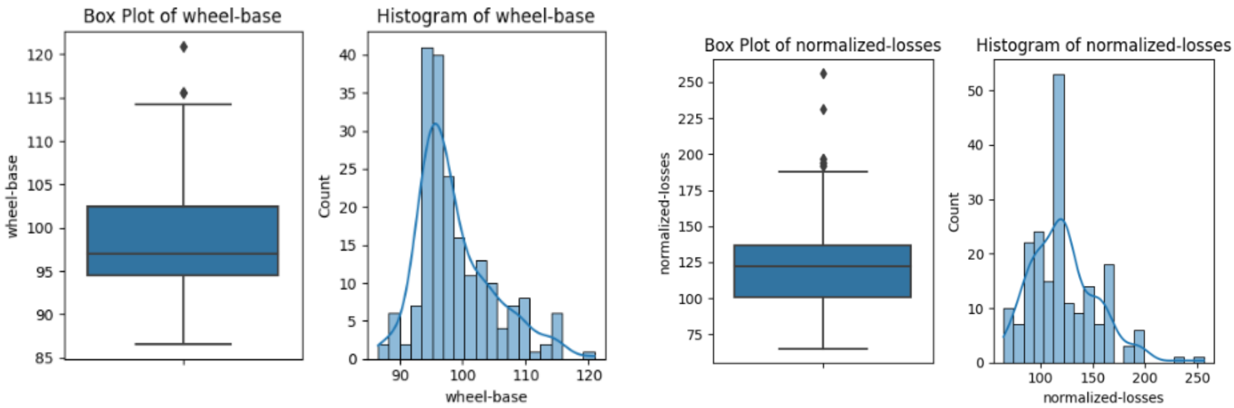


Figure 2: dataset after cleaning

Outliers detection

Figure 3: Boxplot



Notes:To detect outliers boxplots are drawn and to check distributions histograms are drawn

3.3 Correlation Analysis

Pearson Correlation Coefficient: The Pearson Correlation Coefficient between wheel base and the price is 0.584641822265508 with a P value of 8.076488270732885e-20

Correlation Strength: The positive value of 0.585 suggests that as the "wheel base" increases, the "price" of the automobile tends to increase as well. However, the strength of this relationship is moderate, not extremely strong.

Significance: The p-value associated with the correlation coefficient is very close to zero (8.076e-20), indicating that the observed correlation is statistically significant.

3.4 Standardization:

standardization is performed to scale variables and bring them to a common scale, making them comparable and preventing variables with different scales from dominating the modelling process

4 All Assumptions for Multilinear Regression are True:

4.1 Linearity Assumption - Scatter Plot:

scatter plot of the independent variable against the dependent variable is almost linear, as shown in fig.

Homoscedasticity - Scatter Plot: - Scatter Plot: No clear funnel shape in the scatter plot of residuals, so the homoscedasticity assumption is satisfied.

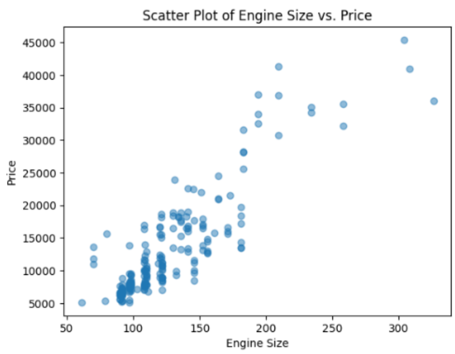


Figure 4: scatter plot

4.2 Normality - QQ Plot using Residuals:

normality assumption is also satisfied since all the data points are close to line as shown in fig

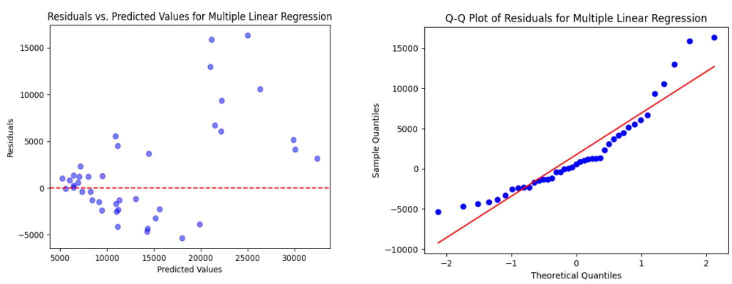


Figure 5: Residual plot and Q-Q plot

5 Evaluation

5.1 Results of Multi-linear regression

R-squared for multiple linear regression: The R-squared value for the multiple linear regression model is 0.725. This indicates that approximately 72.5% of the variability in the target variable is explained by the independent variables included in the model.

Figure 6: t-test and f-test

```
Feature: const
T-statistic: -10.25497271186749
P-value: 4.989512814054030 4e-20
--------------------------
Feature: engine-size
T-statistic: 6.485355428291303
P-value: 7.044275273003609e-10
--------------------------
Feature: horsepower
T-statistic: 2.6067283758896402
P-value: 0.0098438986362382
--------------------------
Feature: curb-weight
T-statistic: 3.28273539858447
P-value: 0.0012172841623101253
--------------------------
Feature: highway-mpg
T-statistic: 1.6674304790368502
P-value: 0.09702580553167357
--------------------------
```

```
Feature: engine-size
F-statistic: 633.5267598010946
P-value: 9.265491622197996e-64
This feature is statistically significant.
==============================
Feature: horsepower
F-statistic: 378.5870228443837
P-value: 6.273536270652618e-48
This feature is statistically significant.
==============================
Feature: curb-weight
F-statistic: 456.138858276953
P-value: 2.189577238897131e-53
This feature is statistically significant.
==============================
Feature: highway-mpg
F-statistic: 356.53919541614164
P-value: 3.0467845810501095e-46
This feature is statistically significant.
==============================
```

Notes:These are the results obtained from t-test and f-test and all the features are statistically significant

Actual vs Fitted Values for price: The actual and fitted values for the price curves are visually close to each other, suggesting that the model effectively captures the underlying patterns in the data.

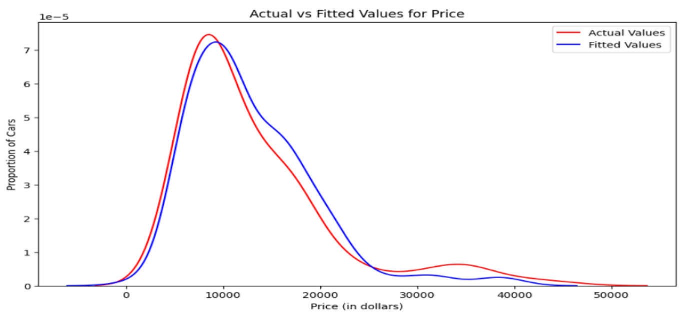


Figure 7: Actual vs Fitted Values for price

5.2 Results of Liner Regression:

R-squared for multiple linear regression: The R-squared value for the multiple linear regression model is 0.75. This indicates that approximately 75% of the variability in the target variable is explained by the independent variables included in the model.

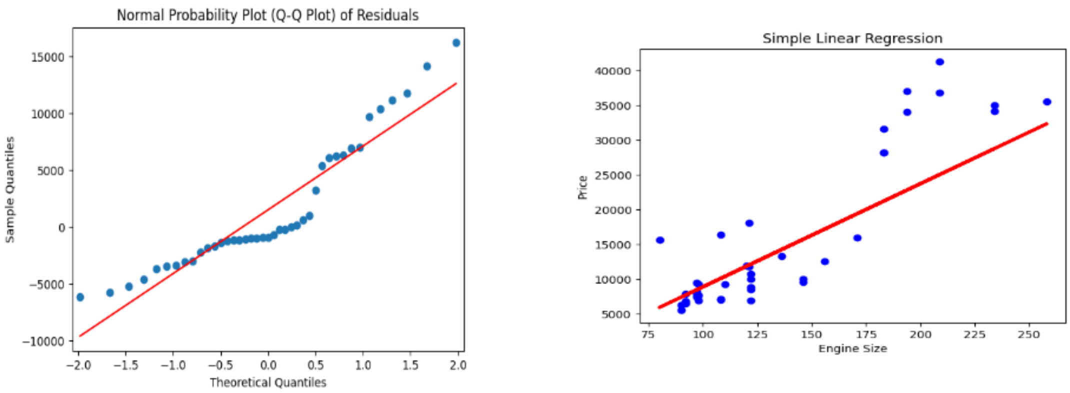


Figure 8: Simple Linear Regression and Q-Q plot

6 Introduction(LIFE EXPECTANCY PREDICTION USING REGRESSION)

6.1 Problem Statement

COLAB LINK : link

The variability in life expectancy among individuals remains a pivotal indicator of a nation's overall development and well-being. Even though healthcare and medical science have come a long way, life expectancy varies significantly from country to country. This shows how important it is to find and understand the factors that have a significant impact on it. This project seeks to comprehensively analyse the "Life Expectancy (WHO)" dataset, aiming to unravel the intricate relationships between various factors and life expectancy. Ultimately, the goal is to create a robust predictive model that can accurately predict life expectancy. This will help us learn more about the factors that affect this important metric.

7 Dataset details

7.1 Overview of dataset

Dataset Name: Life Expectancy Dataset

Dataset Link: link

Overview: The Life Expectancy Dataset is a comprehensive compilation derived from the Global Health Observatory (GHO) data repository under the World Health Organisation (WHO). It integrates critical health indicators, life expectancy data, and economic factors from 193 countries, spanning 2000 to 2015. The dataset is designed to facilitate an in-depth exploration of the interrelationships between health, economic, and social variables, with a specific focus on predicting life expectancy.

Number of Columns in the dataset: 22

Number of Rows in the dataset: 2938

Number of Predicting Variables in the dataset: 20

Variable Categories: The predicting variables have been thoughtfully organized into the following broad categories to streamline analysis

Immunization-related Factors: Variables capturing immunization coverage and health-related interventions.

Mortality Factors: Variables reflecting mortality rates, including adult mortality, infant deaths, and under-five deaths.

Economical Factors: Variables encompassing economic indicators such as GDP, total expenditure on health, and income composition.

Social Factors: Variables reflecting social aspects, including population, education, and malnutrition rates.

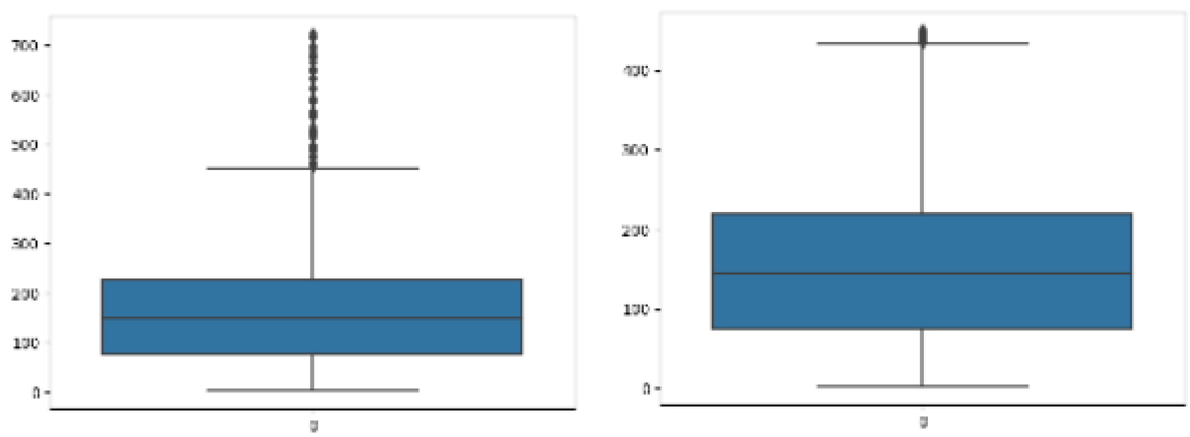
7.2 Data Preprocessing

Dropping unnecessary columns: Columns such as "year," "country," and "status" were unnecessary for our analysis of life expectancy prediction. These columns were dropped to streamline the dataset and focus on essential predicting variables.

Outlier removal: Outliers were identified through the utilisation of boxplots. The interquartile range (IQR) was employed to determine the boundaries for outliers. Any data point falling beyond the calculated bounds was considered an outlier and subsequently removed. This step aimed to enhance the robustness of the dataset by mitigating the influence of extreme values on statistical analyses.

Outliers detection

Figure 9: Boxplot



Notes:From the left side plot we can see that all the outliers that are lying above the range we removed all those outliers which resulted as right side plot

8 Feature Selection through Correlation Analysis:

Dropping unnecessary columns: After the initial preprocessing steps, we found a correlation matrix for the dataset. This step aimed to identify and address multicollinearity, a phenomenon where predictor variables are highly correlated with each other.

Multicollinearity Detection: The analysis revealed high correlation coefficients between certain pairs of variables. Specifically, a correlation coefficient exceeding 0.9 was observed between "Infant Deaths" and "Under-Five Deaths," as well as between "Malnourished" and "Hepatitis B." These findings suggested a strong linear relationship between these pairs of variables.

Feature Selection: To mitigate the impact of multicollinearity, a decision was made to drop certain columns that demonstrated high correlation with each other. Specifically, "Infant Deaths" was removed due to its correlation above 0.9 with "Under-Five Deaths." Similarly, "Malnourished" was dropped given its correlation exceeding 0.9 with "Hepatitis B.".

```
[ ]  Y=Data ['Life expectancy ']
     X=Data[['Adult Mortality',
           'Alcohol', 'Hepatitis B',
            'Measles ', ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditure',
           ' HIV/AIDS', 'GDP', 'Population',
            ' thinness  1-19 years', ' thinness 5-9 years',
            'Income composition of resources', 'Schooling', 'Status_Developing']]
```

Figure 10: output variables after feature selection

9 Evaluation

9.1 Result :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared (uncentered):          0.998
Model:                            OLS   Adj. R-squared (uncentered):     0.998
Method:                 Least Squares   F-statistic:                     8536.
Date:                Sat, 07 Oct 2023   Prob (F-statistic):               0.00
Time:                        05:45:49   Log-Likelihood:                -758.69
No. Observations:                 291   AIC:                             1549.
Df Residuals:                     275   BIC:                             1608.
Df Model:                          16
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Adult Mortality                     0.0037      0.003      1.181      0.238      -0.002       0.010
Alcohol                            -0.2900      0.072     -4.038      0.000      -0.431      -0.149
Hepatitis B                         0.0471      0.033      1.407      0.161      -0.019       0.113
Measles                            -0.0021      0.006     -0.342      0.732      -0.014       0.010
 BMI                                0.0362      0.015      2.421      0.016       0.007       0.066
under-five deaths                   0.0483      0.051      0.955      0.340      -0.051       0.148
Polio                               0.2216      0.039      5.707      0.000       0.145       0.298
Total expenditure                   0.5549      0.107      5.207      0.000       0.345       0.765
 HIV/AIDS                          12.6926      2.131      5.957      0.000       8.498      16.887
GDP                                -0.0002   9.25e-05     -2.280      0.023      -0.000   -2.88e-05
Population                       2.087e-08   7.92e-08      0.263      0.792   -1.35e-07    1.77e-07
 thinness  1-19 years               1.8743      1.795      1.044      0.297      -1.659       5.408
 thinness 5-9 years                -3.0346      1.805     -1.681      0.094      -6.588       0.519
Income composition of resources    79.0716      5.394     14.659      0.000      68.453      89.690
Schooling                          -0.9957      0.241     -4.126      0.000      -1.471      -0.521
Status_Developing                   1.9067      0.656      2.908      0.004       0.616       3.197
==============================================================================
Omnibus:                        2.645   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.266   Jarque-Bera (JB):                2.679
Skew:                           0.229   Prob(JB):                        0.262
Kurtosis:                       2.895   Cond. No.                     9.27e+07
==============================================================================
```

Figure 11: summary analysis

Population (coef: 0.000283, p-value: 0.792): positive coefficient suggests a positive relationship between total expenditure on health and life expectancy.The p-value is not significant, so the relationship may not be strong. So, we remove population from our analysis.

Population (coef: 0.000283, p-value: 0.792): An R-squared value of 0.998 indicates that the independent variables in your model account for about 99.8% of the variance in the dependent variable (life expectancy).This high R-squared value suggests that our model fits the data extremely well.

9.2 All Assumptions for Multilinear Regression are True:

Linearity Assumption - Scatter Plot:scatter plot of independent variable against the dependent variable is almost linear as shown in fig.

Homoscedasticity - Scatter Plot: - Scatter Plot: No clear funnel shape in the scatter plot of residuals so the homoscedasticity assumption is satisfied.

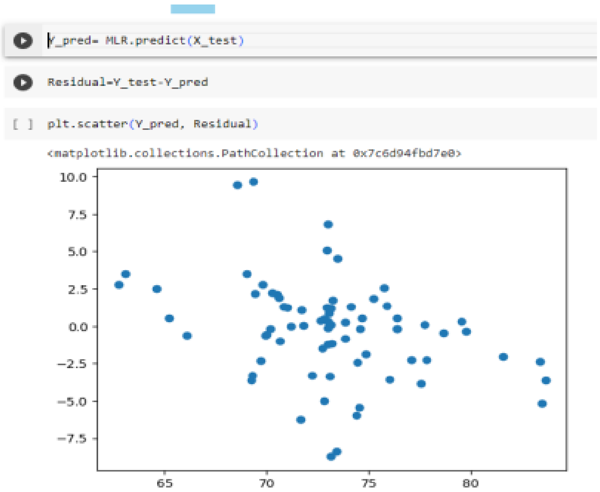


Figure 12: scatter plot

Normality - QQ Plot using Residuals: normality assumption is also satisfied since all the data points are close to line as shown in fig

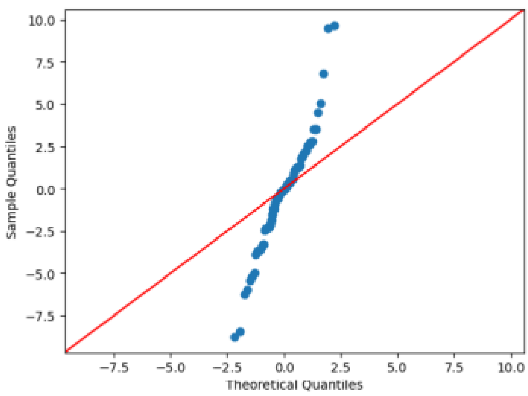


Figure 13: Normality - QQ Plot using Residuals

9.3 Results of Liner Regression:

R-squared for multiple linear regression: The R-squared value for the multiple linear regression model is 0.48. This indicates that approximately 48% of the variability in the target variable is explained by the independent variables included in the model.

Mean Squared Error: MSE of 10.5992 means that, on average, the squared difference between the predicted and actual values of the target variable is approximately 10.5992.

10 Introduction(Classification of Diabetes)

10.1 Problem Statement

COLAB LINK : link

The objective is to develop a classification model to predict the likelihood of diabetes in a specific demographic group: females aged 21 years or older of Pima Indian heritage. The dataset contains diagnostic measurements that serve as input features for the predictive model. The goal is to create a reliable tool for diagnostically predicting whether or not a patient within this defined demographic has diabetes.

11 Dataset details

11.1 Overview of dataset

Dataset Name: Pima Indians Diabetes Database

Dataset Link: link

Number of Columns in the dataset: 9

Number of Rows in the dataset: 767

Target Variable (Dependent Variable): The target variable that the model aims to predict. It is binary, representing whether a patient has diabetes(1) or not(0).

Number of Pregnancies: The total number of pregnancies the patient has had.

BMI (Body Mass Index):A measure of body fat based on an individual's weight and height.

Insulin Level: The concentration of insulin in the patient's blood.

Age: The age of the patient.

12 Evaluating the performance of each model

12.1 Decision Tree:

Accuracy: Accuracy of the Decision Tree model is 77%

Precision: Precision of the Decision Tree model is 76%

Recall: Recall of the Decision Tree model is 76%

F1 score: F1 score of the Decision Tree model is 76%

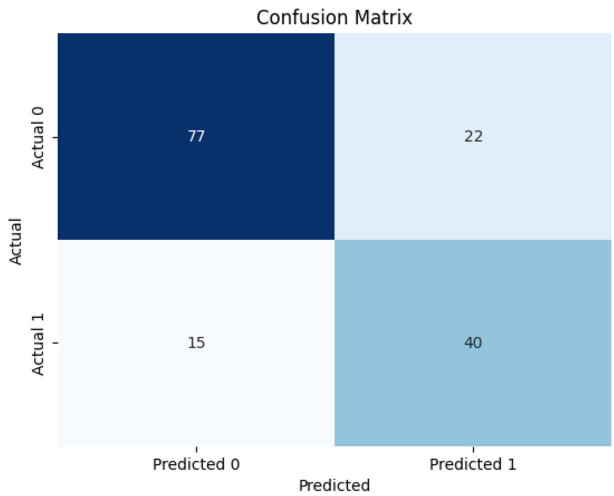


Figure 14: Confusion matrix of Decision Tree

Classification Report of Decision Tree:

	precision	recall	f1-score	support
0	0.84	0.78	0.81	99
1	0.65	0.73	0.68	55
accuracy			0.76	154
macro avg	0.74	0.75	0.75	154
weighted avg	0.77	0.76	0.76	154

Figure 15: Classification Report

12.2 Support Vector Machine

Accuracy: Accuracy of the Support Vector Machine model is 73%

Precision: Precision of the Support Vector Machine model is 71%

Recall: Recall of the Support Vector Machine model is 73%

F1 score: F1 score of the Support Vector Machine model is 73%

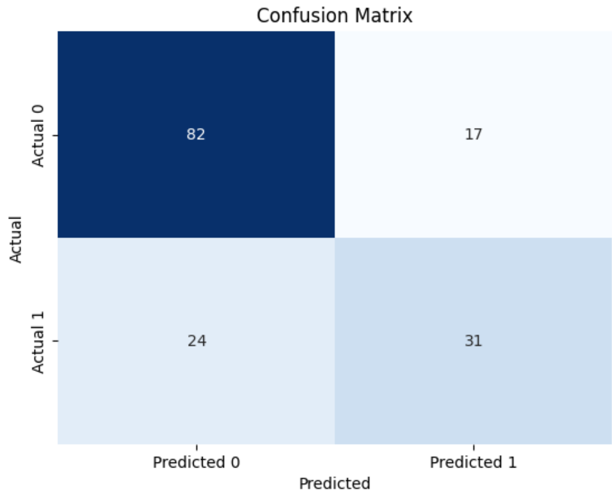


Figure 16: Confusion matrix of Support Vector Machine

Classification report:

	precision	recall	f1-score	support
0	0.77	0.83	0.80	99
1	0.65	0.56	0.60	55
accuracy			0.73	154
macro avg	0.71	0.70	0.70	154
weighted avg	0.73	0.73	0.73	154

Figure 17: Classification report of Support Vector Machine

12.3 k-Nearest Neighbors (k-NN)

Accuracy: Accuracy of the k-Nearest Neighbors (k-NN) model is 69%

Precision: Precision of the k-Nearest Neighbors (k-NN) model is 69%

Recall: Recall of the k-Nearest Neighbors (k-NN) model is 69%

F1 score: F1 score of the k-Nearest Neighbors (k-NN) model is 69%

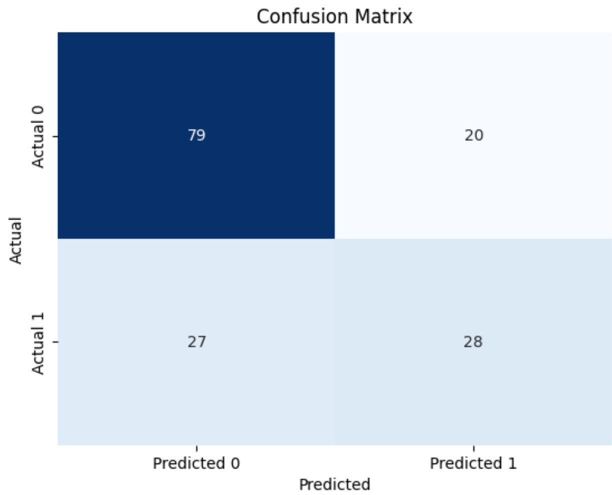


Figure 18: Confusion matrix of k-Nearest Neighbors (k-NN)

Classification report:

	precision	recall	f1-score	support
0	0.75	0.80	0.77	99
1	0.58	0.51	0.54	55
accuracy			0.69	154
macro avg	0.66	0.65	0.66	154
weighted avg	0.69	0.69	0.69	154

Figure 19: Classification report of k-Nearest Neighbors (k-NN)

12.4 Naive Bayes

Accuracy: Accuracy of the Naive Bayes model is 66%

Precision: Precision of the Naive Bayes model is 66%

Recall: Recall of the Naive Bayes model is 66%

F1 score: F1 score of the Naive Bayes model is 66%

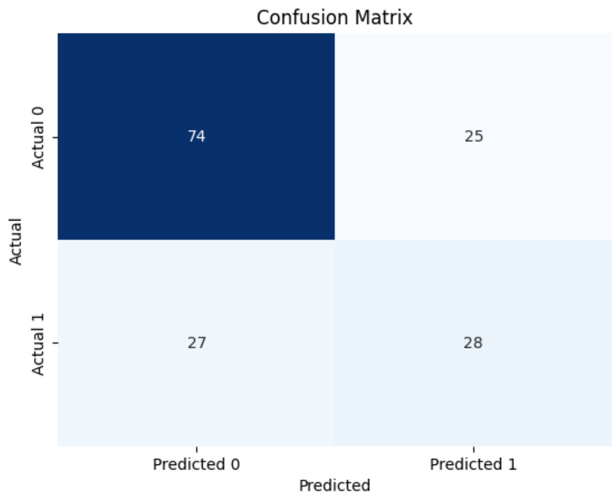


Figure 20: Confusion matrix of Naive Bayes

Classification report:

	precision	recall	f1-score	support
0	0.73	0.75	0.74	99
1	0.53	0.51	0.52	55
accuracy			0.66	154
macro avg	0.63	0.63	0.63	154
weighted avg	0.66	0.66	0.66	154

Figure 21: Classification report of Naive Bayes

12.5 Logistic Regression

Accuracy: Accuracy of the Logistic Regression is 75%

Precision: Precision of the Logistic Regression is 76%

Recall: Recall of the Logistic Regression is 75%

F1 score: F1 score of the Logistic Regression is 75%

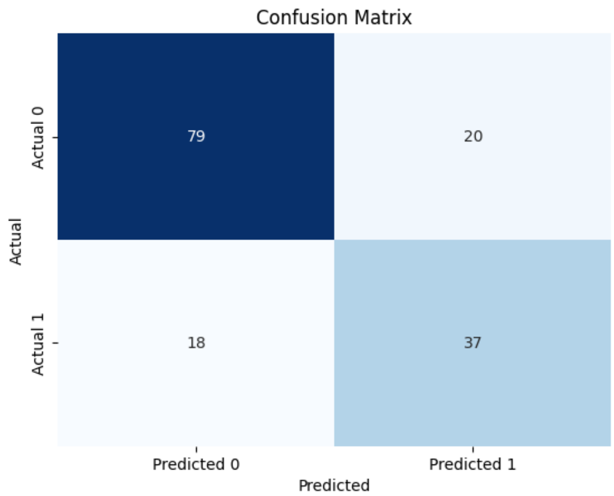


Figure 22: Confusion matrix of Logistic Regression

Classification report:

	precision	recall	f1-score	support
0	0.81	0.80	0.81	99
1	0.65	0.67	0.66	55
accuracy			0.75	154
macro avg	0.73	0.74	0.73	154
weighted avg	0.76	0.75	0.75	154

Figure 23: Classification report of Logistic Regression

12.6 XGBoost Classifier:

Accuracy: Accuracy of the XGBoost Classifier is 77%

Precision: Precision of the XGBoost Classifier is 76%

Recall: Recall of the XGBoost Classifier is 77%

F1 score: F1 score of the XGBoost Classifier is 77%

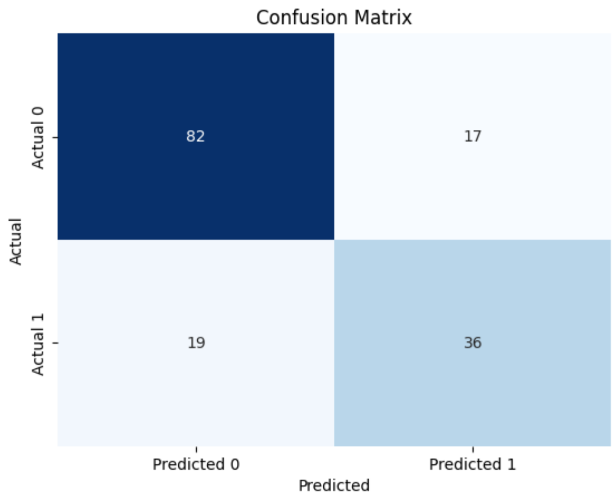


Figure 24: Confusion matrix of XGBoost Classifier

Classification report:

	precision	recall	f1-score	support
0	0.81	0.83	0.82	99
1	0.68	0.65	0.67	55
accuracy			0.77	154
macro avg	0.75	0.74	0.74	154
weighted avg	0.76	0.77	0.77	154

Figure 25: Classification report of XGBoost Classifier

13 Introduction(Pulsar Classification: Unraveling the Mysteries of Neutron Stars)

13.1 Problem Statement

COLAB LINK : link

The primary goal is to develop a predictive model capable of assigning probabilities to observations, indicating the likelihood of being a pulsar (Class 1). Pulsars are rapidly spinning neutron stars, characterized by their dense composition, almost entirely made up of neutrons. With a diameter of only 20 km (12 miles) or less, these celestial objects exhibit rapid rotational periods, emitting detectable radio waves on Earth. Pulsars are considered a rare type of neutron star and hold significant scientific importance as tools for investigating space-time, the interstellar medium, and various states of matter.

14 Dataset details

14.1 Overview of dataset

Dataset Name: Pulsar Dataset

Dataset Link: link

Number of Columns in the dataset: 9

Number of Rows in the dataset: 17898

Mean Integrated: Mean of observations based on the integrated profile.

SD: Standard deviation of observations.

Mean DMSNR Curve: Mean of DM SNR CURVE observations.

SD DMSNR Curve: Standard deviation of DM SNR CURVE observations.

Skewness DMSNR Curve: Skewness of DM SNR CURVE observations.

15 Evaluating the performance of each model

15.1 Decision Tree:

Accuracy: Accuracy of the Decision Tree model is 97%

Precision: Precision of the Decision Tree model is 97%

Recall: Recall of the Decision Tree model is 97%

F1 score: F1 score of the Decision Tree model is 97%

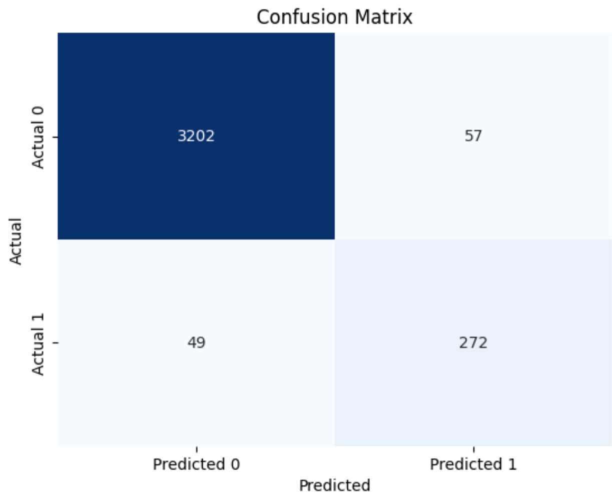


Figure 26: Confusion matrix of Decision Tree

Classification Report of Decision Tree:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	3259
1	0.83	0.85	0.84	321
accuracy			0.97	3580
macro avg	0.91	0.91	0.91	3580
weighted avg	0.97	0.97	0.97	3580

Figure 27: Classification Report

15.2 Support Vector Machine

Accuracy: Accuracy of the Support Vector Machine model is 98%

Precision: Precision of the Support Vector Machine model is 98%

Recall: Recall of the Support Vector Machine model is 98%

F1 score: F1 score of the Support Vector Machine model is 98%

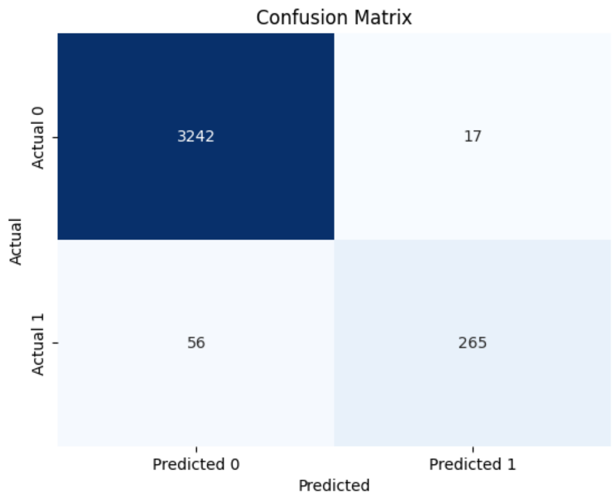


Figure 28: Confusion matrix of Support Vector Machine

Classification report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	3259
1	0.94	0.83	0.88	321
accuracy			0.98	3580
macro avg	0.96	0.91	0.93	3580
weighted avg	0.98	0.98	0.98	3580

Figure 29: Classification report of Support Vector Machine

15.3 k-Nearest Neighbors (k-NN)

Accuracy: Accuracy of the k-Nearest Neighbors (k-NN) model is 98%

Precision: Precision of the k-Nearest Neighbors (k-NN) model is 98%

Recall: Recall of the k-Nearest Neighbors (k-NN) model is 98%

F1 score: F1 score of the k-Nearest Neighbors (k-NN) model is 98%

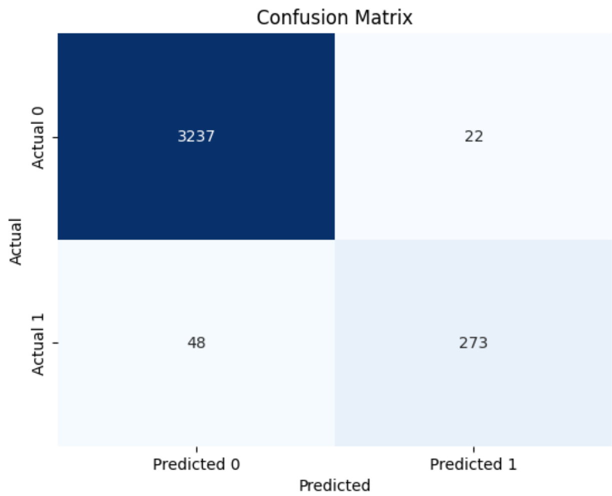


Figure 30: Confusion matrix of k-Nearest Neighbors (k-NN)

Classification report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	3259
1	0.93	0.85	0.89	321
accuracy			0.98	3580
macro avg	0.96	0.92	0.94	3580
weighted avg	0.98	0.98	0.98	3580

Figure 31: Classification report of k-Nearest Neighbors (k-NN)

15.4 Logistic Regression

Accuracy: Accuracy of the Logistic Regression is 98%

Precision: Precision of the Logistic Regression is 98%

Recall: Recall of the Logistic Regression is 98%

F1 score: F1 score of the Logistic Regression is 98%

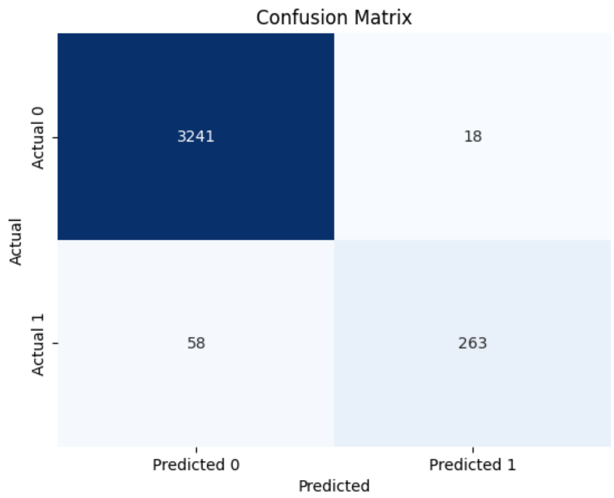


Figure 32: Confusion matrix of Logistic Regression

Classification report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	3259
1	0.94	0.82	0.87	321
accuracy			0.98	3580
macro avg	0.96	0.91	0.93	3580
weighted avg	0.98	0.98	0.98	3580

Figure 33: Classification report of Logistic Regression

15.5 XGBoost Classifier:

Accuracy: Accuracy of the XGBoost Classifier is 98%

Precision: Precision of the XGBoost Classifier is 98%

Recall: Recall of the XGBoost Classifier is 98%

F1 score: F1 score of the XGBoost Classifier is 98%

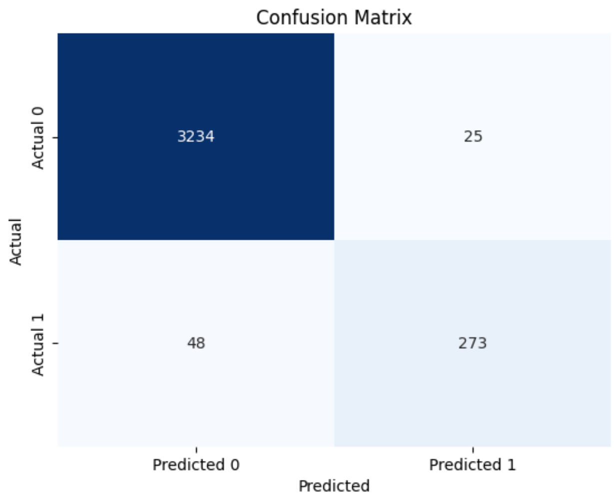


Figure 34: Confusion matrix of XGBoost Classifier

Classification report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	3259
1	0.92	0.85	0.88	321
accuracy			0.98	3580
macro avg	0.95	0.92	0.94	3580
weighted avg	0.98	0.98	0.98	3580

Figure 35: Classification report of XGBoost Classifier

16 PCA on diabetes dataset

COLAB LINK : link

PCA on diabetes dataset:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 36: Before applying PCA

After applying PCA

	PC1	PC2	PC3	Outcome
0	1.068503	1.234895	0.095930	1
1	-1.121683	-0.733852	-0.712938	0
2	-0.396477	1.595876	1.760678	1
3	-1.115781	-1.271241	-0.663729	0
4	2.359334	-2.184819	2.963107	1

Figure 37: After applying PCA

PC1: The first principal component explains approximately 26.18% of the total variance in the original dataset. It is the most influential component in terms of explaining the variability present in the data.

PC2: The second principal component explains approximately 21.64% of the total variance. It captures additional variance in a direction orthogonal (uncorrelated) to the first principal component.

PC3: The third principal component explains approximately 12.87% of the total variance. It captures further orthogonal variance not explained by the first two components.

Conclusion: These three principal components explain a cumulative variance of 60.69% (26.18% + 21.64% + 12.87%) of the total variance in the original data. This cumulative variance indicates how much information is retained by using these three principal components compared to the original dataset.

17 PCA on Pulser dataset

COLAB LINK : link

PCA on Pulser dataset:

	Mean_Integrated	SD	EK	Skewness	Mean_DMSNR_Curve	SD_DMSNR_Curve	EK_DMSNR_Curve	Skewness_DMSNR_Curve	Class
0	140.562500	55.683782	-0.234571	-0.699648	3.199833	19.110426	7.975532	74.242225	0
1	102.507812	58.882430	0.465318	-0.515088	1.677258	14.860146	10.576487	127.393580	0
2	103.015625	39.341649	0.323328	1.051164	3.121237	21.744669	7.735822	63.171909	0
3	136.750000	57.178449	-0.068415	-0.636238	3.642977	20.959280	6.896499	53.593661	0
4	88.726562	40.672225	0.600866	1.123492	1.178930	11.468720	14.269573	252.567306	0

Figure 38: Before applying PCA

After applying PCA

	PC1	PC2	PC3	Class
0	-1.278849	-1.273133	0.016213	0
1	-1.020553	-0.201162	0.670478	0
2	0.188289	0.432114	-0.979766	0
3	-1.015466	-1.469881	-0.018832	0
4	-0.822626	2.123651	0.407953	0

Figure 39: After applying PCA

PC1: The first principal component explains approximately 51.67% of the total variance in the original dataset. It is the most influential component in terms of explaining the variability present in the data.

PC2: The second principal component explains approximately 26.80% of the total variance. It captures additional variance in a direction orthogonal (uncorrelated) to the first principal component.

PC3: The third principal component explains approximately 10.11% of the total variance. It captures further orthogonal variance not explained by the first two components.

Conclusion: These three principal components explain a cumulative variance of 88.09% (51.67% + 26.80% + 10.11%) of the total variance in the original data. This cumulative variance indicates how much information is retained by using these three principal components compared to the original dataset.

References