Statistical Insights into Diabetes Risk Factors and Their Implications

Nagateja Ravilla
StudentId 11644716
nagatejaravilla@my.unt.edu

Jaswanth Srivatsav Mandadi
StudentId 11657612
Jaswanthsrivatsavmandadi@my.unt.edu

Brahma Teja Mule
StudentID 11648398
BrahmaTejaReddymule@my.unt.edu

Mounish Seeni
StudentId
mounishseeni@my.unt.edu

***Abstract*—The above comprehensive analysis might be analysis of diabetes risk factors and their further implications. The given dataset and advanced statistical methodologies and finding out their complex relationship. Here we are using machine learning techniques, EDA and data analysis. Based upon the statistical insights, The techniques might provides a further understanding of the factors of diabetes risk. And let us take our dataset and we will compare the how many people are diagnosis compared to who doesn't . considering the dataset and constructing the interval for the proportion of the people with the diabetes patient with this data we have taken we analyze deeply using the statistical methods this results to draw meaningful conclusions. And to understand the data easier we will create the graphs. we will use machine learning method and this also called as linear regression and this help us to make prediction easier and this will help to find the diabetes and after training model with our data we get the important things weights and intercept. Here we use statistical method to understand the data and machine learning to detect the details about the diabetes.**

***Index Terms*—**

I. Introduction

Here according to daily life the diabetes of risk factor Might increasing more day by day. In our project called " Statistical Insights into Diabetes Risk Factors and their Implications". It aims to know that people who might getting more diabetes and how to prevent from that. based upon the performing some models like EDA techniques, machine learning techniques to find out the value for how much their infected. The diabetes was increasing in the upcoming populations. Here we are collected the dataset from the kaggle source and based upon the given data we have performed the data cleaning because to remove the noise and some of the null value which has been present in the dataset.

In the dataset some of the parameters like pregnancies, glucose, Blood pressure, skin thickness, insulin, BMI, Diabetes pedigree function, Age and outcome. With each parameters we can find out the accuracy of the diabetes factor rate.

It data analysis involving in the quantative manner and their aspects. In the dataset we used feature selection to remove the column which is not useful for finding the diabetes. The diabetes might get influenced by some of like genetics fron ancestors, decreasing of the insulin resistance, and also risk factor of the diabetes might be increases more based upon the age because based upon the age increasing the resistance power slowly decrease so that risk factor may be increases.

The main goal of our project to find out the risk factor of diabets of the person and have to improve from that by preventing some unsual less activities, have to adopt the balanced diet are beyond our controls. Based upon the insights of IEEE papers and from our research aim to perform the implementation to finding out the risk factor and its accuracy to improve their rate. It helps to preventation based upon the rate of Diabetes .

Identify applicable funding agency here. If none, delete this.

II. Goals and Objectives

A. Motivation

Diabetes will increases the chances of the severe health issues possessing like life threating risks, and this associated complication will significantly endanger the lives of the person individually. By conducting the various analysis on the available data and aim to provide appropriate solution to the problem which decreases the problems. aim to perform the implementation to finding out the risk factor and its accuracy to improve their rate.

B. Significance

This significance of the diabetes is in the widespread in the society which impact the health . And diabetes is the long – term condition and it will impacts how the body will handles the sugar in the blood , if it is not properly controlled then it will cause many different health problems .

Example : suppose a person has the family history of the diabetes , has high sugar and has the low physical activity and overweighted based on the calculation on the BMI . And this study will implies how the person will affect with diabetes or how it gets worse if the person already haves it. By concluding the family history like the physical activities mentioned in the above lines . The study will draw the conclusion about their impact on chance of developing diabetes.

C. Objectives

It is used to identify the risk factor of diabetes based upon using some model like Statistical Method and machine learning methods. The based upon the given data and the finding out the accuracy so that we can identify the early which might get highly risk individuals. The implementation of the each model is plotted in the graph as we can see in the below.

D. Features

Diabetes will increases the chances of the severe health issues possessing like life threating risks, and this associated complication will significantly endanger the lives of the person individually. By conducting the various analysis on the available data and aim to provide appropriate solution to the problem which decreases the problems. aim to perform the implementation to finding out the risk factor and its accuracy to improve their rate.

E. Reporting

In this study we focused about the diabetes and our primary aim is to analyse various data driven methodologies to understand the complexity of the disease . We exactly processed and provide the dataset and encompossing the features like glucose levels, pregnancies, blood pressure, skin thickness, insulin levels, BMI, age, and diabetes pedigree function. Take on the range of the machine learning models such as logistic regression and decision tree . We evaluate the predicting of the diabetes onset. And the our findings will be highlighted glucose levels, BMI, and age as pivotal factors in predicting diabetes. And the showcasing of the model accuracy exceeding 85To identify the crucial things . We have plan to share our findings through the documentation and aiming to reach to improve the management strategies of the diabetes.

F. Improvement

In our project we have taken dataset and performed various EDA techniques like data cleaning, data preprocessing, data visualization and analysis methods like statistical inference , univariate analysis, multivariate analysis and In our next increment we will perform various machine learning models on our dataset features to predict whether that particular person will get diabetes and not we compare all those model and will show the model that has the best accuracy score.

III. BACKGROUND WORK

[1] Together, the cited studies represent a sizable body of work in the field of diabetes prediction, demonstrating a wide range of techniques and strategies. Perveen et al. (2016) conducted a thorough analysis of data mining classification algorithms, clarifying the finer points of algorithmic performance, such as support vector machines' adaptability and decision trees' resilience. Kumari et al. (2021) add to the discussion with an ensemble method that combines the predictive power of many models using a soft voting classifier. This demonstrates how different algorithms can be combined to create potential synergies that could lead to increased accuracy in the prediction of diabetes mellitus. [2]Han et al. (2018) investigate the interpretability of diabetes prediction models, particularly decision tree-based models. Their study emphasizes the value of models that provide accurate prediction along with lucid insights into the decision-making process. However, Tafa et al. (2015) offer an intelligent system created especially for diabetes prediction, emphasizing not just the system's predictive capabilities but also its all-encompassing architecture and features that increase its efficacy. This all-encompassing perspective provides a roadmap for combining various components, boosting the prediction model's overall effectiveness. [3]Sisodia and Sisodia (2018), who investigate a range of classification algorithms and offer a full understanding of how they might be applied to the difficult issue of diabetes prediction, eventually broaden the study's scope. Based on the characteristics of their datasets and the desired outcomes, practitioners and researchers can utilize their study, which provides relevant information on the benefits and drawbacks of various algorithms, to identify the most effective method for diabetes prediction. Together, these articles offer a range of perspectives that advance the subject of diabetes prediction and offer guidance for next research projects and applications in the fields of predictive analytics and healthcare.

IV. DATASET

Name : Statistical Insights into Diabetes Risk Factors and Their Implications.
. Number of features: 9
. Number of rows : 2000
This dataset is downloaded from the Kaggle and this dataset contains the information about the patients . This dataset called diabetes contains the patients information and this dataset includes the columns such as pregnancies , glucose , Blood pressure , skin thickness , insulin , BMI , Diabetes pedigree function , Age and outcome . And this each columns implies about the details about the patient. This dataset has provides the information regarding the glucose level , this glucose provides the information concentration of glucose in the blood and this increasing glucose level indicates the diabetes or the risk of developing the diabetes.
Pregnancies(this variable implies the number pregnancies the person has had .According to some study the difference between the number of pregnancies and the risk taking of developing the diabetes) . Blood pressure(This significance refers the blood exerted by the blood on the walls of the blood vessels and this high blood pressure can lead to the diabetes. Skin thickness (This refers to thickness that skinfolds and it is commonly measures at the various bodies . And in some cases if the skin thickness is increases and it is associated with the insulin resistance
And coming to next step insulin it implies that the amount of insulin present in the bloodstream . And the low insulin and the insulin resistance develop the diabetes . BMI (body mass index) this is the process that will calculated based on the body weight and height and higher BMI will increases the

risk of the type 2 diabetes. Diabetes pedigree function this process provides the information about the hereditary risk of the diabetes that will calculated by using the family history and genetics.
Age (the age of the individual person in the dataset and this significant risk in the diabetes as the chances of developing the conditions increasing with the age). And last coming to the outcome this implies indicates the presence or absence of the diabetes.

V. DETAILED DESIGN OF THE FEATURES

A. Feature Selection

In this step we remove features that are not necessary or have no positive impact on our output result. This step also improves model accuracy as there are less features overall and also helps reduce overfitting where a model has good accuracy on well trained data and worse on new data. In our project out of 9 features we have we dropped the 'skin Thickness' column. The reason is the thickness of skin has zero impact on diabetes.
Feature Engineering In order to improve readability and for faster coding we have also renamed the column of Diabetes Pedigree function to DPF. This column contains a value of a function that predicts the chance of a person inheriting this disease from his family members. But this is not the only factor that influence the occurance of diabetes as there are many. Feature engineering also helps in creating new columns that are a combination of existing columns .

B. Analysis

In this section we have describes each column of the data separately using describe() function which gives us values like minimum values, maximum value, mean , std dev and the three quartiles. Using this data we can see where the most data lies and their median . We can also perform some preprocessing techniques for data points that doesn't make sense like having BMI of zero which is not possible.
So we could replace values with better values like median of the data. In this step we also remove outliers in our data for more accurate results and to reduce overfitting
We visualize the data so we can understand the distribution of data using distribution plot that is nothing but a histogram that divides the data into chunks and draws a KDE curve that is used to better understand the smooth distribution of data . This plot tells us the density of the data in which interval most of the data lies .
Another visualization used is a box plot that is used for understanding outliers in our data. This plot consists of box divided into two part with first part representing the first quartile and second box representing the third quartile and line in the middle is 2nd quartile or median of the data and the lines to the left of the box is the minimum value and to right tells us the maximum value . While points outside the whisker tells us the outliers.
To remove the outliers we have used the standard based function which calculates the mean and standard dev of the

```
[191] df['Insulin'].describe()
      # provide the description of the given column like mean, 3 quartiles, min and max values

count    2000.000000
mean       80.254000
std       111.180534
min         0.000000
25%         0.000000
50%        40.000000
75%       130.000000
max       744.000000
Name: Insulin, dtype: float64
```

Fig. 1. Describe about insulin

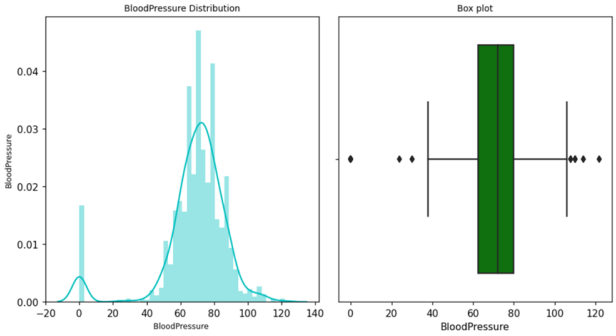


Fig. 2. Box plot and Insulin Distribution

column and we have given the 3*std dev as the cutoff value . we calculate the upper and lower bounds for this and we allow values that are only in between these bounds and remove all other values in the column as we consider them as outliers Here is what we did for one of the column 'Insulin'
Here's the visualization of the column before replacing the zeroes with median and before removing null values.
The visualization below is after performing operations like replacing zeroes and removing null values.

we calculate the upper and lower bounds for this and we allow values that are only in between these bounds and remove all other values in the column as we consider them as outliers.

The rest of the visualization are in the code file.

Inferential statistical analysis In this step we have performed the correlation matrix that give us correlation of each pair of the columns in our dataset. This step also splits the data

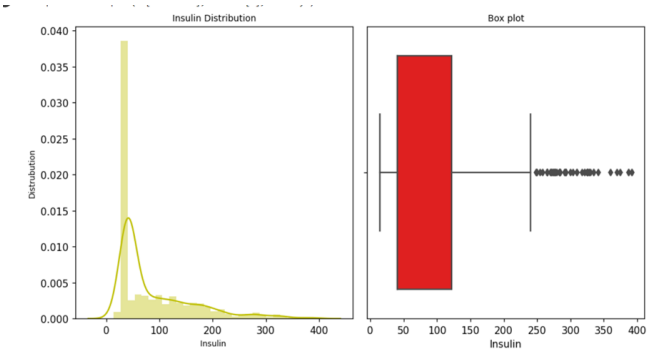


Fig. 3. Performing EDA For Insulin

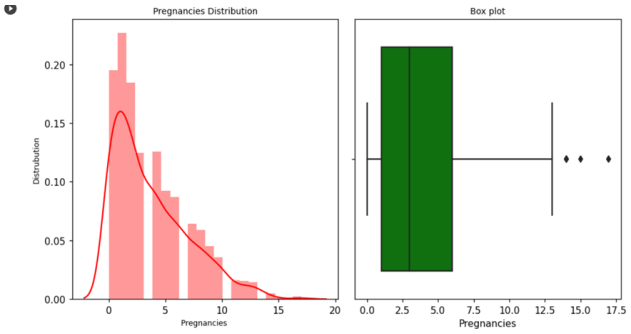


Fig. 4. Before Pre-Processing for Pregnancies Distribution, Box.

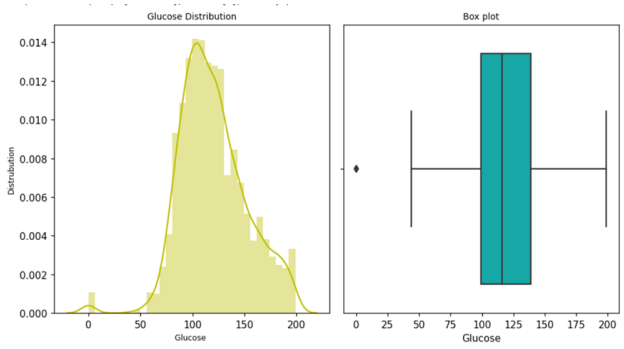


Fig. 5. After Pre=Processing For Pregnancies Distribution, Box.

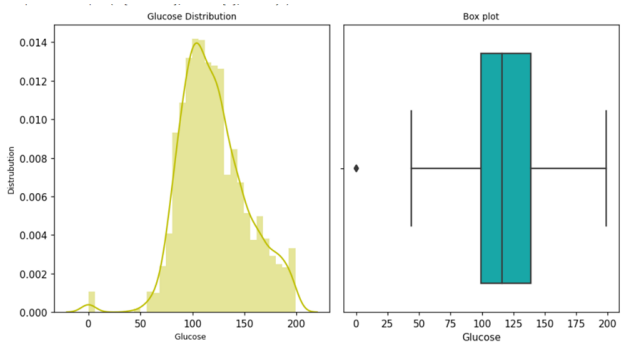


Fig. 6. Before Pre-Processing for Glucose

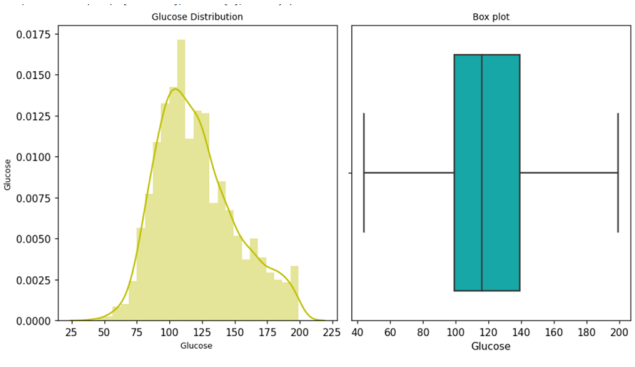


Fig. 7. After Pre-Processing For Glucose.

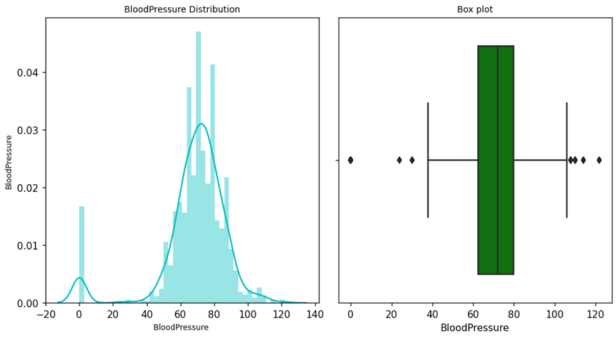


Fig. 8. Representing BP Distribution and Box plot

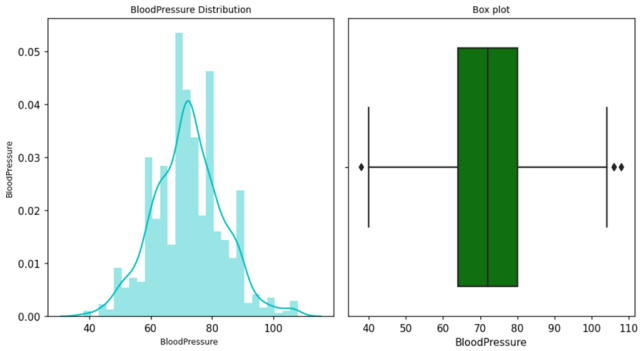


Fig. 9. After Pre-Processing

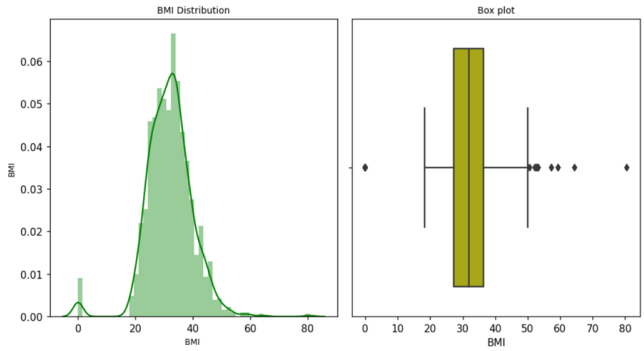


Fig. 10. BMI Distribution and Box Plot

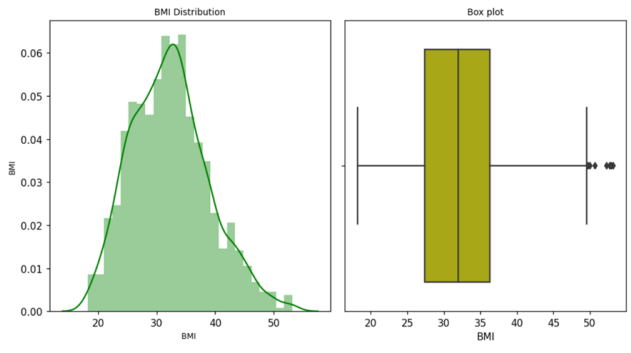


Fig. 11. After Pre-Processing

```
Correlation Matrix:
                Pregnancies   Glucose  BloodPressure   Insulin       BMI  \
Pregnancies        1.000000  0.123720       0.206999 -0.059094  0.030017
Glucose            0.123720  1.000000       0.186335  0.331657  0.199520
BloodPressure      0.206999  0.186335       1.000000 -0.048198  0.282666
Insulin           -0.059094  0.331657      -0.048198  1.000000  0.200951
BMI                0.030017  0.199520       0.282666  0.200951  1.000000
DPF                0.020666  0.072307       0.025924  0.188449  0.123485
Age                0.543612  0.284326       0.348239 -0.030592  0.056780
Outcome            0.217060  0.493214       0.189646  0.126631  0.284074

                     DPF       Age   Outcome
Pregnancies     0.020666  0.543612  0.217060
Glucose         0.072307  0.284326  0.493214
BloodPressure   0.025924  0.348239  0.189646
Insulin         0.188449 -0.030592  0.126631
BMI             0.123485  0.056780  0.284074
DPF             1.000000  0.057485  0.186170
Age             0.057485  1.000000  0.259839
Outcome         0.186170  0.259839  1.000000

T-test for Glucose between Outcome groups:
T-statistic: -24.465374889282756
P-value: 7.455723690693177e-115

ANOVA for Age between Outcome groups:
F-statistic: 134.8176380806184
P-value: 3.846378295336311e-30
```

Fig. 12. Correlation Matrix

into two columns based on the outcome of diabetes. We also perform sample t test for the column glucose column between two possible outcomes that tells us how much does that column has an impact on the data outcome. The output of this is the T-statistic value and p value. A larger T-statistic value of either positive or negative means there is great difference between the mean of variables. P-value is probability of getting that result if the null hypothesis is true. Next we perform ANOVA Test on age column and outcome variable. This test tells us if there is a difference between mean of ages and outcomes. The output of this is F-statistic that tells if there is difference between mean of age and outcome variable and p-value is the probability of getting such high value if the null hypothesis is true..

Univariate Analysis This is more focused on how each variable performs individually and how data is distributed in that variable and also calculate values like minimum, maximum, 25

Multivariate Analysis In this we analyze relations between many variables simultaneously. In our project we fit a multiple regression model to our dataset. We get an output and it helps to understand how each independent variable relates to the probability of diabetes occurrence, considering the effect of other variables in the model. The coefficients' significance, odds ratios, and confidence intervals are essential in understanding the impact of each variable on the outcome.

This first visualization creates a horizontal bar plot of the coefficients we got from the logistic regression model. It helps visualize the impact that can be either positive or negative of each predictor variable on the odds of the outcome variable.

In the above graph we show the odds ratio with the confidence intervals. This ratio tells us how chance of outcome variable changes with change in the predictor variable . With the help of both these visualizations we get better understanding of the logistic regression output.a

We also performed PCA on this dataset which is used to convert higher dimension data to lower dimension data and identifying patterns between the data . We calculated two

From the above T-test result of the glucose variable T-statistic of -24 means there is significant difference between mean of glucose and outcome variable and the low p-value says that there is strong evidence against null hypothesis. And ANOVA test result says that F-statistic value of 134 is too high that there is a significant difference between mean of age and outcome variables. Small p-value of 3.84 e-30 says that there is high evidence to reject null hypothesis.

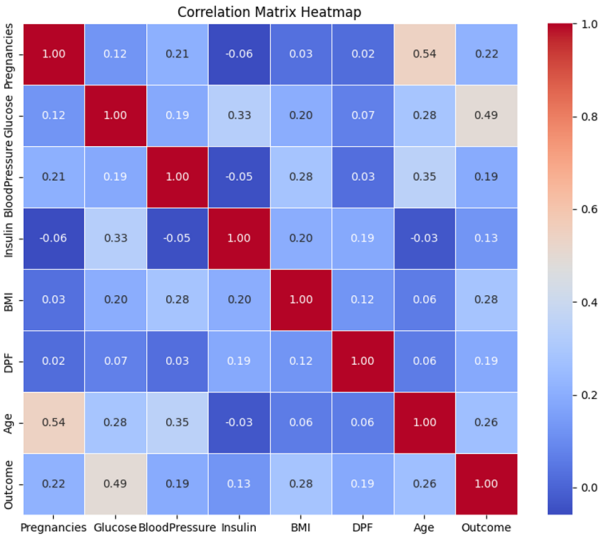


Fig. 13. HeatMap

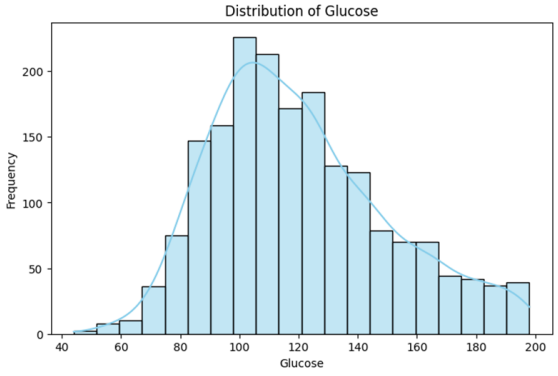


Fig. 14. Glucose Distribution

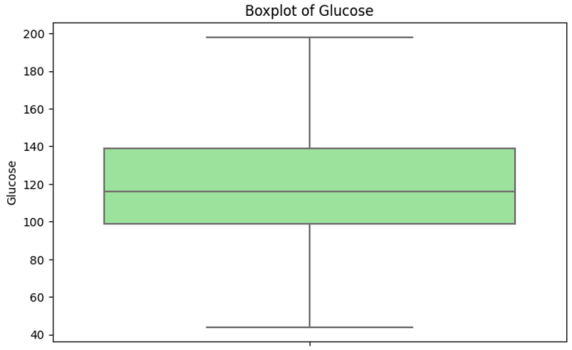


Fig. 15. Box Plot for glucose

```
Descriptive Statistics for 'Insulin':
count    1864.000000
mean       88.582618
std        73.064201
min        15.000000
25%        40.000000
50%        40.000000
75%       122.000000
max       392.000000
Name: Insulin, dtype: float64
```

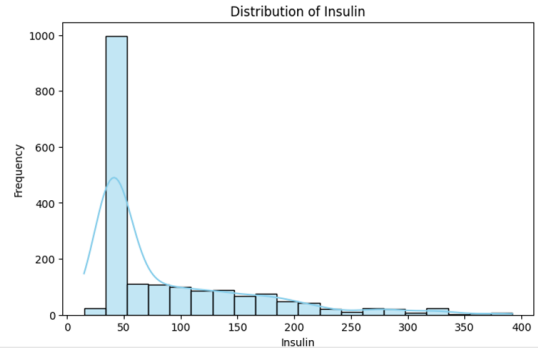


Fig. 16. Distribution of insulin

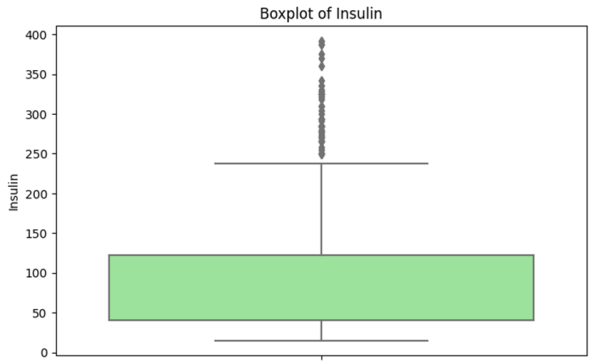
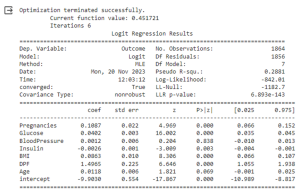


Fig. 17. After Pre-Processing Insulin



```
Optimization terminated successfully.
         Current function value: 0.451721
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                 1864
Model:                          Logit   Df Residuals:                     1856
Method:                           MLE   Df Model:                            7
Date:                Mon, 20 Nov 2023   Pseudo R-squ.:                 0.2881
Time:                        12:03:12   Log-Likelihood:                -842.01
converged:                       True   LL-Null:                       -1182.7
Covariance Type:            nonrobust   LLR p-value:                6.859e-143
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Pregnancies     0.1087      0.022      4.969      0.000       0.066       0.152
Glucose         0.0402      0.003     10.002      0.000       0.035       0.045
BloodPressure  -0.0012      0.006     -0.204      0.838      -0.010       0.013
Insulin        -0.0026      0.001     -3.009      0.003      -0.004      -0.001
BMI             0.0863      0.010      8.306      0.000       0.066       0.107
DPF             1.4065      0.225      6.646      0.000       1.055       1.758
Age             0.0118      0.006      1.821      0.069      -0.001       0.025
Intercept      -9.0030      0.554    -17.867      0.000     -10.089      -8.017
==============================================================================
```

Fig. 18. Multivariate Analysis

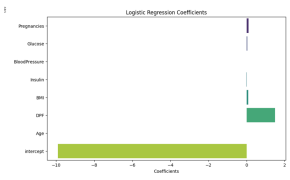


Fig. 19. Logistic Regression

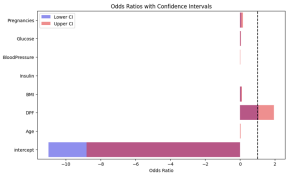
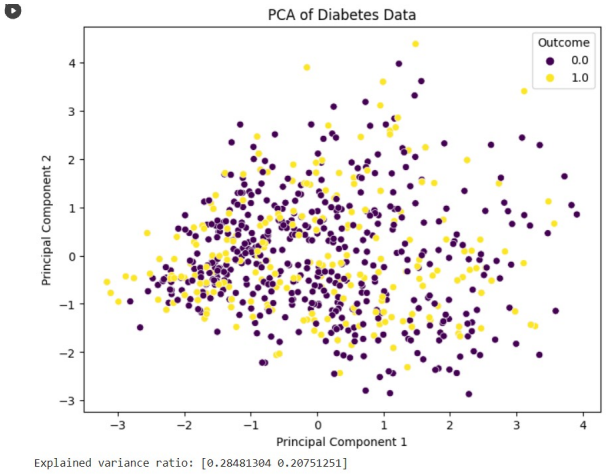


Fig. 20. Coefficient Intervals




```
Explained variance ratio: [0.28481304 0.20751251]
```

Fig. 21. PCA of Diabetes

variables PC1 and PC2 which are new variables created from original variables. PC1 is the variable with most dominant structure in the data and it has the highest variance compare to other principle component variables and PC2 has the second highest variance. The plot below is the scatterplot of two variable with outcome variable used to differentiate them by color. The output variance is the variance of selected principle component variables.

VI. IMPLEMENTATION AND PRELIMINARY RESULTS

VII. PROJECT MANAGEMENT

C. Work completed:

The Data is cleaned and remove the null values and taking the exploratory Data Analysis and performing the various Statistical methods and Machine learning techniques have been implemented.

D. Responsibility:

Nagateja Ravilla: Data collection,cleaning and pre-processing. Jaswanth Mandadi: Box plots, Distribution plots during outlier removal. Brahma Teja Mule: Statistical Inferences, Univariate analysis and their visualizations Mounish Seeni: Multivariate analysis and their visualizations.

E. Contributions:

Nagateja Ravilla-25percentage, Jaswanth Srivatsav Mandadi-25 percentage, Brahma Teja Mule-25 percentage, Mounish Seeni:25 percentage

F. Work to be completed

Teja : Decision Tree Jaswanth : SVM, Random Forest Brahma Teja : XGBoost Mounish Seeni : KNN,Performance evaluation methods

Issues : We are yet to implement these models and have to find out the best performing one among them

VIII. REFERENCES

[1]. Perveen S., Shahbaz M., Guergachi A., Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science . 2016;82:115–121. doi: 10.1016/j.procs.2016.04.016. [CrossRef] [Google Scholar] [2]. Kumari S., Kumar D., Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering . 2021;2:40–46. doi: 10.1016/j.ijcce.2021.01.001. [CrossRef] [Google Scholar] [3]. Han J., Rodriguez J. C., Beheshti M. Discovering Decision Tree-Based Diabetes Prediction Model. Proceedings of the International Conference on Advanced Software Engineering and its Applications; December 2018; Jeju Island, Korea. Springer; pp. 99–109. [Google Scholar] [4]. Tafa Z., Pervetica N., Karahoda B. An Intelligent System for Diabetes Prediction. IEEE Explore ; Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO); June 2015; Budva, Montenegro. pp. 378–382. [Google Scholar] [5]. Sisodia D., Sisodia D. S. Prediction of diabetes using classification algorithms. Procedia Computer Science . 2018;132:1578–1585. doi: 10.1016/j.procs.2018.05.122. [CrossRef] [Google Scholar]

https://github.com/mandadi1999/emp$_proj$