Analyzing Diabetes Risk Factors through Statistical Methods

|  |  |  |  |
| --- | --- | --- | --- |
| Supriya R | Mounika D | Rishika K | Charan V |
| 11644767 | 11652326 | 11699010 | 11718112 |

***Abstract*—** **Diabetes diagnosis can be a complex medical procedure. is one of the most difficult medical diagnoses. The World Health Organization says that the patient population grows each year. In this project, we are calculating the ratio of diabetics to non-diabetics in the dataset and, we build 95% interval of confidence for individuals with diabetes and the next step we will be performing inferential statistical analysis, which will be assisted by graphical representations. In the end, we will employ by using algorithms of machine learning, like** **KNN, Decision Tree, SVC, XGBoost Classifier and linear regression. Additionally, we will be training the model by suggesting the model's coefficients and intercept.**

1. Introduction

The worldwide prevalence of diabetes, which is a long-term metabolic condition marked by high blood glucose levels, has increased dramatically. Knowing how the many risk variables interact is critical to developing efficient preventative and therapeutic methods as the incidence of diabetes rises. This research uses strong statistical techniques to pinpoint and examine important diabetes hazards to explore the intricate web of these variables.

Diabetes is a complex disease that is impacted by numerous ecological, habits, it takes a methodical strategy to determine the corresponding significance of various aspects, as well as statistical approaches offer an effective toolkit for these kinds of studies. This study aims to provide significant insight into the complex links between different risk factors with the prevalence of diabetes through utilizing data-driven analysis.

The results of this study have a chance to improve risk evaluation instruments, direct tailored actions, and provide guidance regarding public health campaigns. Healthcare practitioners may establish focused plans for lifestyle changes

safeguards by identifying risk factors that can be improved.

Furthermore, by comprehending the genetic basis of diabetes danger, medical advances and increasingly individualized therapies are possible for those who are prone to this disease.

The importance of this study is in its ability to provide a thorough description of the variables affecting diabetes risk, rather than merely a summary. Our goal is to facilitate well-informed choices for physicians, officials, and individuals by means of statistical research, so adding to the ongoing conversation on preventative healthcare. We hope to shed light on potential avenues for a more thorough and efficient strategy to address this widespread worldwide health issue as we set out to investigate diabetes risk factors.

To understand the intricate relationships that lead to the development of diabetes, this project sets off on a statistical voyage. Our goal is to identify patterns and relationships as we examine the abundance of accessible data, offering a more comprehensive understanding of the intricate dynamics of diabetes risk. In this endeavor, the study turns into an important first step towards a better educated and practical strategy for reducing the worldwide diabetes pandemic.

II. GOALS AND OBJECTIVES

*A. Motivation*

Diabetes causes people to develop various illnesses, putting their health in peril. Therefore, to solve this dilemma, we will do various analyses of available data and provide relevant solutions to this challenge, which will lower its intensity. Furthermore, the knowledge derived from statistical studies is essential for preventing illness. Comprehending the importance of various risk factors facilitates the creation of efficacious preventative interventions. This could involve putting prevention programmers, teaching promotions, including behavioral change measures into action. Scientists and physicians can proactively treat the fundamental causes of diabetes by identifying its primary drivers, which will lower the disease's overall occurrence.

Addressing a major worldwide health concern is the driving force behind the statistical analysis of diabetes indicators. Scientists and physicians can advance into a more individualized and successful strategy for diabetes controlling and preventing it by utilizing statistical data to design targeted efforts and effectively deploy assets. Statistical tools provide a practical way to optimize healthcare assets in setting economic distribution of resources.

1. *Significance*

The project’s significance is to show various types of features that determine diabetes in each human. For example, a person possesses low glucose levels, low blood pressure, and high skin thickness. Hence these overall studies give us conclusion. Statistical tools provide a practical way to optimize healthcare assets in setting economic distribution of resources. Healthcare systems can better distribute resources and prioritize activities through determining what are the most relevant risk variables. This ensures that treatments are focused where they are most required, which is especially important in environments with scarce funds.

1. *Objectives*

The objective for this project is to determine the causes influencing diabetics through statistical evaluation and AI techniques. This project intends to identify associations, trends, and relationships between numerous risk variables using sophisticated statistical analysis, offering important information in the formulation of medical policies, focused treatments, including prevention strategies.

1. *Features*

Diabetes has been analyzed by using different statistical

analysis like Inferential Statistical Analysis, Univariate Statistical Analysis, Multivariate Statistical Analysis, and different machine learning algorithms by using the columns present in the dataset. These characteristics are crucial parts of the analysis-related mathematical models.

Information accessibility, credibility in biology, and the amount of current research ought to all be taken into consideration when choosing features. The study precise as well as detailed examination of risk variables for diabetes may be achieved by utilizing a broad collection of characteristics.

Lifestyle factors, such as food and exercise routines, represent the impact of individual decisions in their own unique way. Going further, healthcare history presents sections in an account that explain the relationship between health indicators.

III. Background Work

[1] This paper predicts early forecasting in women employing various techniques for machine learning including as K-Nearest Neighbor, a Random Forest, and Decision Tree.

[2] This work describes the development of a diabetic forecasting system utilizing decision trees. Han, Rodriguez, and Behesti performed the research, which strengthens the growing field of sophisticated programming within medical devices through offering details about the creation and effectiveness of statistical models based on trees for diabetes diagnosis.

[3] With this article, Kumari, Kumar, and Mittal suggest a combined technique for obesity detection and prediction using an approximate voting classifier. The research project advances computational thinking for medical purposes by examining the effectiveness of ensemble approaches in increasing disease prediction correctness.

IV. DATA SET

* Name of the dataset: Diabetes dataset
* Number of columns:9
* Number of rows:768

<https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data>

The data that was used in this dataset was provided by the National Institute for Diabetes and Digestive and Kidney Diseases. The goal of this data set aims to figure out whether a person develops diabetes based on specific diagnostic criteria contained during its dataset. Requirements were employed for choosing such instances within a larger database. Each of those receiving treatment at this facility were Pima Indian women above the age of 21. Pregnancy, which plasma glucose levels during a 2-hour tolerance for oral glucose examination, systolic blood pressure, triceps folds on the skin its thickness, 2-hourly rate plasma insulin production, the body mass index, diabetes genealogy operation, and age are the 8 characteristics that comprise this data set. The initial 786 diabetic information in this set has been decreased to 392 when data that is absent has been removed. This (.csv) document contains quite a few elements, a few of those that are independent (many healthcare prediction elements) and just a single variable that is dependent (Results).

The dataset was first composed of 786 diabetes-related cases; records with values lacking were eliminated and the final count was 392. The information is contained in a.csv file and includes a number for distinct variables that are pertinent to healthcare forecasting in addition to a single dependent variable called "Results," which is the focus of the research's focus on the onset developing diabetes. The careful management of incorrect information demonstrates a dedication to accuracy and guarantees the dataset's dependability in guiding important medical choices. This dataset is a living asset that will continue to provide chances for improving our understanding of diabetes and developing targeted solutions as technological and scientific advances.

The dataset is a living example for the growing relationship between information and medical services, showing that techniques based on data may advance the field of medicine.

V. DETAILED DESIGN OF THE FEATURES

*A. Analysis*

1. Exploratory Data Analysis: EDA is a method that examines the given dataset to characterize the attributes and display visually. The phase is pivotal specifically for machine learning model development. EDA involves the creation of different types of plots like histograms, box plots and many more. Generally, data exploration takes more time, but it plays an important role in formulating the problem statement or articulating the definition for our data collection, which holds paramount significance.

2. Inferential Statistical Analysis: Inferential statistical analysis generates predictions and inferences regarding the population using an accurate representation of data given by individuals. After extending a vast dataset, it employs probabilistic to arrive at an inference. Its sole purpose is to communicate the value of analysis of statistics. Its sole objective is to research, analyze, and then draw conclusions. Inference statistics are primarily related to testing hypotheses, the primary objective that aims to reject the idea of a null. Hypothetical testing involves an inference procedure which analyses and evaluates the reliability of a group's premise using sample data. Inferential statistics are frequently used to evaluate the strength of an association among a set of data. Still, compiling an individual list and selecting a sample at random is time-consuming.

3. Univariate Statistical Analysis: There is the sole statistic in this data collection. Since the knowledge only must do on a single variable that changes, the assessment of single-variate data constitutes the quickest sort of review. It is not concerned with origins or interactions, as the primary goal of the investigation is to characterize the data and identify themes throughout it. When examining connections in the context of multivariate analyses, it is essential to grasp the underlying characteristics of each factor through univariate testing.

4. Multivariate Statistical Analysis: Multifunctional study of data is a class of prediction approaches which investigate trends within data that are multifaceted by taking many different characteristics into account with the correct sequence. This method is an extension from bidirectional research, using just two variables for its analyses. Because multivariate models can allow for additional factors, we can study a broader variety of features that uncover trends in datasets which tend to be more related to reality.

5. KNN: K-Nearest Neighbors represents a straightforward yet critical categorization method for the field of machine learning. This comprises a supervised instruction technique utilized for identification of patterns, data extraction, as well as detection of intrusions. Although KNN is simple to operate and doesn't need an initial training phase, it may be dependent on redundant or noisy data.

6. Decision Tree Classifier: A decision tree is a highly practical or often employed forecasting or classification technique. A decision tree is an organized design typically resembles a schematic, having every inner node conveying an experiment upon a component, all branches displaying a testing conclusion, then each leaf node (last node) keeping a class name. They have been used in many different machine learning scenarios because of their popularity, comprehension, and capacity for handling data that is numerical as well as categorical.

7. SVC: SVC is a supervised learning technique that may be utilized for categorization or regression. It seeks a hyperplane with the greatest possible margin across classes. Support vectors are important points that influence the surface of the hyperplane. Employing the kernel's method, SVM is capable of handling non-linear sequences. This parameter strikes an equilibrium between accurate categorization with margins width. It is commonly used in picture or text classification. To greatest efficiency, kernel functions and parameters need to be carefully tuned.

8. Feature Selection: The ease of installation and implementation of Recursive Feature Elimination (RFE) renders it an effective method during choosing features. When forecasting the desired factor, it does exceptionally well at determining which training dataset factors (columns) were most important. The quantity of characteristics that are chosen as well as the method for selecting features constitute two important design choices that must be made throughout RFE deployment. It is not necessary for either one of the input values to be specified exactly for the technique to be effective, even though they may both be evaluated.

9. Voting Classifier: Through the integration of multiple designs, collaborative learning improves the goals of machine learning. It makes use of the combined predictive ability of several models as opposed to depending only on one. The basic idea is to assemble a group of classification specialists and enable them to vote on each other's contributions as a group. Voting classifiers are useful to feed a variety of classification and regression assignments, and they frequently improve total model efficacy by utilizing the advantages of various base systems. The collective setup and initial classifier selection are important factors to consider in applications in real life. Additionally, their accuracy frequently outperforms that of separate models.

10. Random Forest Classifier: Utilizing decision trees for different purposes such as classification and regression, the Random Forest method, also called the Random Decision Forest, the supervised machine learning technique. A randomly chosen part of the training data set is used by the Random Forest classifier to create a succession decision tree. To establish the correct forecast, it basically consists of an assortment independent decision tree (DT) selected free arbitrary across the data set. The responses from these DT are subsequently combined. Voting classifiers are useful to feed a variety of classification and regression assignments, and they frequently improve total model efficacy by utilizing the advantages of various base systems.

11. XGBoost Classifier: Using this strategy, decision trees are created in a predetermined order. The use of loads in XGBoost is crucial. The use of weights is assigned to each independent factor and included in a choice tree to forecast outcomes. Qualities that the decision tree mis predicted have a higher weight, so these factors are subsequently included onto the subsequent decision tree. After that time, these many classifiers/predictors are put together to produce an algorithm that is more accurate as well as efficient. The method's exceptional results and accomplishments in numerous machine learning contests are credited with its wide appeal.

VI. IMPLEMENTATION AND PRELIMINARY RESULTS

*A. Data cleaning:* It originated in the medical services sector and is made up of 768 instances containing 9 characteristics. The variable that determines the result represents the goal variable. It doesn't seem to have an error in the data, since the data types offered were Boolean operators float, and integer. Now descriptive research indicates thus the parameters Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI possess a minimal value of 0 which is illogical; these numbers may be incorrect or anomalies however I'm not going to alter any until I can investigate true data statistics. Regarding the Pregnancies section, I saw in the lowest number is 0 (possibly indicating not a pregnancy), which is quite noteworthy, though the maximum time for pregnancy is 17, making it irrelevant. The separation between each of the multiple response factors. further exploration of the data should involve investigating any missing values or outliers that might impact the overall integrity of the dataset. Conducting summary statistics for each variable could reveal patterns or anomalies that require attention. Additionally, examining the distribution of the target variable across the dataset is crucial for understanding the balance or imbalance in the outcome classes, which is essential for building predictive models. Correlations or

interdependence within variables may be found by examining

correlation matrix based on their characteristics; this could reveal potential crossover challenges that might hinder the accuracy of the correlation matrix based on their characteristics; this could reveal potential crossover challenges that might hinder the accuracy of prediction models.

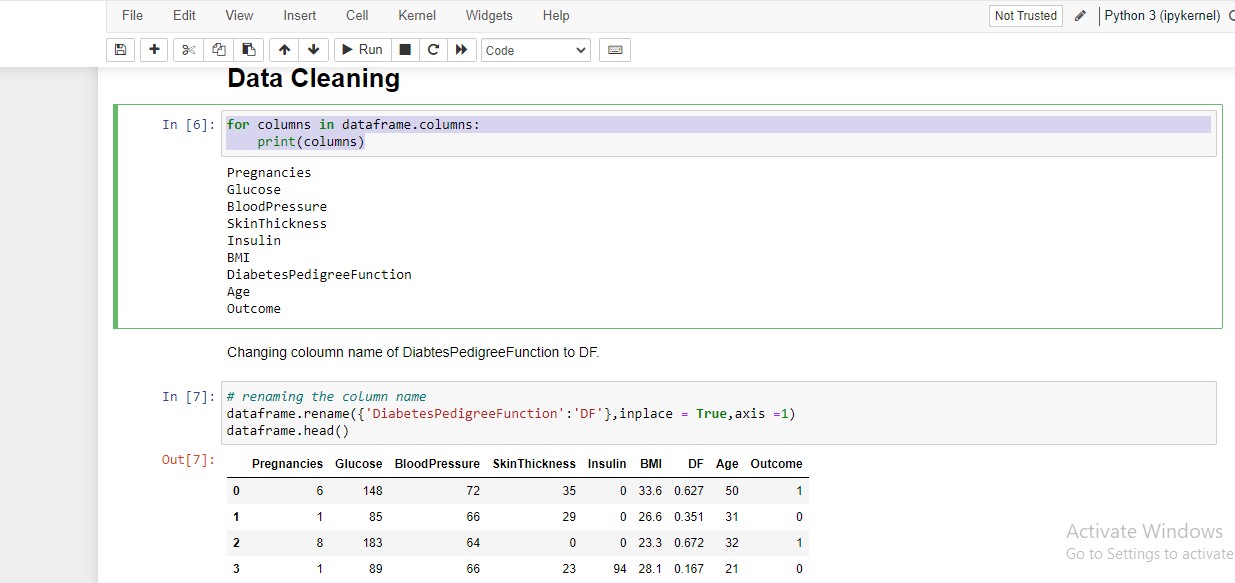


Fig. 1. Data cleaning

First, we need to perform data cleaning so after successfully performing then we need to proceed with data preprocessing.

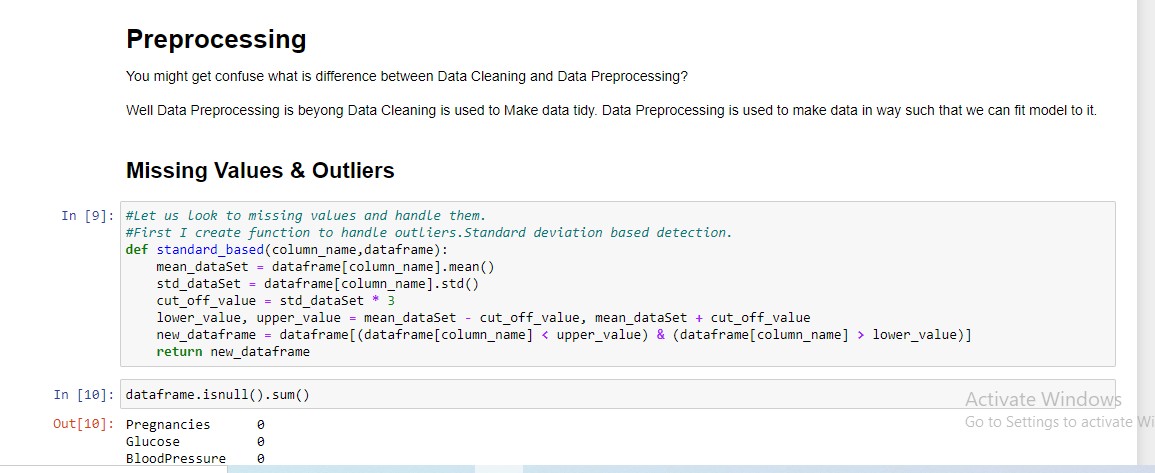


Fig. 2. Data Preprocessing

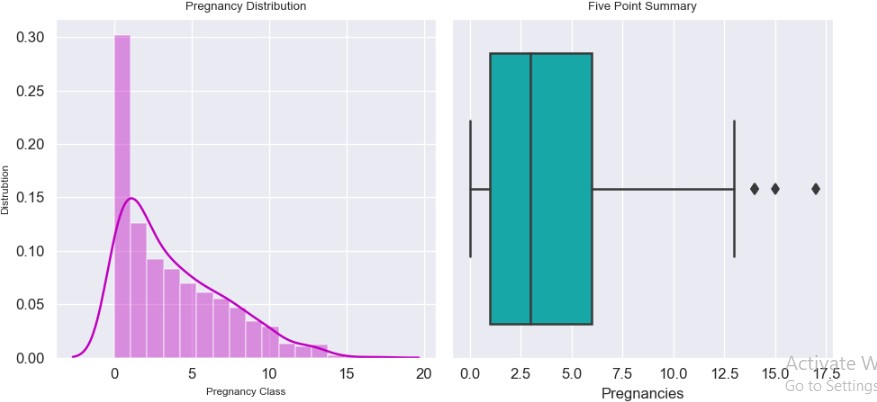
A few of these values were excessive for the count in pregnancy. Therefore, we'll display the pregnancy frequency visualization.

Fig. 3. Outlier for pregnancy

In this step, we are now including the outlier for pregnancy

frequency.

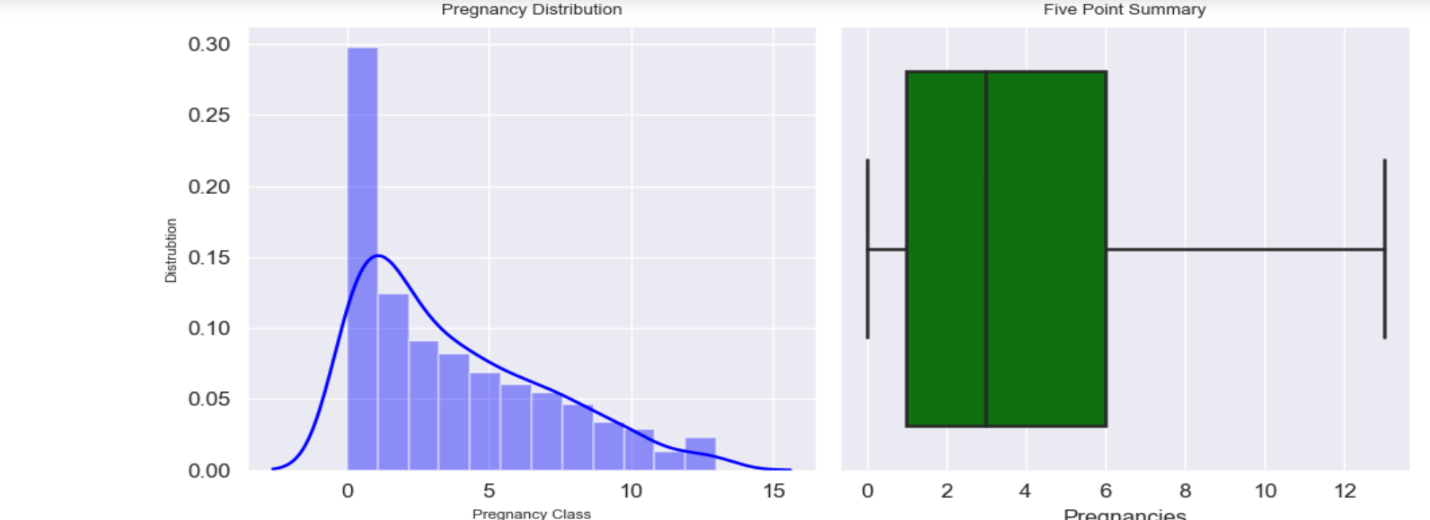


Fig. 4. Pregnancy class

.

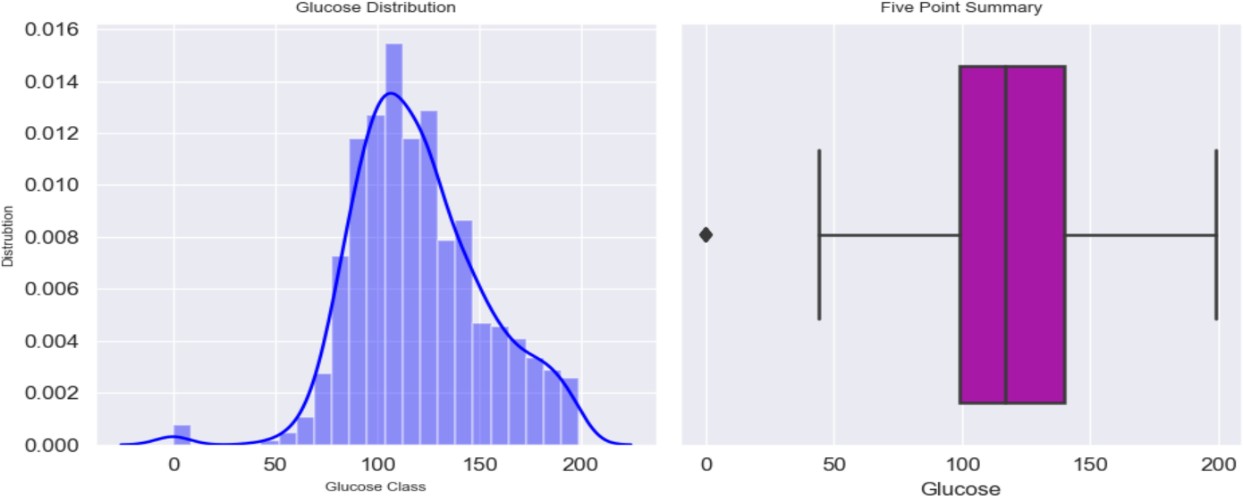


Fig. 5. Glucose class

A few among the glucose levels in the provided plot are 0, this doesn't make any reason. As a result, we're substituting 0 by the mean value.

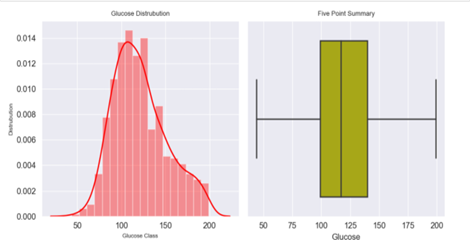


Fig. 6. Blood sugar class

The graphic shown depicts the blood sugar range once data have been replaced by the mean.

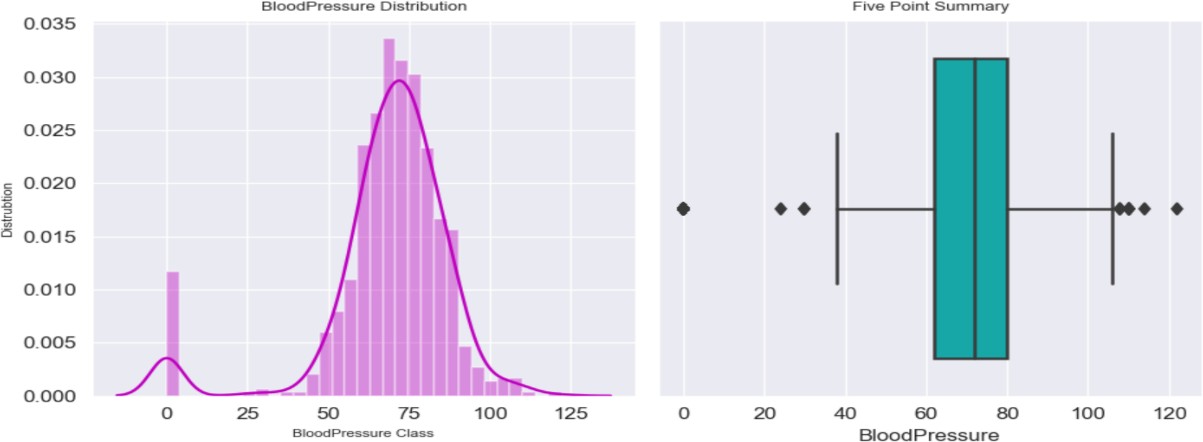


Fig. 7. Outlier for Blood Pressure

Since the figures ranging from 0 to 25 in the preceding Blood-pressure outlier doesn't make sense, we will substitute them by the median.

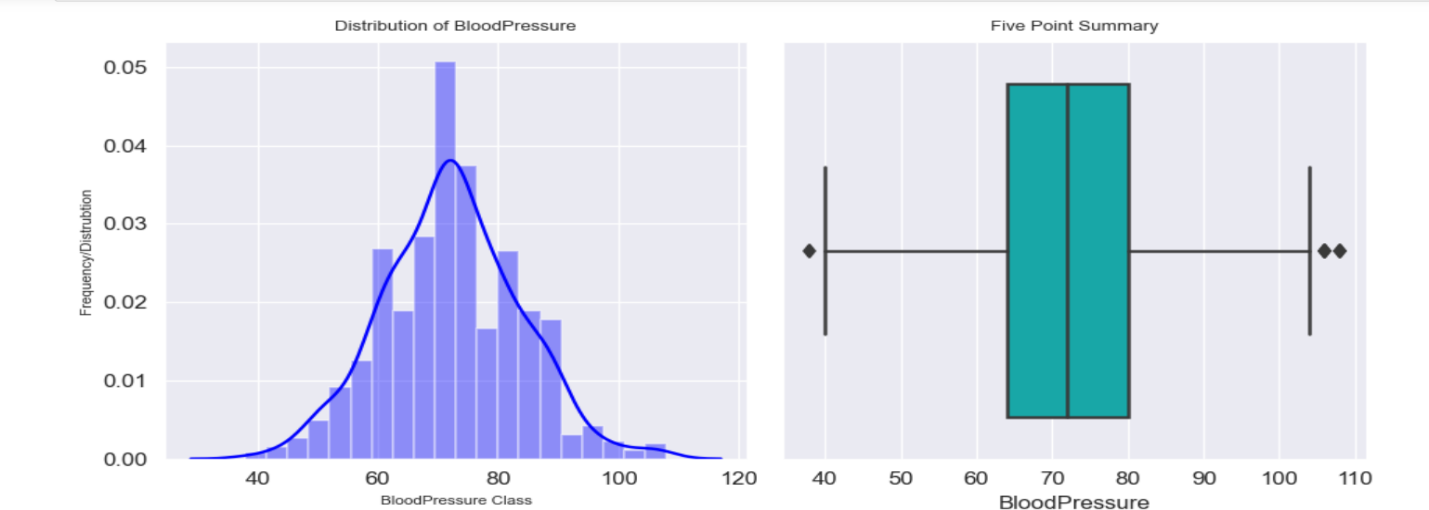


Fig. 8. Blood Pressure class

The graph will appear like this following updating the data that is missing by the median. We are currently starting with the statistical evaluation.

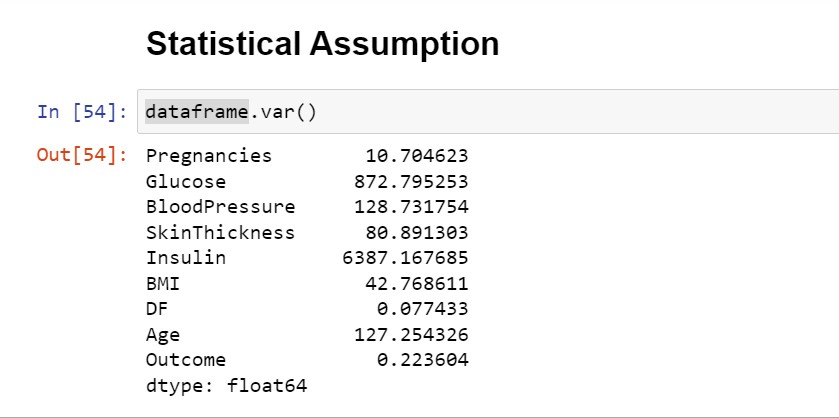


Fig. 9. Statistical Assumption

We are removing the DF category as the variance is quite minimal.

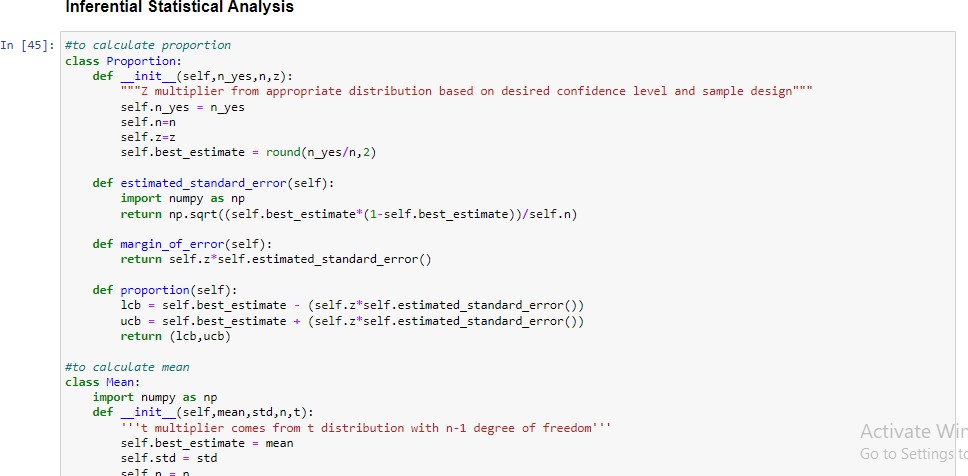


Fig. 10. Inferential Statistical Analysis

We are calculating the entire diabetes score using the given dataset. We are currently determining its inferential statistical analysis results.

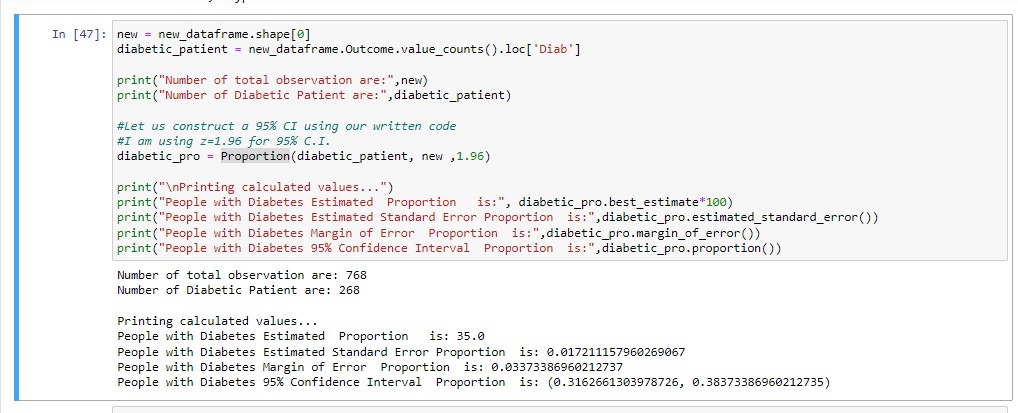


Fig.11. mean and p-value.

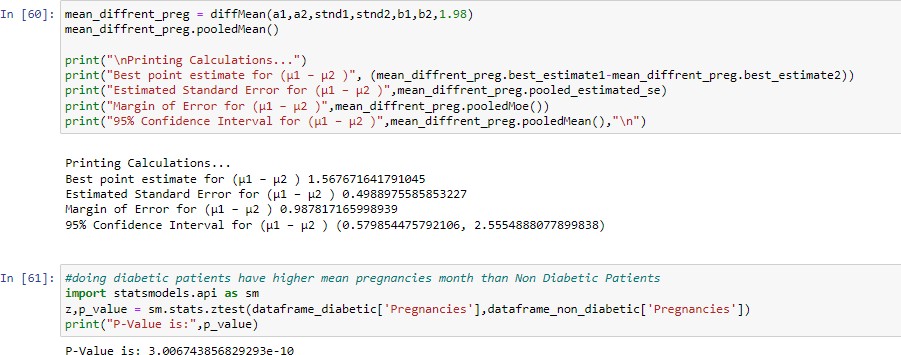
Following that, we will divide the diabetics into those who are pregnant and those who are not pregnant, computing the disparity in mean and determining the p-value.

Fig.12. Calculation of p-value

We reject the null hypothesis since the obtained p-value is below 0.05.

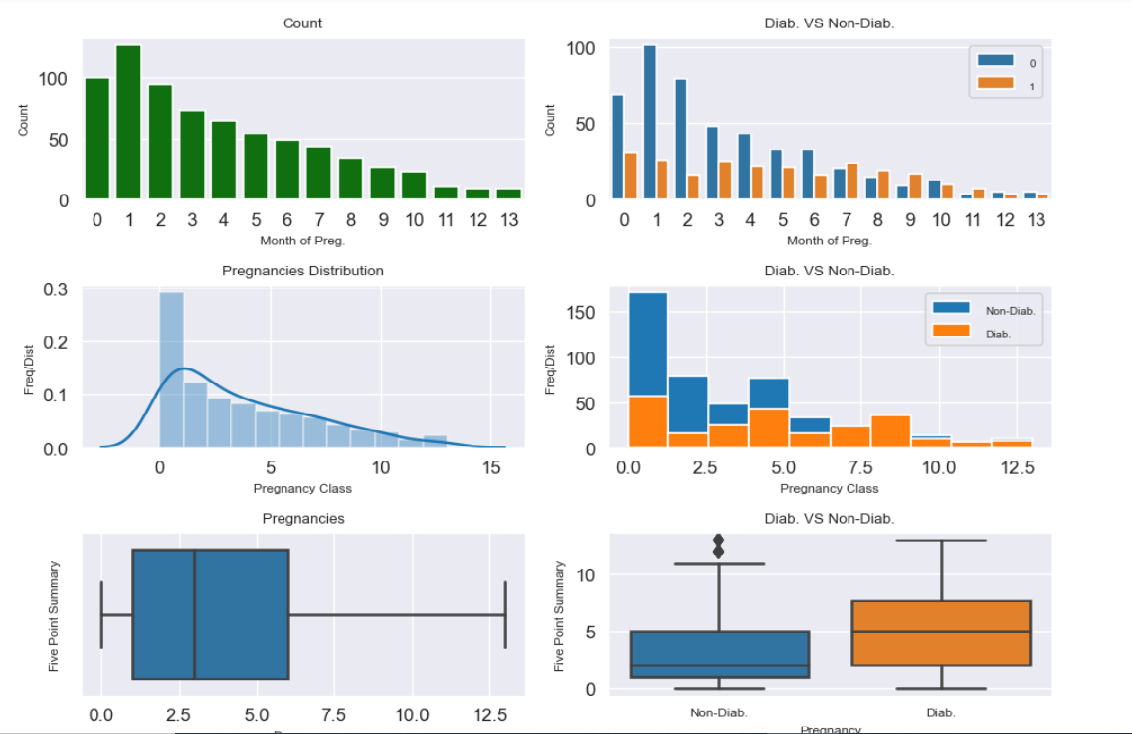


Fig.13. Univariate Statistical Analysis for pregnancy column

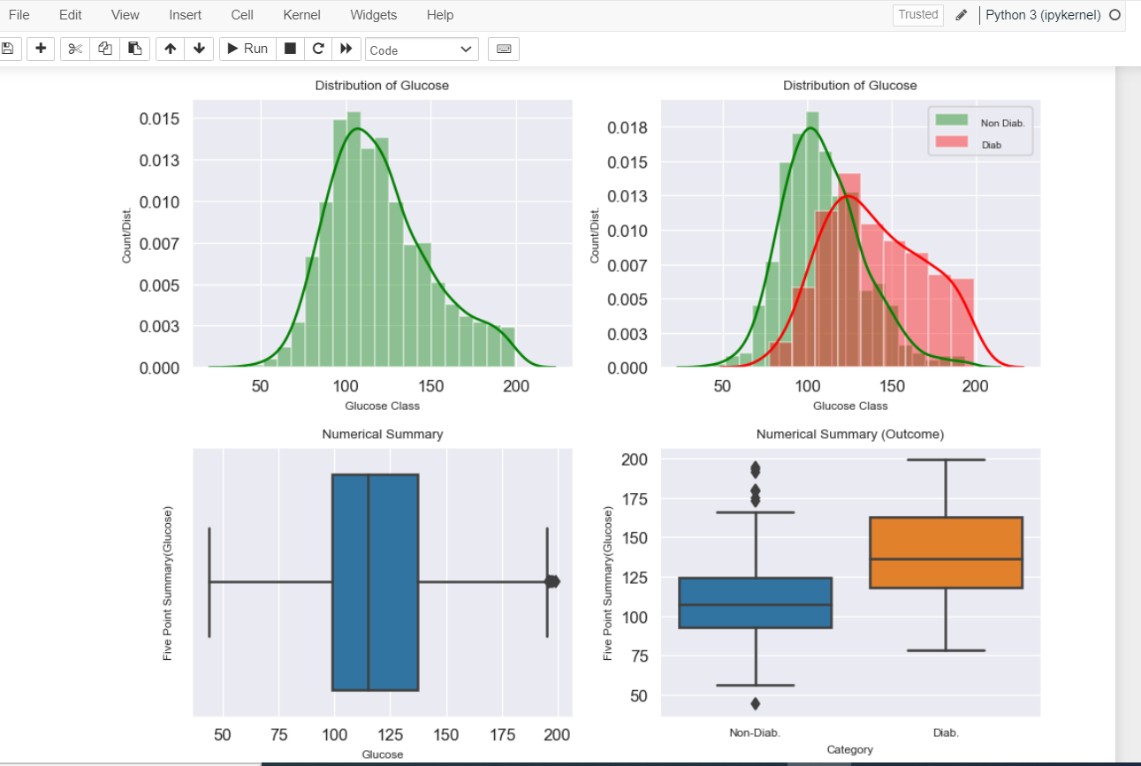
We have conducted a univariate statistical study on pregnancy column.

Fig. 14. univariate statistical analysis on glucose column

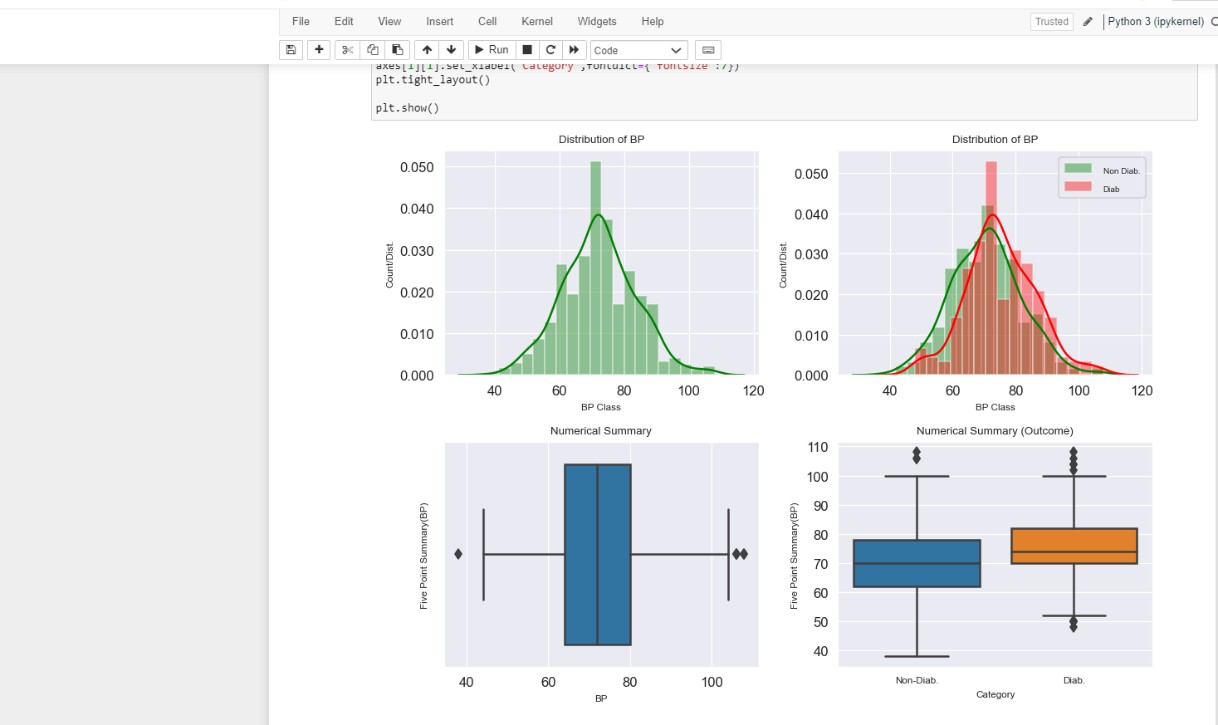
We have conducted a univariate statistical analysis on glucose column.

Fig.15. Univariate statistical analysis on blood pressure column.

Next, we conducted a univariate statistical analysis on blood pressure column.

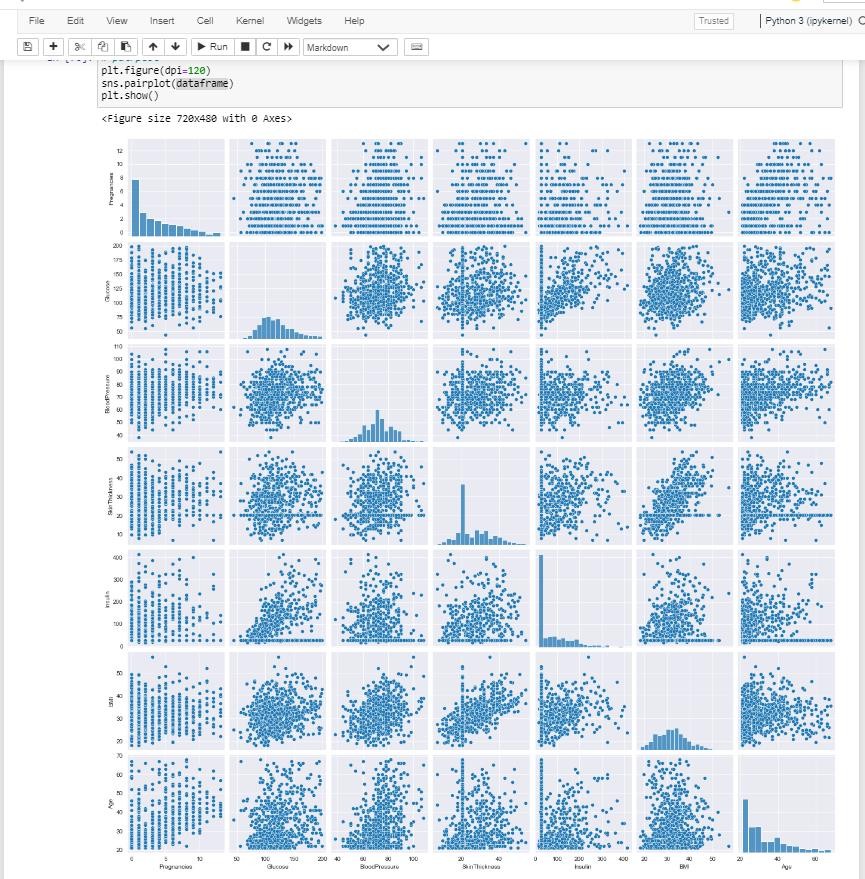


Fig.16. pairplot using multivariate statistical analysis.

At last, we will create a pairplot using multivariate statistical analysis.

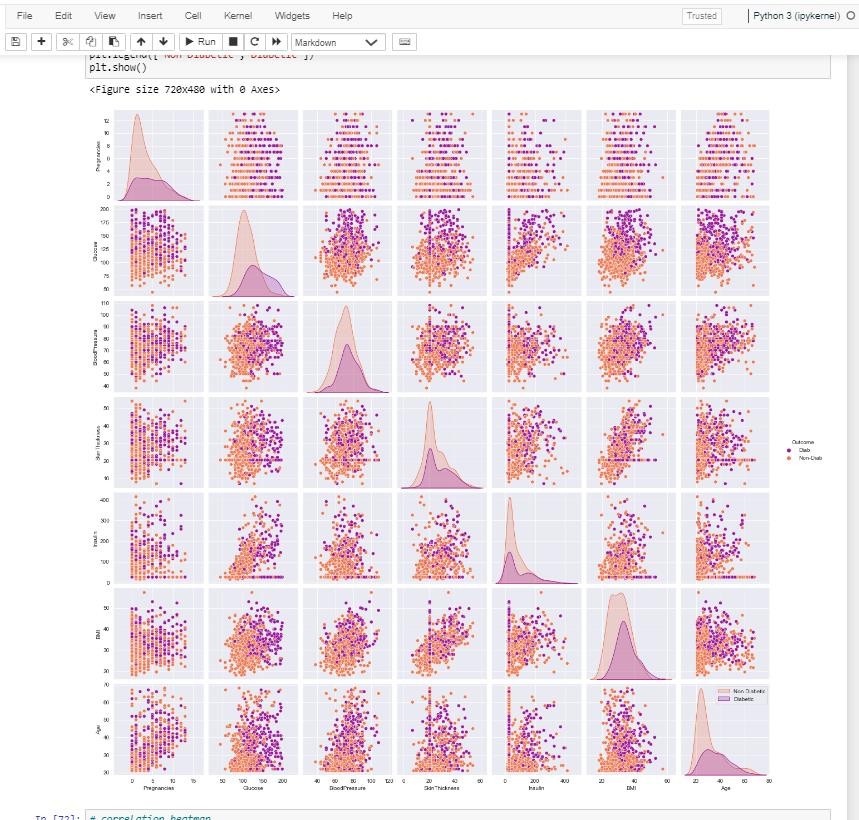


Fig. 17. pairplot based upon the obtained result.

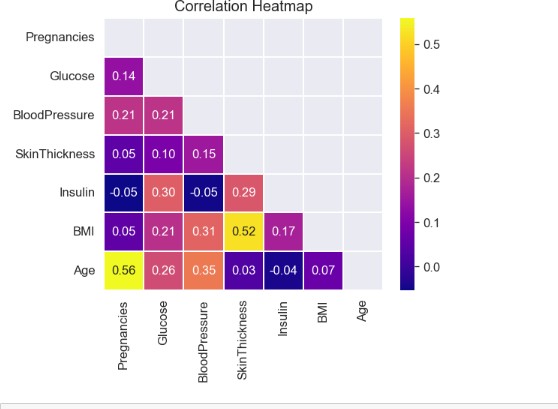
This figure shows the pairplot based upon the obtained result.

Fig. 18. Correlation heatmap

Correlation heat map is plotted between all features.

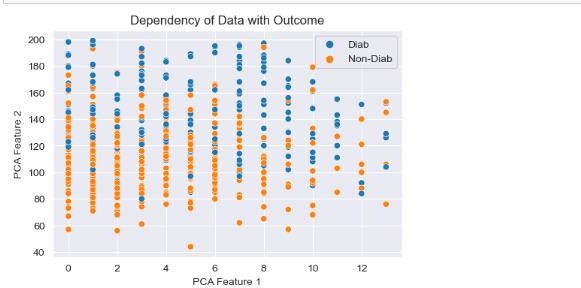


Fig. 19. Dependency of data with outcome

# A screenshot of a computer Description automatically generated Using Machine Learning algorithms:

 Fig. 20 predictive modelling analysis

Fig.21 Using KNN algorithm.

We are dividing the data into 80% for training and 20% for testing, and we are experimenting with various methods. To begin, we used KNN and obtained an accuracy of 79% for the train dataset and 77% for the test dataset.

Fig.22 Using Decision Tree Classifier

Next, we use a decision tree classifier, whose accuracy for the train dataset is 83% and 69% on the test dataset.



Fig. 23. Using SVC

We use the SVC algorithm, which has an accuracy of 78% on the train dataset and test dataset.

A screenshot of a computer code

Description automatically generated

Fig. 24. Using Voting Classifier

Now, we used voting classifier algorithm and obtained an accuracy of 77% on the test set.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

Fig. 26. Random Forest Classifier

Here, we have randomly chosen 5 features from our dataset and calculated training and testing by doing preprocessing.

A screenshot of a computer code

Description automatically generated

Fig. 27. Random Forest Classifier

Now we are dividing the data into 80% for training and 20% for testing, and we are experimenting with various methods. To begin, we used KNN and obtained an accuracy of 79% for the train dataset and 77% for the test dataset.

A screenshot of a computer

Description automatically generated

Fig. 25. Using Feature Selection

Here, we used Feature Selection algorithm for different features in our dataset and obtained the accuracy for each individual feature.

Fig. 28. XGBoost Classifier

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Fig. 29. XGBoost Classifier

We are dividing the data into 80% for training and 20% for testing, and we are experimenting with various methods. To begin, we used KNN and obtained an accuracy of 77% for the train dataset and 77% for the test dataset.

VII. PROJECT MANAGEMENT

1. *Implementation Status Report*
2. *Work Completed:* Description: All necessary datasets, namely research papers, have gathered to be able to carry out the analysis and achieve the intended outcomes. We used few algorithms sections like KNN, Decision Tree Classifier, SVC, Univariate analysis along with Multivariate Statistical Analysis, Exploratory Analysis of Data, Inferential Statistical Analysis, XGBoost Classifier, Voting Classifier, Random Forest Classifier and Feature Selection finished. By extensive commitment in participation, every member from the team made a comparable part in the task at hand as a way fulfil the responsibilities and tasks that were assigned to individuals.
3. *Responsibility and Contribution:*

A table of information

Description automatically generated with medium confidence

VIII. CONCLUSION

* When performing a p-test Our statistic is less significant below the 0.05 cutoff. demonstrates the fact that the null value may be rejected.
* During the Univariate Statistical Analysis process prevalence of Pregnancies, the average number of pregnancies in the data is unimodal, which is unequal at the right and in center and the value of 1 varies from 0 to 15. On the upper top, there exist outliers.
* The range of glucose levels in the information is predictable and single modal. centered at approximately 115 and fluctuates from 90 to 140.On the bottom end, there exist outliers.
* The information's course of blood pressure is predictable and unimodal. fluctuates between 60 and 90, with a center around 65. On the lower conclusion, there exist outliers.
* The Multivariate Analysis with Pair Plot visualization reveals asymmetrical distributions for age, DF, insulin, and pregnancies. Most machine learning models require data distribution.
* Typically. They must grow what we must accomplish this, however rather than expanding it, we assume these are regularly dispersed. Eliminating the anomaly as well.
* The data set cannot be separated by linearity; hence the algorithm is going to fail. This kind of data tree architecture or neural network-based systems will produce high-quality output.
* Following model runs using a variety of algorithms, they can observe that SVC outperforms KNN and Decision Tree, XGBoost and Random Forest classifiers are performing the best out of all SVC.

VIII. REFERENCES

1. Alzaabi, Ashraf, et al. "Prevalence of diabetes and cardio‐metabolic risk factors in young men in the United Arab Emirates: A cross‐sectional national survey." *Endocrinology, diabetes & metabolism* 2.4 (2019): e00081.
2. Malik, Sumbal, Saad Harous, and Hesham El-Sayed. "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women." *International Symposium on Modelling and Implementation of Complex Systems*. Cham: Springer International Publishing, 2020.
3. Han, Jianchao, Juan C. Rodriguez, and Mohsen Beheshti. "Discovering decision tree- based diabetes prediction model." *Advances in Software Engineering: International Conference, ASEA 2008, and Its Special Sessions, Sanya, Hainan Island, China, December 13-15, 2008. Revised Selected Papers*. Springer Berlin Heidelberg, 2009.
4. Kumari, Saloni, Deepika Kumar, and Mamta Mittal. "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier.”

*International Journal of Cognitive Computing in Engineering* 2 (2021): 40-46.

1. Alawadi, Fatheya, et al. "Emirates Diabetes Society consensus guidelines for the management of type 2 diabetes mellitus–2020." *Dubai Diabetes and Endocrinology Journal* 26.1 (2020): 1-20.
2. Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." *Frontiers in genetics* 9 (2018): 515.
3. Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." *Frontiers in genetics* 9 (2018): 515

**GITHUB LINK: -**

[**https://github.com/supriya455/emperical**](https://github.com/supriya455/emperical)