Name: Shinza Jabeen

Student Number: 5012207852

Supervisor name: Ceni Babaoglu

Project Title: Diabetes Prediction Using Machine Learning

Date: 20th October,2023

Introduction:

Diabetes is a long-term health issue in Canada, affecting people of all ages and causing severe health problems like heart disease, vision loss, kidney failure, nerve damage, and amputation. Diabetes has three types: type 1, 2, and gestational diabetes. In 2013-2014, 3 million Canadians had diabetes, with older males having higher rates. Managing diabetes involves maintaining a healthy lifestyle, taking medication if needed, and monitoring blood sugar levels to prevent complications. Type 1 diabetes, a childhood condition requiring insulin, is prevalent in adults, while type 2 is more prevalent in adults and often linked to obesity or inactivity. Gestational diabetes increases the risk of type 2 diabetes, with 3 million Canadians diagnosed in 2013-2014, higher among older males. Maintaining a healthy lifestyle (eating healthy, exercising, and average weight) is crucial for reducing risk. According to the study, females aged 10 to 54 are diagnosed with diabetes 120 days before or 180 days after conception. The Canadian Public Health Agency works with provinces to track occurrences and establish policy. (Pelletier et al., 2012) [7]

Revised Version of abstract:

Diabetes prediction using Machine learning:

Based on the medical data of patients, I would like to predict the likelihood of diabetes in patients. This is a classification problem. I am using the Diabetes prediction dataset that is publicly available on Kaggle. It can be downloaded. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset .The dataset consists of: Age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and diabetes. It has 100,000 rows and 9 columns. The Research Question is: What are the medical and demographic factors

that influence the likelihood of developing diabetes? And the Research question 2: Based on patients' history, identify if they might be at a risk of developing diabetes?

I will be primarily working with python as my programming language. My objective is to predict the likelihood of diabetes in patients, so I'll utilize machine learning methods focused on classification. The libraries I plan to use include matplotlib for data visualization, scikit-learn for constructing machine learning algorithms, and pandas for data ingestion and manipulation. The models I will be using are Logistic Regression, Decision Tree, Random Forest, and an ensemble approach. Using random forest, I can get the importance of each feature and identify the most important factors in determining the likelihood of diabetes. I will use classification metrics like Accuracy, Precision, Recall, F1 score, Confusion Matrix, and AUC to measure how good the model is doing.

Summary of Article 1:

"Diabetes Prediction Using Machine Learning Techniques"

The authors of this study used multiple machine-learning algorithms to try and predict diabetes. If diabetes is not identified and treated at an early stage, it can cause a number of complications. The scientists used numerous machine learning classification and algorithmic ensembles on a dataset, notably the Pima Indian Diabetes Dataset, to achieve this goal. KNN (K-Nearest Neighbour), Decision Tree, SVM (Support Vector Machine), Logistic Regression, Gradient Boosting, and Random Forest, were among the algorithms utilized. The dataset included several factors that are linked to diabetes, including age, blood pressure, BMI, glucose, pregnancy, skin thickness, and insulin levels. The data was divided into sets for testing and training and preprocessed to eliminate any missing values. According to the experimental findings, Random Forest outperformed other machine learning methods in terms of accuracy. This

shows that, based on the provided dataset, Random Forest was the most successful model for predicting diabetes. The authors also examined the significance of several factors in the procedure for predicting. In conclusion, the study showed how machine learning methods might be used to predict diabetes and offered insightful information on the significance of various features. The findings can help with early diabetes diagnosis and better diabetes treatment, which will eventually improve patient outcomes. (Soni & Varma, 2020) [5]

Summary of Article 2:

"Diabetes Prediction Using Machine Learning Algorithms"

This study discusses the important subject of Diabetes Mellitus, a condition that significantly affects a large number of people globally. Age, lifestyle, obesity, high blood pressure, and genetics are just a few of the elements that might cause diabetes, which can have serious health repercussions. The accuracy of traditional diagnostic techniques isn't always the best and they require numerous time-consuming procedures. The standard features of glucose, body mass index (BMI), age, and insulin are combined with external factors in this article to improve diabetes prediction. The study follows a defined methodology, beginning with dataset collecting and scaling and handling missing values. Then, patients are divided into diabetes and non-diabetic groups using K-means clustering. The research then investigates several machine learning techniques, such as DTC (Decision Tree Classifier), RFC (Random Forest Classifier), SVC (Support Vector Classifier), and others, for diabetes prediction. Notably, Logistic Regression produced a maximum accuracy of 96%, and when a pipeline model was used, it improved even more to 97.2%. The superiority of the suggested model is demonstrated by comparing it with the PIMA Diabetes Dataset. In order to minimize false-negative and false-positive diagnoses, that can result in unnecessary treatments or missed interventions, the study addresses the expanding

diabetes problem. The authors offer a potential remedy to improve accuracy and lessen human effort in predicting diabetes by automating the diagnosis procedure using machine learning. This study demonstrates the effectiveness of data-driven prediction models in enhancing patient care and results, which has ramifications for the healthcare sector. The possibility that those who are not already diabetic will become diabetic in the future could be the subject of future research. (Mujumdar & Vaidehi, 2019)[6]

Summary of Article 3:

"Current Techniques for Diabetes Prediction: Review and Case Study"

This study offers a thorough analysis of current advances in diabetes prediction, concentrating on DL (deep learning), ML (machine learning), and hybrid models since 2013. The publication reviews over 40 research investigations, providing insights into different DL and ML algorithms and classification methods, going beyond prior surveys in terms of classifier diversity in recognition of the global impact of diabetes and the necessity for early identification. The study discusses recent DL and ML studies, such as Random Forest, SVM (Support Vector Machine), Deep Neural Networks, Naive Bayes, and Convolutional Neural Networks, in the related works section. Datasets such as the Pima Indian Diabetes dataset are used to assess these approaches. For possibly more accurate predictions, the study emphasizes integrating classifiers and considers a combined dataset. Results vary between investigations and are dependent on the classifiers' accuracy levels. The report offers a thorough view of the advantages and disadvantages of each strategy. In a related study, the authors used the Pima Indian Dataset to investigate uncommon classifiers for diabetes prediction. The REPTree achieved 74.48% accuracy, KStar excelled at managing noisy data (68.23%), and oneR successfully classified fresh instances with 70.83% accuracy. These classifiers were divided into numerous kinds. SMO, which is used to train Support

Vector Machines, had an accuracy rate of 72.14%, while BayesNet had a rate of 73.83%. These classifiers showed competitive accuracy, and it is advised to utilize them in studies on prediction. The REPTree decision tree method in particular shows promise in a variety of applications. The authors advise further investigation, using these infrequently used classifiers on additional datasets, and considering their fusion with other DL, ML, and AI methods to improve prediction accuracy. Overall, by addressing a variety of methodologies and datasets and suggesting the construction of a centralized repository for future study, this work provides an invaluable resource for scholars as well as practitioners in the discipline of diabetes prediction. (Marie-Sainte et al., 2019)[4]

Summary of Article:4

"Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers"

This study uses the PID (Pima Indian Diabetes) dataset to construct a complex ensemble model to study diabetes prediction. The study highlights the crucial part that preprocessing plays in improving prediction quality, with an emphasis on rejecting outliers and filling in missing variables. The distribution features of the dataset's attributes are significantly improved by these preprocessing processes, which take care of problems like kurtosis and skewness. To strengthen the relationship between features and the desired result, feature selection approaches are used, such as correlation-based attribute selection. Although data standardization is proven to have less of an impact on the efficiency of tree-based classifiers, it is still crucial to maintain data consistency.

To verify the reliability of classifiers such as MLP (Multilayer Perceptron), Extreme Gradient Boosting (XB), and the suggested ensemble model, the study applies rigorous 5-fold cross-validation. To improve the learning capacities of various classifiers, hyperparameter tuning using

grid search optimization is used. The utilization of AUC (Area Under the ROC Curve) as a weighting method for creating a potent ensemble classifier is the research's fundamental innovation. For datasets with substantial inter-class redundancy and non-linear separability, like the PID dataset, random tree-based classifiers are effective. The suggested framework performs better than previous approaches, according to comparative results, indicating that it has the ability to predict diabetes from the PID database. Given their low correlation, the optimal method is the combination of both boosting-type classifiers, XB (Extreme Gradient Boosting) and AB (Adaptive Boosting). The study highlights the potential of its suggested structure for diabetes prediction in its conclusion and proposes possible directions for further investigation in this area. (Hasan et al., 2020)[3]

Summary of Article:5

"Prediction of Diabetes Using Classification Algorithms"

The crucial problem of early diagnosis of diabetes using machine learning classification algorithms is addressed in this research study. Early detection is essential for efficient management of diabetes, a severe chronic condition characterized by increased blood sugar levels. The SVM (Support Vector Machine), Decision Tree, and Naive Bayes classification algorithms are the three that the study concentrates on. The PIDD (Pima Indians Diabetes Database), a dataset taken from the UCI machine learning repository, is used to test these algorithms. Various measures, such as Accuracy, Precision, Recall, F-Measure, and ROC (Receiver Operating Characteristic) curves are used in the research to evaluate the algorithms' performance. The findings show that Naive Bayes outperforms the other algorithms in terms of accuracy, with a rate of 76.30%. The research also uses 10-fold cross-validation to guarantee the evaluation's robustness. The study emphasizes the utilization of the WEKA tool for carrying out the tests, and the dataset used includes medical data

of 768 female patients. According to the assessed performance measures, the research concludes that Naive Bayes is the most reliable machine learning classifier to predict diabetes. This approach helps to identify diabetes early, which is important for preventing the complications linked to the disease. The system and machine learning algorithms created in this research could be expanded in further work to forecast and identify different diseases. The results of the study point to the possibility of automating diabetes analysis and broadening the use of machine learning in healthcare. (Sisodia & Sisodia, 2018)[8]

Summary of Article:6

"Data Mining Models Comparison for Diabetes Prediction"

The goal of this study is to use data mining techniques to predict diabetes early on. In order to effectively manage diabetes, which is a rising health concern, early detection is essential. On the PID (Pima Indians Diabetes) dataset, the study uses three different types of data mining algorithms: Naive Bayes, Decision Tree, and KNN (K-Nearest Neighbour). Women who are at least 21 years old and live in Phoenix, Arizona, USA, are represented in the PID dataset. It is a classification in a binary dataset, where '0' denotes a diabetes test result that is negative and '1' denotes a diagnosis that is positive. The dataset is first pre-processed in the study, which includes managing missing values and categorizing numerical data. With different parameters, the three methods are then applied to the dataset. The accuracy of the Decision Tree was the highest, coming up at 75.6 percent. When compared to the other algorithms, KNN had the lowest accuracy (71.74% for Naive Bayes). The study shows that by using data mining techniques, it is possible to accurately forecast diabetes using patient data. The study emphasizes the value of data mining for early disease prediction in the healthcare industry. It provides insights into which approach could be most appropriate for pattern identification and early diagnosis in the context of medical care by

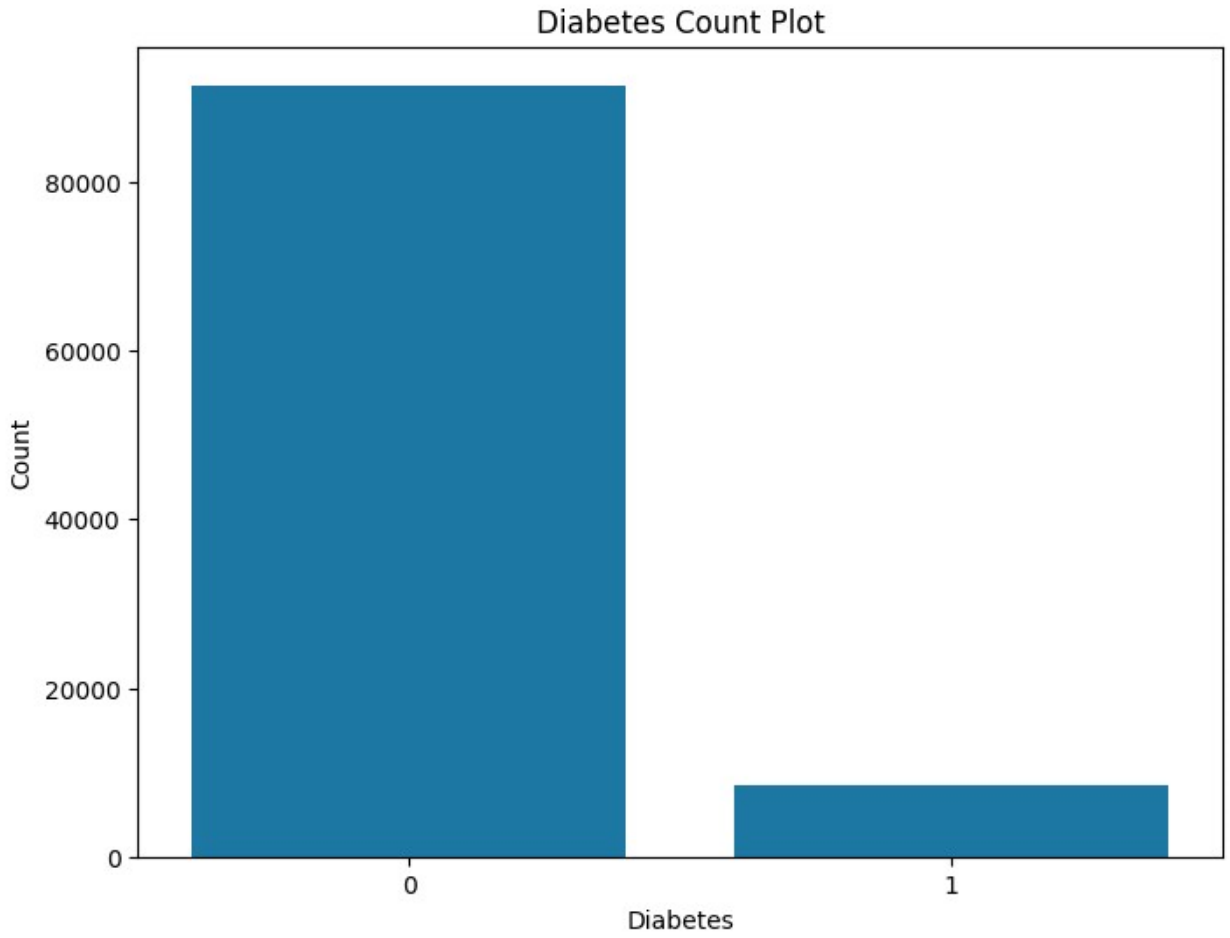
comparing these algorithms. The particular dataset and features used may influence the best algorithm to use. RapidMiner is used in the study's testing and validation, using a 70:30 split between training and test datasets. With a 75.65% accuracy rate, the study's findings generally imply that Decision Tree is the best algorithm for this specific dataset. (Azrar et al., 2018)[2]

Summary of Article:7

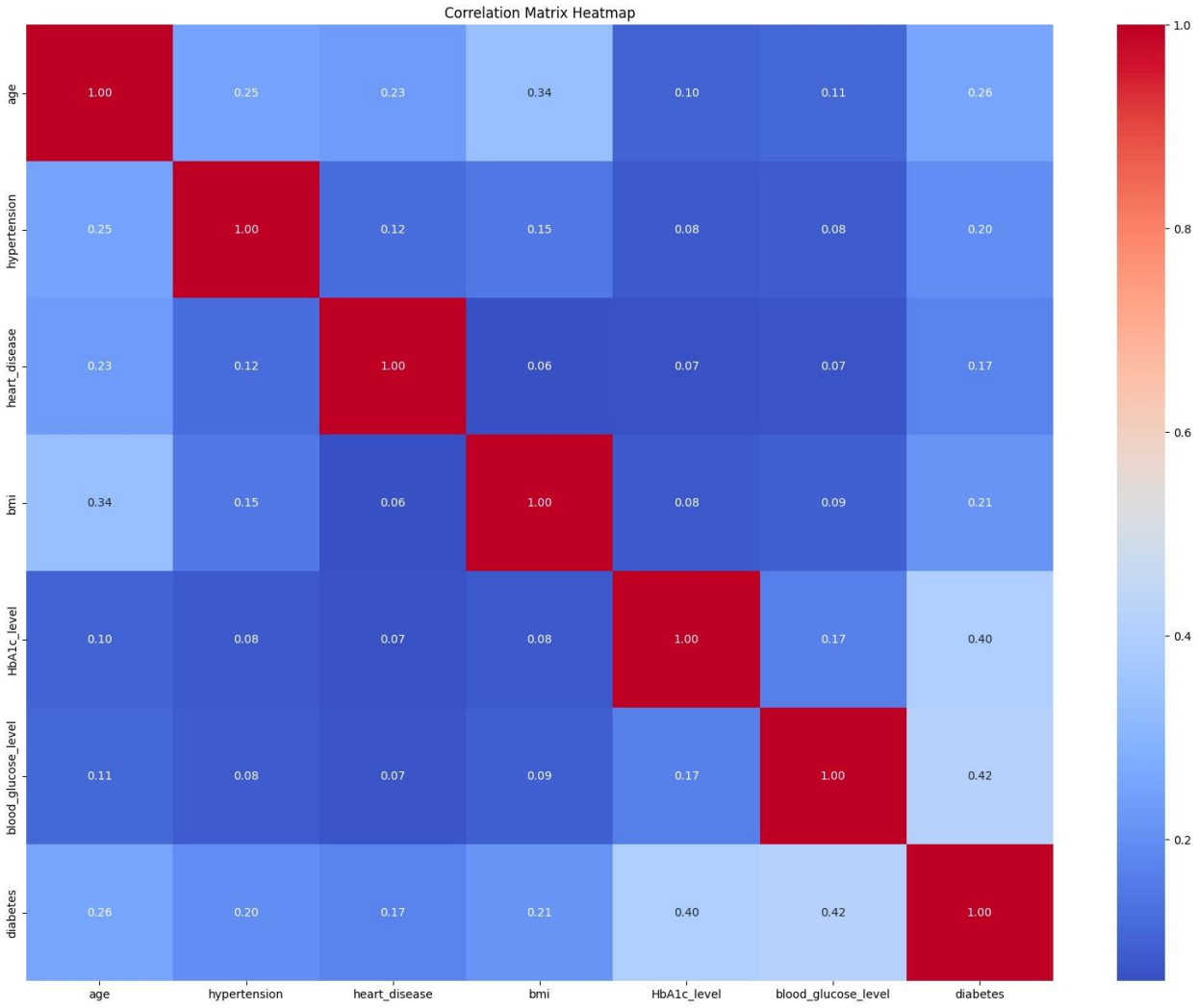
"Diabetes Prediction: A Deep Learning Approach"

The article "Diabetes Prediction: A Deep Learning Approach", which was published in 2019, addresses the critical issue of diabetes, a condition that is common and severe throughout the world, particularly in Bangladesh. The authors suggest a system that uses deep neural networks to accurately diagnose diabetes. For this study, they use the PID (Pima Indian Diabetes) dataset from the UCI machine learning library. The findings of this study show that deep learning can accurately predict diabetes, with a five-fold cross-validation prediction rate of 98.35% being particularly noteworthy. The system also achieves a ten-fold cross-validation accuracy of 97.11%. These findings show that the deep learning strategy beats other diabetes prediction techniques currently in use. The essay offers a thorough description of its approach, which covers data collecting, data preparation, and deep neural network implementation. Additionally, it assesses the system's effectiveness using a number of criteria, including F1 score, accuracy, sensitivity, specificity, and sensitivity to change. The deep learning model shows encouraging results, making it an important tool in the early identification of diabetes. The article's conclusion makes the case that the method of deep learning can significantly advance diabetes detection. It has a high level of resilience and accuracy, particularly in the five-fold cross-validation, which might be an important addition to the medical industry. This study supports current initiatives to combat diabetes, a rising worldwide health issue. (Ayon & Islam, 2019)[1]

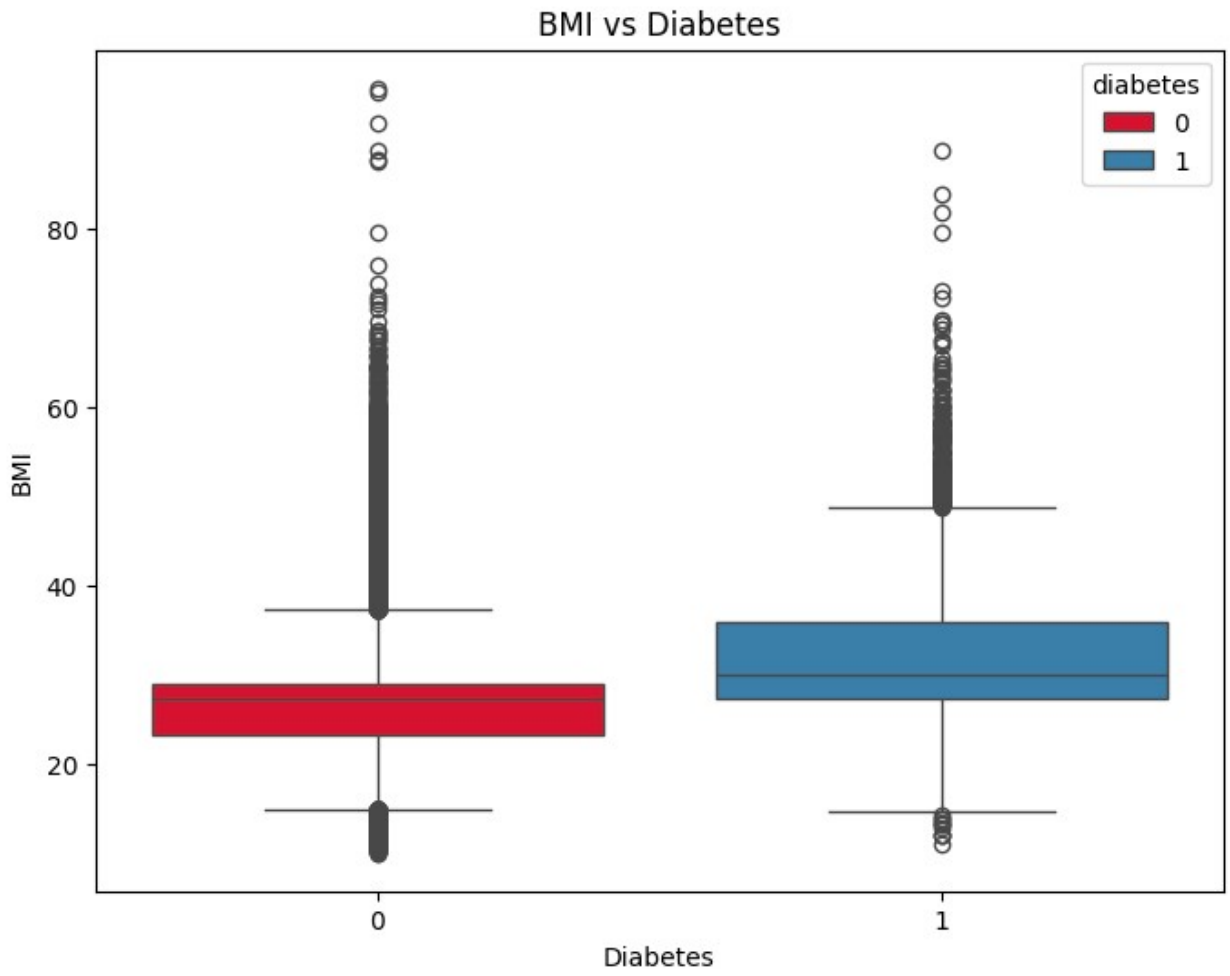
Descriptive statistics:



The dataset contains 8500 of 1s and 91500 of 0s for our dependent variable Diabetes. This data is highly imbalanced meaning there are much more rows with value 0 than there are for 1. We will need to use strategies to take care of data imbalance when running machine learning models.



Diabetes is moderately positively correlated with blood glucose level and Hba1c Level.



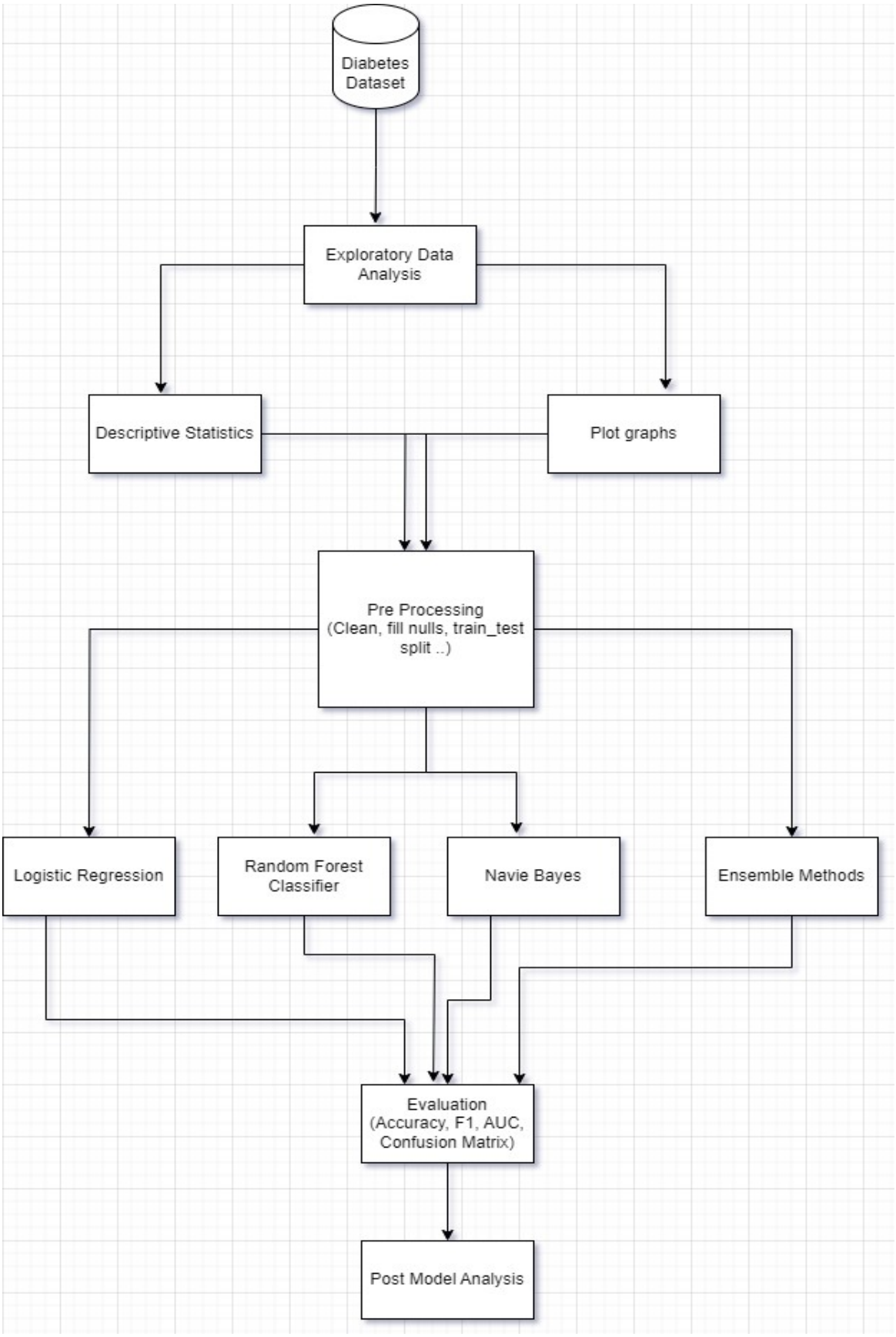
The people who have diabetes have higher bmi than people who do not have Diabetes.

I have checked for the missing values for the dataset and there are no missing values.

Link for the GitHub Repository:

https://github.com/ShinzaJabeen/Diabetes_prediction

Tentative methodology:


Diabetes Dataset
Exploratory Data Analysis
Descriptive Statistics
Plot graphs
Pre Processing (Clean, fill nulls, train_test split ..)
Logistic Regression
Random Forest Classifier
Navie Bayes
Ensemble Methods
Evaluation (Accuracy, F1, AUC, Confusion Matrix)
Post Model Analysis

Exploratory analysis:

In the exploratory data analysis, I will do the 2 parts one is doing the visualization through the graphs like bar plots. In the descriptive statistics I will find the mean, median, mode, value count, standard deviation, and quartiles to analyze the data and to know the relation of between the variables.

Preprocessing:

In the preprocessing, I will clean the data and check for the null values and train the dataset for further machine learning algorithms.

Algorithms:

I will use the algorithms logistic regressions, Random Forest Naive bayes theorem to predict the diabetes based on features. In Ensemble I will combine the predictions and find the mode of it by combining them.

Evaluation:

After performing the machine learning algorithm, I will evaluate which algorithm performed better by Accuracy, F1, AUC and confusion matrix. and through post model analysis, i will analyze the feature selection and feature importance.

References

[1] Ayon, S., & Islam, M. (2019). Diabetes Prediction: A Deep Learning Approach. *I.J. Information Engineering and Electronic Business, 2*. Retrieved from https://j.mecs-press.net/ijieeb/ijieeb-v11-n2/IJIEEB-V11-N2-3.pdf

[2] Azrar, A., Awais, M., Ali, Y., & Zaheer, K. (2018). Data Mining Models Comparison for Diabetes. *International Journal of Advanced Computer Science and Applications, 9*. Retrieved 2023, from https://pdfs.semanticscholar.org/3be5/c172b0c99e83cfd7a6e0b01ad89ed7363cfd.pdf

[3] Hasan, K., Alam, A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access , Vol.8*. Retrieved 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9076634

[4] Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019, 10 29). Current Techniques for Diabetes Prediction. *Applied Sciences*. Retrieved 2023, from https://www.mdpi.com/2076-3417/9/21/4604/pdf?version=1572920776

[5] Soni, M., & Varma, S (2020, 09). Diabetes Prediction using Machine Learning. *International Journal of Engineering Research & Technology, Vol. 9*(Issue 09). Retrieved 2023, from https://d1wqtxts1xzle7.cloudfront.net/64739619/diabetes_prediction_using_machine_learning_techniques_IJERTV9IS090496-libre.pdf?1603351804=&response-content-disposition=inline%3B+filename%3DIJERT_Diabetes_Prediction_using_Machine.pdExpires=1697291953&Sig

[6] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science, Volume 165*. Retrieved 2023, from https://www.sciencedirect.com/science/article/pii/S1877050920300557?ref=pdf_download&fr=RR-2&rr=8169344f7ff1a253

[7] Pelletier, C., Dai, S., Roberts, K. C., Bienek, A., Pelletier, L., & Onysko, J. (2012). Report Summary - Diabetes in Canada: facts and figures from a public health perspective. Retrieved 2023, from https://www.researchgate.net/publication/234085745_Report_Summary_-_Diabetes_in_Canada_facts_and_figures_from_a_public_health_perspective

[8] Sisodia, D., & Sisodia, D. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science, 132*. Retrieved 2023, from https://www.sciencedirect.com/science/article/pii/S1877050918308548?ref=pdf_download&fr=RR-2&rr=8169ba691849a232