DATA 606: Data Science Capstone

Masters in Data Science UMBC

# Analysis and Comparison Of Early-Stage Diabetes Risk Prediction and Diabetes Prediction

Utkarshika Baidya

Pranavi Gadikoppula

## ABSTRACT

Diabetes is a chronic disease with numerous health complications that affect millions of people globally.[1] Timely detection and management of diabetes can significantly improve health outcomes and lower the risk of complications. Increase In diabetes can cause other disease like kidney issues, damaged blood vessels etc. Delay in predicting these can lead to loss of life. Early prediction can save life. Many people are unaware about the impacts of the disease. There are many automated techniques to predict. This study proposes a machine learning-based approach to predict early-stage diabetes risk and diabetes prediction. Data was collected from a large population of people with and without diabetes, covering demographic, clinical, and lifestyle factors. Several machine learning models were trained, including logistic regression, random forest, and support vector machines, to predict diabetes risk based on the data. The analysis also identified several important diabetes risk factors, including age, BMI, family history, and physical activity. These findings could help healthcare professionals identify high-risk individuals early and take preventative measures. The proposed approach could be a valuable tool for diabetes risk prediction, helping to reduce the burden of this chronic disease on individuals and healthcare systems.


## INTRODUCTION:

Diabetes is a condition that results from high levels of blood glucose, also known as blood sugar. Blood glucose is the body's primary source of energy and comes from the food we consume. Insulin, a hormone produced by the pancreas, facilitates the transport of glucose from food into cells to be utilized for energy. However, in some cases, the body doesn't produce sufficient insulin or is unable to utilize insulin effectively, resulting in the buildup of glucose in the bloodstream and a failure to reach the cells.[2]

As of 2015, 30.3 million people in the United States, or 9.4 percent of the population, had diabetes. More than 1 in 4 of them didn’t know they had the disease. Diabetes affects 1 in 4 people over the age of 65. About 90-95 percent of cases in adults are type 2 diabetes.[3] Diabetes is a multifactorial metabolic disease, its diagnostic criteria is difficult to cover all the ethology, damage degree, pathogenesis, and other factors, so there is a situation for uncertainty and imprecision under various aspects of medical diagnosis process. With the development of Data mining, researchers find that machine learning is playing an increasingly important role in diabetes research. Machine learning techniques can find the risky factors of diabetes and reasonable threshold of physiological parameters to unearth hidden knowledge from a huge amount of diabetes-related data, which has a very important significance for diagnosis and treatment of diabetes.

According to WHO, In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. Another 460 000 kidney disease deaths were caused by diabetes, and raised blood glucose causes around 20% of cardiovascular deaths. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes. In lower-middle-income countries, the mortality rate due to diabetes increased 13%.[4]

**Paper Review:**

Diabetes Mellitus is a disease which is caused due to multiple causes it is a metabolic disease with chronic Hyperglycemia. The Primary cause is because of insulin secretion. The symptoms of this disease are polyuria, polydipsia, polyphagia, and weight loss in addition itching can also be caused. In addition to these the disease can also show its effect on eyes, kidneys, nerves, heart, and blood vessels. Diabetic Ketoacidosis can also occur if the diabetes is more and under stress. Currently, there are three types of diseases one is immune mediated which is very slow and almost at the early stage there is idiopathic which comes under type 1 which has no evidence of autoimmunity. The second type of Diabetes is insulin resistance, and the third type is gestational diabetes. there are many factors for the cause of diabetes they are dietary factors and chemical poisons. This disease can show impact on blurred vision, shortness of breath numbness and many more. To predict the diabetes there are many methodologies in this paper they have choose SVM Machine learning Technique. The algorithm involves Dataset with patient information then applying a, Naïve Bayes and Light GBM prediction algorithm performance is evaluated with the confusion matrix. Then finds the best algorithm for the best algorithm for prediction.

The dataset for this they have collected 520 patients at Sylhet disease hospital where it has considered many factors like gender, age, weakness, visual blurring, partial paresis etc. The Support Vector Machine algorithm is a linear classifier it is based on supervised technique it performs binary classification. It has robustness and sparsity it can also perform nonlinear classification with kernel method. The accuracy we get is 91.18%. Naïve Bayes classifier is based on bayes theorem it is a probability classifier this is helpful for predicting the probability of the diabetes. The accuracy of this method gave 93.27%. The Light GBM is a gradient boosting framework this is based on decision tree it is faster in training efficiency, less memory usage the accuracy of this algorithm is 88.46%. The early detection can play an important role to treat the disease and reduce the risk of the disease out of all the methods SVM gives the highest accuracy. In this way using different algorithm techniques we can predict diabetes at early stage.[5]

**Paper review:**

Diabetes is most prone disease, and it is life threatening almost 422 m people are suffering with this disease. This disease is caused due to increase in blood Glucose which can harm heart, blood vessels, eyes, kidneys, and nerves. This disease is also the major cause of death over the year. There are three different types of diabetes one disease attacks pancreatic beta cells which causes no insulin production, The second one uses insulin. The third one is Gestational Diabetes which is caused due to pregnancy. It involves lot of cost from diagnosing the disease to treat the disease as it need long duration for treatment. As the disease is asymptotic in nature it will be difficult to predict at early stage. There are few established methods like Oral Glucose Tolerance Test and HbA1c which may not be available for many people as it is costly.  To predict this diabetes few methodologies such as feature selection can detect the disease at early stage and this process doesn't cost much. This methodology helps in removing overfitting and redundant data.

The method involves Pre-Processing, Feature Selection where it is done with four algorithm which involves Ranker search method. The classification is done using classifier such as KNN, Naïve Bayes, Random Forest. The evaluation of results is done based on confusion Matrix and accuracy, ROC, AUC and F-measure metrics. The whole process is done on weka platform. They collected data of 520 individuals with different factors into consideration like glucose level, BMI, Blood pressure level, male, female. Data Processing is done by removing the duplicates and finding out the missing value. Feature selection is the process of recognizing a subset of data from a large dataset. With the use of Symmetrical Uncert Attribute Evaluator, Info Gain Attribute Evaluator, Gain Ratio Attribute Evaluator which is together worked with ranker search method. After all this process cross validation and percentage split are used for test mode option. ANN method gave 75.7 % accuracy, random forest gave 74.7% and k means gave 73.4% accuracy. If the disease is not predicted at early stage there can be chances of kidney problems and produces more urine, there are chances of getting Partial paresis when body is not able to control the sugar in the blood. Using all this machine learning technique detection of diabetes at early stage is done and can reduce the cost of diagnosis.[6]

**Paper Review:**

Diabetes is the most prevalent illness; it affects 8.5% of individuals over the age of 18, and it contributes to over 1.6 million deaths annually. Over 18% of people die from diabetes worldwide. Heart disease, renal issues, and many more difficulties are brought on by this condition. It will be simpler to live a longer, healthier life if the condition can be recognized early. In this study, supervised machine learning models such as Decision Tree, Naive Bayes, K-NN, Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine were used to predict the presence of diabetes. They also added a few processing techniques, such as label encoding and normalization, to further improve accuracy.

The risk variables have been identified and prioritized while applying the procedures. They created a web application in addition to using several datasets to test the accuracy. In this project, the data was preprocessed using various preprocessing techniques, and the performance of the algorithms was evaluated using parameters like precision, recall, f1-score, roc curve, and accuracy. Finally, they identified a few key factors, such as chi square and correlation, and then created a web application based on these findings. Two datasets containing various characteristics, such as height, weight, and BMI, were gathered for the data collection. Outliers were then removed, missing values were handled, and label encoding was used to make predictions, which were then verified using performance analysis. The accuracy of the dataset1 is 78.95% for NB, 76.32% for DT, 80.25% for RF, 80.26% for SVM, 77.63% for LR, 78.95% for GB, and 75% for KNN. The accuracy of dataset 2 is 5.57% for NB, 12.81% for DT, 2.71% for RF, 4.99% for SVM, and 13.13 percent for KNN. They have also examined the relationship between the risk factors for diabetes. These techniques allow for the prediction of diabetes.[7]

**Paper Review:**

Diabetes mellitus is a fatal condition that can cause damage to the kidneys, the heart, and the nerves. It is a metabolic condition brought on by an increase in blood sugar. In this study, a few machine learning algorithms that are useful for categorization, early-stage diabetes prediction, and diabetes prediction are employed. Additionally, an IOT-based system is developed to determine the state of health of those with diabetes. Long Short-Term Memory, Moving Averages, and Linear Regression are used for the prediction analysis, while Random Forest, Multi-Layer Perceptron, and Logistic Regression are used for classification.

The dataset contains 768 entries with various attributes, including gender, blood pressure, age, insulin, and more. The findings of this study suggest that MLP has the greatest performance at 86.6%, 85.1% precision, and 86.083% accuracy. A prediction accuracy of 87.26% was attained by LSTM. A limited amount of technology has also been employed to deploy IOT for healthcare. They made use of JSON, MongoDB, NoSQL, and Kafka.[8]

**METHODS:**

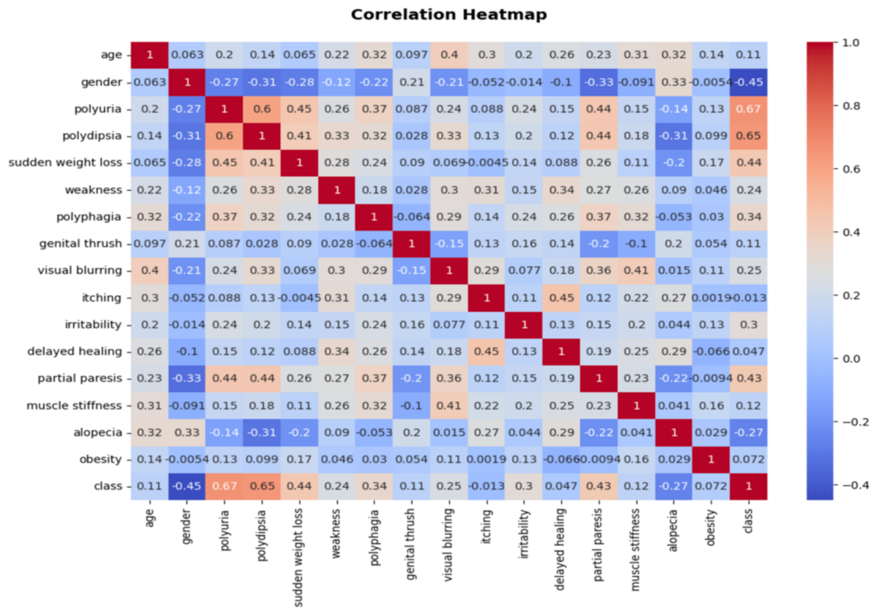
The method employed for predicting early-stage and diabetes using machine learning encompasses several detailed steps as follows:

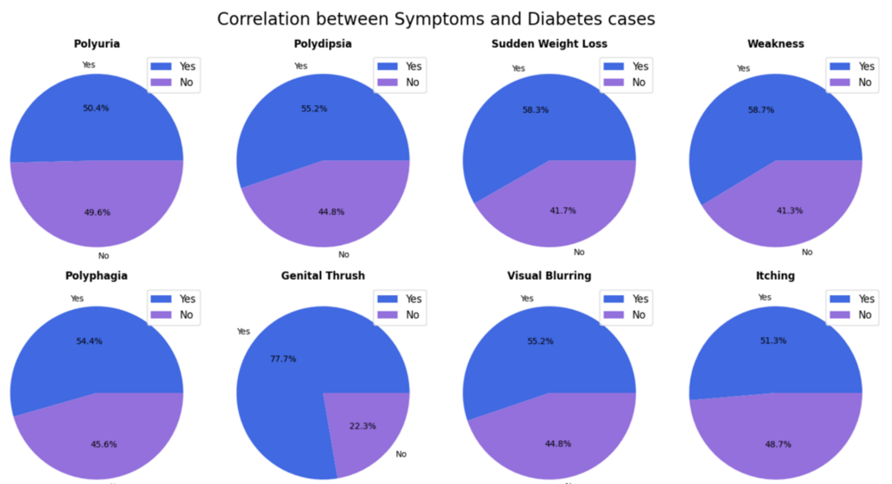
1. Data Loading: The diabetes data was loaded into the notebook using the pandas library. This involved importing the dataset and organizing it into a suitable data structure for further analysis.
2. Data Cleaning and Preprocessing: To ensure data quality, we performed data cleaning and preprocessing. This included identifying and addressing missing data, handling duplicated records, and dealing with inconsistent data values. Techniques such as mean substitution and

outlier detection were applied to address these issues effectively.

3. Data Exploration: This step involved exploring the dataset to gain a better understanding of its characteristics. Various visualizations, such as histograms and scatter plots, were created to uncover patterns, relationships, and potential insights within the data.
4. Correlation Analysis: Correlation maps were generated to identify significant correlations between different features within the dataset. This analysis helped in understanding the relationships between variables and identifying potential predictive factors for early-stage diabetes.
5. Data Splitting: The dataset was divided into training and test sets using the train_test_split function from the scikit-learn library. This separation allowed us to train the models on a subset of the data and evaluate their performance on unseen data.
6. Machine Learning Methods: Classification algorithms such as logistic regression, decision trees, random forest, support vector classification (SVC), K-neighbors, Gaussian Naive Bayes, and Gradient Boosting were employed to predict the likelihood of early-stage diabetes. These algorithms leverage patterns and relationships within the data to make predictions.
7. Model Training and Evaluation: The models were trained using the training data and evaluated using the test data. Performance metrics such as accuracy, precision, recall, and F1-score were utilized to assess the models' predictive capabilities and determine their effectiveness in identifying early-stage diabetes.

Some of the visualizations are given below:




In summary, the process involved loading and preprocessing the diabetes data, exploring the dataset through visualizations and correlation analysis, splitting the data into training and test sets, applying various machine learning algorithms, training and evaluating the models, and optimizing their hyperparameters. This comprehensive approach enabled us to predict early-stage diabetes with high accuracy, providing valuable insights for healthcare providers to take preventive measures, manage the condition effectively, and ultimately improve health outcomes.

**RESULTS AND ANALYSIS:**

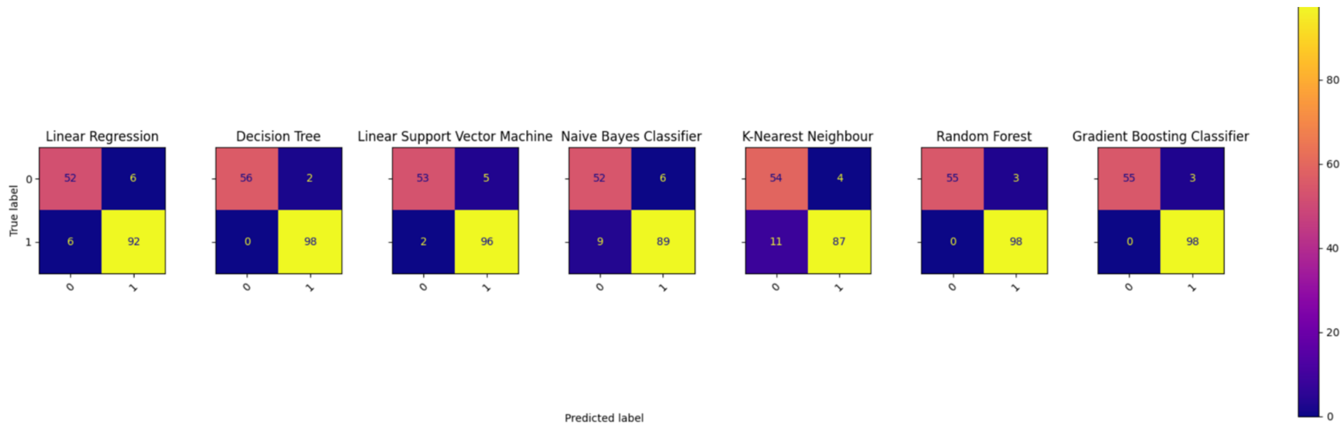
| Model | Accuracy |
|---|---|
| Logistic Regression | 92.31% |
| Decision Tree | 98.08% |
| Support Vector Machine | 91.67% |
| Naive Bayes Classifier | 90.38% |
| K Nearest Neighbour | 90.38% |
| Random Forest | 97.44% |
| Gradient Boosting Classifier | 94.23% |

Table1: of Model and Accuracy Early Stage Diabetes Risk Prediction

The table above summarizes the results of several machine learning models that have been trained to predict early-stage diabetes risk based on a set of features. Here's a brief analysis of the table:

- The table shows the accuracy scores of six different machine learning models: Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes Classifier, K Nearest Neighbour, Random Forest, and Gradient Boosting Classifier.
- The models have been trained and evaluated on a dataset containing the features Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity, and the class (indicating diabetes or not).
- The accuracy scores range from 90.38% to 98.08%, with Decision Tree having the highest accuracy, followed closely by Random Forest.
- Logistic Regression, Support Vector Machine, Naive Bayes Classifier, K Nearest Neighbour, and Gradient Boosting Classifier all have moderate accuracy scores in the range of 90% to 94%.

- The table provides a useful comparison of the performance of different machine learning models on this particular problem, but it's important to note that accuracy is not the only metric that should be considered when selecting a model. Other factors such as interpretability, training time, and the nature of the problem being solved may also be important. Therefore, the choice of the best model should be based on multiple factors and may require further evaluation and validation.




The confusion matrix is a table that visualizes the performance of a classification model by displaying the counts of true positive, true negative, false positive, and false negative predictions. It is a useful tool for evaluating the accuracy of classification models.

In the generated plot, each subplot represents a different machine learning model. The x-axis of each subplot represents the predicted labels, while the y-axis represents the true labels. The cells of the confusion matrix contain the counts or percentages of instances classified into different categories.

The title of each subplot corresponds to the name of the respective machine learning model. The color map used to visualize the confusion matrix is 'plasma'.

Additionally, the plot includes a color bar representing the color scale used in the confusion matrix. The color bar provides a reference for the color mapping of the confusion matrix cells.

Overall, this plot allows for a visual comparison of the performance of different machine learning models based on their confusion matrices.

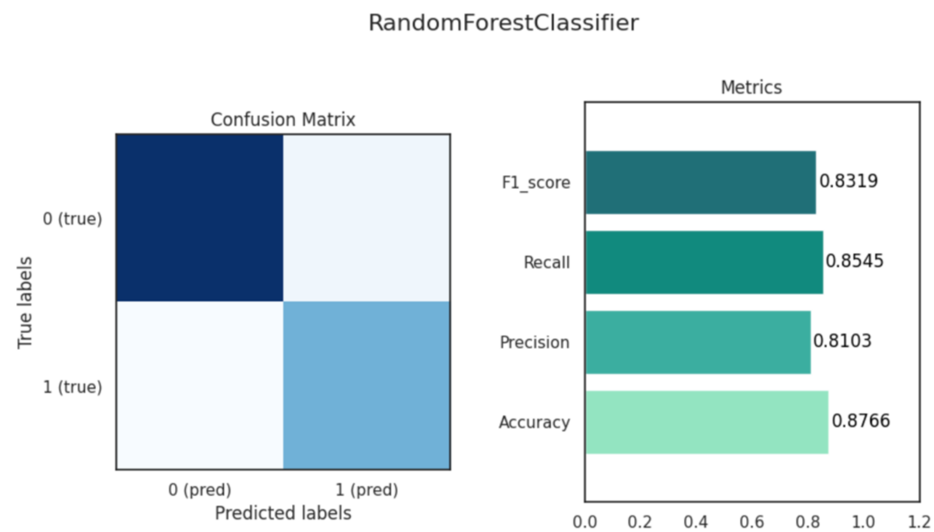
| Classifier | Accuracy |
|---|---|
| GaussianNB | 70.13% |
| DecisionTreeClassifier | 80.52% |
| SVC | 82.47% |
| RandomForestClassifier | 87.66% |
| LogisticRegression | 80.52% |
| KNeighborsClassifier | 82.47% |
| GradientBoostingClassifier | 84.42% |

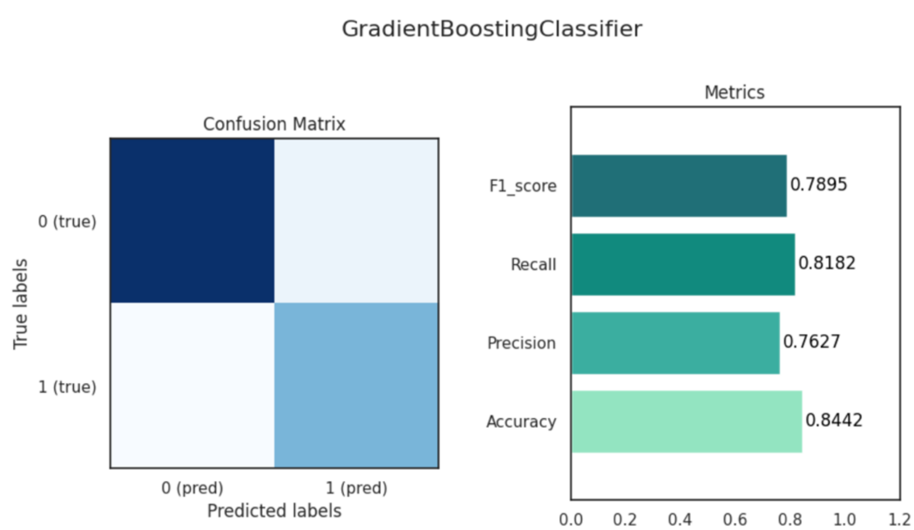
Table 2: Diabetes Prediction

The table depicts the performance of different classifiers in a classification task, specifically their accuracy. Here is a breakdown of the table:

- The highest accuracy is achieved by RandomForestClassifier, which achieves 87.66%. When compared to the other classifiers in the table, this classifier outperforms them in accurately predicting the target variable.
- GradientBoostingClassifier comes in second place with an accuracy of 84.42%. It performs well and is the second most accurate classifier.
- The accuracy of SVC (Support Vector Classifier) and KNeighborsClassifier is the same (82.47%). These classifiers perform similarly, indicating their effectiveness in the classification task.
- The accuracy of DecisionTreeClassifier and LogisticRegression is 80.52%. They perform well but are slightly less accurate than the top classifiers.
- GaussianNB has the lowest accuracy (70.13%) of the classifiers in the table. Despite having the lowest accuracy, it can still provide some classification performance.

- Overall, the table shows that RandomForestClassifier is the most accurate classifier for the task, closely followed by GradientBoostingClassifier. The other classifiers also perform well but fall short of the top two in terms of accuracy. These accuracy metrics can be used to select the most suitable model for classification purposes, depending on the specific requirements and constraints of the task.



**CONCLUSION:**

Finally, the combination of early stage diabetes risk prediction as a preliminary screening tool with machine learning models for predicting the possibility of high-risk patients getting diabetes holds tremendous promise in healthcare.

By identifying those who are at a higher risk of acquiring diabetes at an early stage, healthcare providers can intervene and slow the disease's progression. Individuals with these risk characteristics can be targeted for more rigorous monitoring, specialized therapies, and lifestyle changes. These tailored methods have the potential to postpone or even prevent the onset of diabetes, hence decreasing the burden on individuals and healthcare systems.

The application of machine learning algorithms enables the combination of several risk variables and pertinent data to forecast the chance of high-risk individuals developing diabetes within a specific timeframe. Healthcare providers can receive significant insights into personalized risk assessments and prognosis by employing these models, allowing them to make informed decisions about intervention options.

Furthermore, the integration of these models facilitates ongoing monitoring and reassessment of individuals' risk profiles over time. As new data becomes available, the predictive models can be refined and personalized, allowing for adaptive and dynamic intervention strategies tailored to each individual's changing risk factors and circumstances.

In conclusion, the combination of early stage diabetes risk prediction and machine learning models presents a powerful framework for identifying high-risk individuals and implementing timely interventions. By utilizing these predictive tools, healthcare providers can optimize their efforts in diabetes prevention and management, ultimately improving the overall health and well-being of individuals at risk of developing diabetes.

**'Appendix A: ChatGPT explorations'**

1. How can we predict early-stage diabetes using machine learning?

   Machine learning algorithms can be used to predict early-stage diabetes by analyzing large datasets of demographic, clinical, and lifestyle factors. The following are some common steps involved in using machine learning for early-stage diabetes prediction:

- Data Collection: A dataset of relevant demographic, clinical, and lifestyle factors is collected from a large population, including information such as age, gender, family history, body mass index (BMI), blood pressure, cholesterol levels, physical activity levels, and dietary habits.
- Data Preprocessing: The collected data is preprocessed to remove outliers, handle missing values, and transform the data into a suitable format for machine learning algorithms.
- Feature Selection: The most relevant features for predicting early-stage diabetes are identified from the preprocessed data using feature selection techniques.
- Model Training: A machine learning model, such as a logistic regression, support vector machine, or decision tree, is trained on the preprocessed and selected features of the dataset using a labeled dataset.
- Model Evaluation: The trained model is evaluated on a separate labeled dataset to assess its predictive performance, and the performance metrics, such as accuracy, precision, recall, and F1-score, are calculated.
- Deployment: The trained machine learning model is deployed to predict early-stage diabetes risk in new individuals, using their demographic, clinical, and lifestyle data.
- By analyzing large datasets of demographic, clinical, and lifestyle factors, machine learning algorithms can identify individuals at high risk of developing early-stage diabetes, enabling healthcare providers to take preventive measures and manage their condition to reduce the risk of complications and improve health outcomes.

2. What features do you need to predict early-stage diabetes using machine learning?

To predict early-stage diabetes using machine learning, the following features can be used:

- Age: Age is a crucial factor in determining diabetes risk, as older people are more likely to develop diabetes.
- Gender: Women are more prone to developing diabetes than men, and gender can be an essential factor to consider in predicting diabetes.
- Body Mass Index (BMI): BMI is a measure of body fat based on height and weight, and it is a crucial factor in diabetes prediction, as obese or overweight individuals are at higher risk of developing diabetes.
- Blood pressure: High blood pressure can be a contributing factor to diabetes, and measuring blood pressure can help in diabetes prediction.
- Cholesterol levels: High levels of LDL cholesterol, also known as bad cholesterol, and low levels of HDL cholesterol, also known as good cholesterol, can contribute to diabetes risk.
- Physical activity levels: Sedentary lifestyles and lack of physical activity can increase the risk of developing diabetes.
- Family history: Individuals with a family history of diabetes are at higher risk of developing diabetes.
- Dietary habits: Poor dietary habits, such as consuming high amounts of sugar, processed food, and saturated fats, can increase the risk of developing diabetes.
- By using these features and other relevant demographic, clinical, and lifestyle factors, machine learning algorithms can predict early stage diabetes with high accuracy and help healthcare providers take preventive measures to manage the condition and improve health outcomes.

**'Appendix B: ChatGPT improvement'**

Now that you have a better understanding of your project, go back to ChatGPT and ask TWO good questions which can help you in improving the quality of your work.

1. How can we improve the accuracy of the different machine learning modules for a diabetes risk prediction?

   There are several ways to improve the accuracy of machine learning models for diabetes risk prediction. Here are some possible approaches:

- Feature engineering: This involves selecting or creating the most relevant features (or variables) from the dataset to train the model. This can help improve the accuracy by reducing noise and improving the signal-to-noise ratio in the data.
- Hyperparameter tuning: Many machine learning models have hyperparameters that can be adjusted to improve their performance. Tuning these hyperparameters using techniques like cross-validation can help optimize the model's accuracy.
- Ensembling: This involves combining multiple models (such as decision trees or neural networks) to create a more accurate prediction. This can be done through techniques like bagging or boosting.
- Data augmentation: This involves generating additional data by manipulating existing data or generating new data using techniques like oversampling, undersampling, or synthetic data generation. This can help improve the model's accuracy by increasing the diversity and size of the training dataset.
- Regularization: This involves adding a penalty term to the model's objective function to prevent overfitting to the training data. This can help improve the model's accuracy on new data by reducing variance in the model.
- Using advanced machine learning techniques: Advanced techniques like deep learning or reinforcement learning can sometimes provide better accuracy for complex datasets like medical data. However, these techniques also require more data and computing resources.

- Domain knowledge: Incorporating domain knowledge about diabetes risk factors or biomarkers into the model can also help improve the accuracy. This may involve working with medical experts or conducting literature reviews to identify relevant features or variables.

2. What are the limitations of existing diabetes risk prediction models, and how can these be addressed in the development of a new model?

   Existing diabetes risk prediction models have several limitations that can be addressed in the development of new models. Here are some possible limitations and ways to address them:

- Limited generalizability: Some existing models may have been developed using data from specific populations or regions, which may limit their applicability to other populations. To address this, new models can be developed using more diverse and representative datasets, or by using transfer learning techniques to adapt existing models to new populations.
- Limited accuracy: Existing models may have limited accuracy due to factors such as incomplete or biased data, overfitting, or model assumptions that do not hold for all patients. To address this, new models can be developed using advanced machine learning techniques, such as deep learning or reinforcement learning, that can better handle complex or high-dimensional data.
- Limited interpretability: Some existing models may be difficult to interpret, making it challenging to identify which risk factors are most important or to understand how the model arrived at its predictions. To address this, new models can be developed using interpretable machine learning techniques, such as decision trees or rule-based models, or by incorporating domain knowledge or expert feedback into the model.
- Limited real-world applicability: Some existing models may not be easily integrated into clinical workflows or may require additional data or testing that is not feasible in real-world settings. To address this, new models can be developed with a focus on usability, simplicity, and scalability, and can be designed to be easily integrated into electronic health records or clinical decision support systems.
- Ethical and privacy concerns: Developing new diabetes risk prediction models may raise ethical and privacy concerns related to the collection, use, and sharing of patient data. To address these concerns, new models can be developed using transparent and accountable processes, with

appropriate informed consent and data protection measures in place. Additionally, the use of synthetic or de-identified data can be explored as a way to mitigate privacy risks while still allowing for model development and evaluation.

**Appendix C :' MEDICAL TERMS AND DEFINITIONS'**

**Polyuria** is a condition where the body urinates more than usual and passes excessive or abnormally large amounts of urine each time you urinate. Polyuria is defined as the frequent passage of large volumes of urine – more than 3 liters a day compared to the normal daily urine output in adults of about 1 to 2 liters.

**Polydipsia** is a medical name for the feeling of extreme thirstiness. Polydipsia is often linked to urinary conditions that cause you to urinate a lot. This can make your body feel a constant need to replace the fluids lost in urination. It can also be caused by physical processes that cause you to lose a lot of fluid.

**Genital Thrush (or candidiasis)** is a common condition caused by a type of yeast called Candida. It mainly affects the vagina, though may affect the penis too, and can be irritating and painful. Many types of yeast and bacteria naturally live in the vagina and rarely cause problems.

**Partial Paresis** Paresis involves the weakening of a muscle or group of muscles. It may also be referred to as partial or mild paralysis. Unlike paralysis, people with paresis can still move their muscles. These movements are just weaker than normal.

**Polyphagia** also known as hyperphagia, is the medical term for excessive or extreme hunger. It's different than having an increased appetite after exercise or other physical activity. While your hunger level will return to normal after eating in those cases, polyphagia won't go away if you eat more food.

## REFERENCES


[1] What is diabetes? (2023) Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/diabetes/basics/diabetes.html (Accessed: 15 May 2023)

[2] Health Information. (n.d.). Retrieved from NIHH: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[3] CDC. (n.d.). CDC. Retrieved from https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[4] Global Burden of Disease Collaborative Network. (2020). Retrieved from Istitute for health metrics and evaluation: https://vizhub.healthdata.org/gbd-results/

[5] Xue1, J., Min1, F., & Ma1, F. (2020, November 1). IOPscience. Journal of Physics: Conference Series. Retrieved May 2, 2023, from https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012062

[6] Classification models for likelihood prediction of diabetes at early stage using feature selection (no date) Applied Computing and Informatics. Available at: https://www.emerald.com/insight/content/doi/10.1108/ACI-01-2021-0022/full/html (Accessed: 15 May 2023).

[7] Ahmed, N., Ahammed, R., Islam, M. M., Uddin, A., Akhter, A., Talukder, A. H., & Amiri, I. S. (2021). Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering, 2, 229–241. https://doi.org/10.1016/j.ijcce.2021.12.001

[8] Butt, U.M. et al. (2021) Machine learning based diabetes classification and prediction for Healthcare Applications, Journal of healthcare engineering. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744/ (Accessed: 15 May 2023).


Link to GitHub: https://github.com/Utkarshika/606-Capstone