Diabetes Prediction Project Report

1. Introduction:

This report presents a data science project aimed at predicting the occurrence of type 2 diabetes based on various health indicators. The underlying question is to identify the key risk factors associated with diabetes and develop a predictive model that can aid in early detection and intervention. By leveraging regression and machine learning techniques, this project seeks to add value to both the healthcare industry and society by improving diabetes management and preventive strategies.

Question statement:

How can we use regression and machine learning techniques to predict whether an individual has diabetes based on their age, BMI, smoking history, heart disease, hypertension, and other indicators?

2. Background on the subject matter area:

Diabetes is a chronic metabolic disorder that affects millions of individuals worldwide. In United States, diabetes especially type 2 diabetes, is one of the most prevalent chronic diseases The International Diabetes Federation estimates that by 2045, at the current growth rate, 693 million people will have diabetes worldwide. According to the Centers for Disease Control and Prevention (CDC), in 2012, 29.1 million people in the United States were diagnosed with diabetes, making it the seventh leading cause of death in the country. Diabetes puts a high financial burden on the US economy. Studies show the total estimated cost of diagnosed diabetes increased to $327 billion in 2017, including $237 billion in direct medical costs and $90 billion in reduced productivity.

Early detection and accurate prediction of diabetes are crucial for timely intervention, implementing preventive measures, and improving overall health outcomes. This report presents the findings and outcomes of our capstone project focused on diabetes prediction using advanced regression and machine learning techniques to help facilitate early diagnosis and intervention and reduce medical costs.

3. Details on dataset:

The dataset used in this analysis was derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) data provided by the Centers for Disease Control and Prevention (CDC). The dataset comprised 438,693 records with 303 variables. To identify relevant predictors for diabetes, a meticulous feature selection process was conducted, involving domain experts, referencing a research paper on diabetes risk prediction, and consulting the BRFSS 2021 Codebook for variable descriptions.

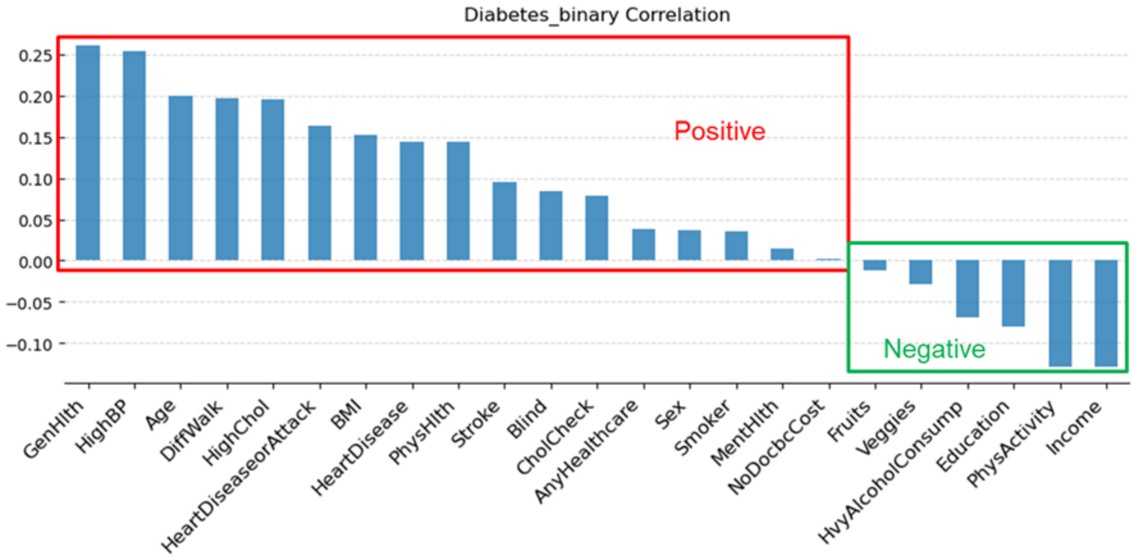
4. Summary of cleaning and preprocessing:

Data preprocessing involved handling duplicated rows, imputing missing values, and converting categorical variables into numerical format. Feature engineering techniques were employed to create new meaningful attributes and capture interactions between variables.

After data preprocessing, cleaning, and imputation of missing values, a cleaned dataset of 215,258 records and 24 informative variables was obtained. The selected features included diabetes status, age, BMI, smoking history, heart disease, hypertension, and other indicators related to health and lifestyle.

5. Insights, modeling, and results:

5.1 The exploration and analysis of the dataset revealed the following key findings:



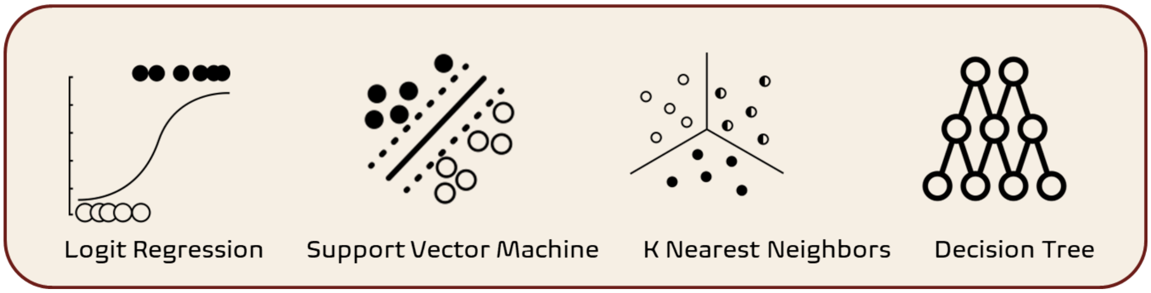
- The dataset analysis highlights the significant correlations between various health indicators and diabetes status. These findings can be used to identify potential risk factors and prioritize preventive measures.
- Significant positive Correlations: Several variables exhibited significant positive correlations with diabetes status. They are GenHlth, HighBP, Age, DiffWalk, HighChol, HeartDiseaseorAttack, BMI, PhysHlth, Stroke, Blind and CholCheck.
- Significant negative Correlations: HvyAlcoholconsump, Education, PhysActivity, and Income demonstrated strong negative correlation with diabetes status.
- Individuals with diabetes perceive a lower level of well-being, indicating the need for comprehensive support and management strategies to improve their quality of life.
- The strong correlation between diabetes and hypertension emphasizes the importance of integrated care approaches targeting both conditions for effective management and prevention.

- The concentration of individuals diagnosed with diabetes in the age group of 50 to 74 years suggests the need for targeted screening and intervention strategies for this age range.
- The higher cholesterol levels and incidence of cardiac conditions among individuals with diabetes underline.

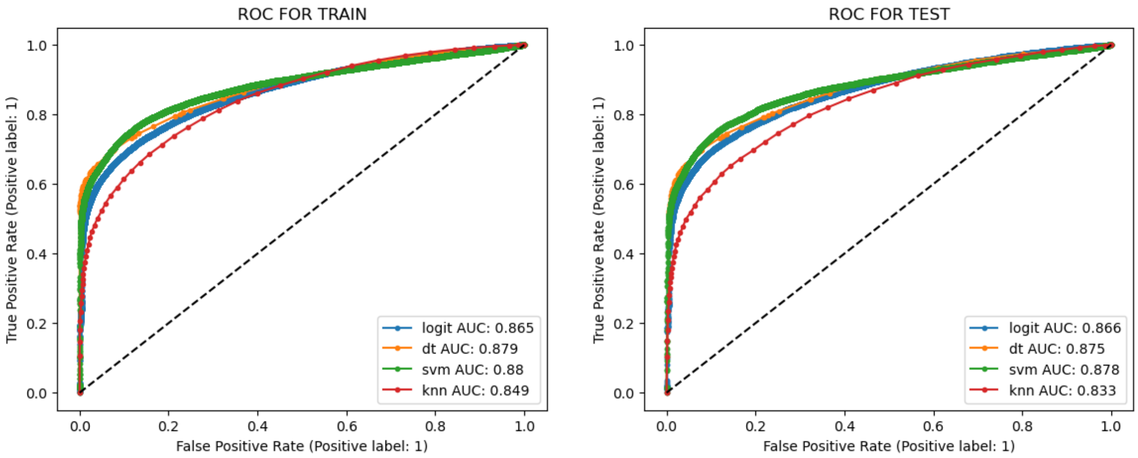
5.2 Modeling

To predict diabetes status, four regression models were employed: Logistic Regression, K-Nearest Neighbors, Decision Tree Classifier, and Support Vector Machine. The dataset was divided into training, validation, and test sets to evaluate model performance on unseen data. The accuracy of prediction for each model as follows:

Train: 79.3%	Train: 68.6%	Train: 75.6%	Train: 81.3%
Valid: 79.2%	Valid: 68.5%	Valid: 74.2%	Valid: 80.2%
Test: 79.3%	Test: 67.9%	Test: 74.4%	Test: 80.7%



Model evaluation metrics, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), were used to assess model effectiveness.



We built 4 machine learning models for predicting type 2 diabetes, including support vector machine, decision tree, logistic regression and K-Nearest Neighbors. We used univariable and multivariable weighted logistic regression models to investigate the associations of potential risk factors with type 2 diabetes.

6. Findings and conclusions:

6.1 Findings

The results of the analysis identified several key risk factors associated with diabetes, including general health, blood pressure, and age. By leveraging advanced data science techniques, this project demonstrated the potential to enhance diabetes prediction and improve public health outcomes. The practical value of the project lies in its ability to facilitate early detection, enabling targeted interventions and preventive measures. The predictive model can aid healthcare professionals in identifying high-risk individuals and designing personalized diabetes management plans.

The performance of each model on the test set was analyzed. The Decision Tree Classifier emerged as the best-performing model, achieving the highest accuracy and AUC-ROC among all models. However, the choice of the best model must consider the project's specific goals, as accuracy alone may not be the sole criterion.

6.2 Conclusions:

The diabetes prediction project has successfully leveraged regression and machine learning techniques to predict diabetes based on health indicators. The findings provide valuable insights into the factors influencing diabetes occurrence and contribute to early detection and prevention efforts. The Decision Tree Classifier demonstrated the best performance in predicting diabetes status, but the choice of the optimal model depends on the project's specific goals and context. Continuous improvements and rigorous validation are crucial for deploying predictive models as valuable tools in the fight against diabetes and for enhancing public health outcomes. The insights gained from this project can pave the way for further research and interventions to improve diabetes management and overall population health.