Predicting Diabetes: A Machine Learning Approach Using the CDC Diabetes Health Indicators Dataset

Advanced Machine Learning (DSC_540)
Nov 2024

Submitted By:

Rohith Perumandla

Professor:

Dr. Casey Bennett

DePaul University, Chicago

Table of Content:

1. Abstract

This project explores the application of machine learning techniques to predict and analyze diabetes using the CDC Diabetes Health Indicator Dataset. The dataset includes key health indicators and lifestyle survey information related to diabetes, such as demographics, physical activities, and daily habits. It consists of 250,000 instances, and preprocessing steps were applied to address data imbalance. The dataset was under-sampled to balance the minority class, resulting in approximately 35,000 instances per class. Feature selection was performed using the chi-square test with respect to the target feature and also through Recursive Feature Elimination (RFE) using Random Forest. Both methods identified the same top 15 features. Multiple machine learning classification algorithms, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and AdaBoost, were trained, evaluated using cross-validation, and compared for their predictive performance. Random Forest outperformed the other algorithms with an accuracy of 0.82 and an AUC of 0.81, followed by Gradient Boosting with an accuracy of 0.79 and an AUC of 0.78. In contrast, AdaBoost, SVM, Logistic Regression, and Decision Tree achieved accuracies below 0.75. The most important five features identified impacted the Random Forest model were GenHlth, HighBP, BMI, Age, and HighChol.

2. Introduction

In 2019, diabetes caused about 4.2 million deaths worldwide, making up 11% of all global deaths. In the U.S. alone, over 38 million people, or 10% of the population, have diabetes. According to the CDC, around 1.5 million new diabetes cases are diagnosed each year, making it the eighth leading cause of death. Diabetes is rising quickly worldwide. A CDC report estimates that 37.3 million people in the U.S., or 11.3% of the population, have diabetes. Of these, 28.7 million are diagnosed, while around 8.6 million have it but don't know it yet. In 2019, 283,000 children and teens under 20 had diabetes, with 244,000 having type 1 diabetes. Among adults over 20, 1.6 million (5.7% of adults with diabetes) reported having type 1 diabetes and using insulin. From 2014-2015, about 18,291 new cases of type 1 diabetes were reported each year in children and teens under 20. Diabetes is more often diagnosed in adults aged 45-64 and those 65 and older compared to younger adults aged 18-44 [7].

Diabetes is increasing not just among older people but also in teenagers. Despite modern technology that can track health, people's lifestyle and eating habits are contributing to the rise in diabetes. While there are tools available, they are not fully applied in real life. Understanding the prevalence and statistics shows the urgent need for effective strategies to detect, manage, and prevent diabetes. In the modern era of Artificial Intelligence, machine learning has become an integral part of various industries, including healthcare. This study leverages machine learning classification algorithms to classify diabetes and also identify and analyze the key factors influencing people's health that contribute to diabetes.

3. Literature Review

Diabetes is a chronic health condition that affects millions of individuals worldwide. Early diagnosis and intervention are crucial to prevent complications, making predictive modeling for easy access is an essential area of research. Machine learning techniques have been widely adopted to classify diabetes risk, leveraging diverse datasets and features including laboratory tests and different lifestyle data. This review synthesizes existing research to identify trends, gaps, and opportunities for improvement.

Qin et al. [1], The study aimed to predict diabetes using machine learning models, specifically CATBoost, XGBoost, Random Forest, Logistic Regressionand Support Vector Machine (SVM), based on lifestyle-related variables. CATBoost achieved the highest performance with an AUC of 0.83 and an accuracy rate of 82.1%. The ranking of the models based on performance was as follows: CATBoost > RF > LR > XGBoost > SVM, indicating that CATBoost is the most effective model for diabetes prediction among the ones tested.

Mujumdar et al. [2] discussed existing research on diabetes prediction, noting that the performance of previous models was not particularly high. They addressed different types of diabetes, including Type 1, Type 2, and Type 3, and incorporated external factors such as age, BMI, glucose levels, and other laboratory data. Using two different datasets, they applied 12 different classification algorithms and achieved an average accuracy of 90%, significantly boosting predictive performance.

Gupta et al [3], authors used Pima India Diabetes dataset from UCI repository with around 900 instances and used Naive bayes and Support Vector Machine algorithms to train the model with feature selection and cross-valdiation and got accuracy 79% and 81% respectively.

Kaviyaadharshani et al. [4] provided an extensive literature review on research and studies related to diabetes prediction using machine learning applications. The paper compared eight studies and their predictive performance on various datasets, including Pima, UCI, and others. The authors concluded that ensemble methods outperformed other algorithms in these studies.

Orabi et al. [5] designed a system for diabetes prediction. The main goal of the study was to predict diabetes in individuals at a particular age. The proposed study was based on machine learning, primarily utilizing a decision tree, which produced good results and performed well in predicting diabetes at specific ages.

Ren et al. [6] worked on predicting diabetes using the Diabetes Health Indicator Dataset collected by BRFSS, sourced from Kaggle, which is different from the dataset used in this project but with same features. Feature selection was performed using the Chi-square test, and the dataset imbalance was addressed using the SMOTE method to generate more instances for the minority class (diabetes). They achieved a test accuracy of 0.866 and a training accuracy of 0.92 with the CatBoost algorithm, as well as a test accuracy of 0.84 and a training accuracy of 0.99 with Random Forest. These results indicate that their models were overfitting the data.

After reviewing the literature, I observed that most studies have employed ensemble methods such as Random Forest, and Gradient Boosting. Additionally, many of these studies focused on datasets with fewer lifestyle-related features. I believe that using lifestyle data, such as that provided by the CDC Diabetes Health Indicator in the US, offers valuable insights for analyzing and predicting diabetes. By utilizing this lifestyle survey data, individuals can better understand how their daily habits affect their health and make informed decisions to improve their overall well-being. Furthermore, this approach allows for easier prediction of diabetes risk based on lifestyle factors."

4. Methodology:

The system architecture shown in figure 1 represents the machine learning pipeline for diabetic data analysis and prediction followed in this study. The process begins with the collection of diabetic data, which is then preprocessed by identifying null values, removing or imputing them, and eliminating duplicate values, then checking the dataset imbalance. The data is then prepared for Exploratory Data Analysis (EDA). EDA is conducted to gain a better understanding of the data and identify patterns. Next, feature engineering is performed to create and standardize the relevant features for model training. Feature selection is also carried out to select the significant features that impact the model using different feature selection techniques. The data is split into training(80%) and test(20%) sets. Machine learning classification algorithms are applied to the training data to build the model, and cross-validation is used to validate the model. Hyperparameter tuning is performed to optimize the model and achieve the best results without overfitting with the training data itself. Then after the hyper parameters are finalized. The final models are then evaluated using the test data. The output achieved with unseen test data are the final performance metrics for those models.

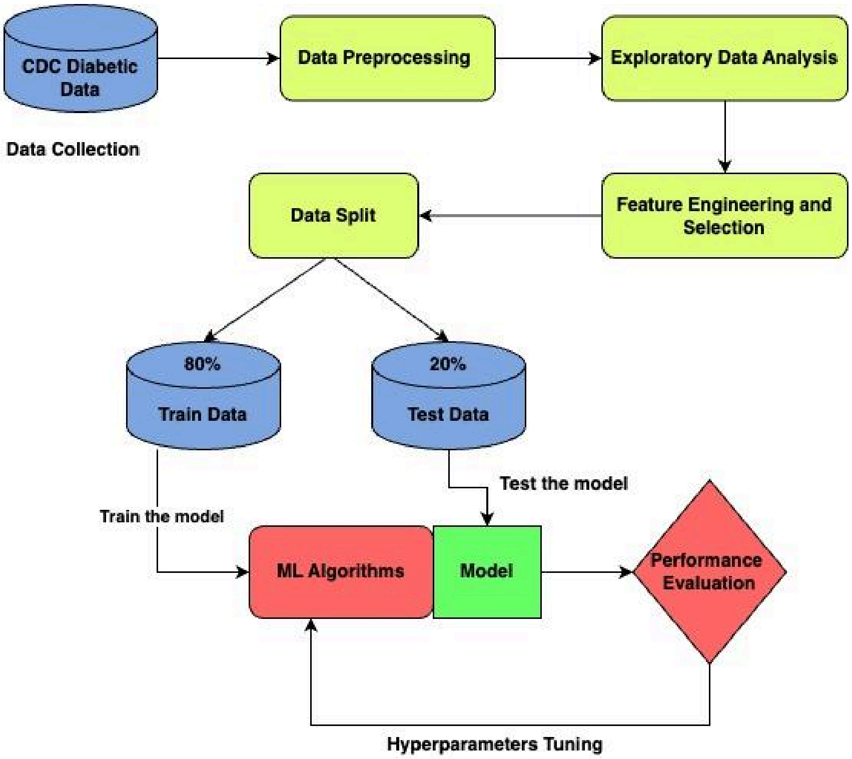


Figure 1

4.1 Dataset:

The dataset, sourced from the UCI Machine Learning Repository, consists of 253,000 instances and 21 features, which include both demographic information and lifestyle factors related to diabetes. The dataset is designed to improve understanding of the relationship between lifestyle and diabetes in the United States. The key features in the dataset includes:

1. **Demographics:** 'Sex', 'Age', 'Education', 'Income'
2. **Health Indicators:** 'HighBP', 'HighChol', 'CholCheck', 'BMI', 'Stroke', 'HeartDiseaseorAttack', 'GenHlth',, 'PhysHlth',
3. **Lifestyle Factors:** 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'Smoker', 'DiffWalk', 'MentHlth'

Features Description:

1. **Diabetes_binary** → 0 = no diabetes 1 = pre diabetes for diabetes
2. **HighBP** → 0 = no high BP 1 = high BP
3. **HighChol** → 0 = no high cholesterol 1 = high cholesterol
4. **CholCheck** → 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
5. **BMI** → Body Mass Index
6. **Smoker** → Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
7. **Stroke** → (Ever told) you had a stroke. 0 = no 1 = yes
8. **HeartDiseaseorAttack** → coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
9. **PhysActivity** → physical activity in past 30 days - not including job 0 = no 1 = yes
10. **Fruits** → Consume Fruit 1 or more times per day 0 = no 1 = yes
11. **Veggies** → Consume Vegetables 1 or more times per day 0 = no 1 = yes
12. **HvyAlcoholConsump** → Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
13. **AnyHealthcare** → Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14. **NoDocbcCost** → Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
15. **GenHlth** → Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
16. **MentHlth** → Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
17. **PhysHlth** → Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
18. **DiffWalk** → Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes'
19. **Sex** → 0 = female 1 = male
20. **Age** → 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older

21. **Education** → Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
22. **Income** → Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

4.2 Data Preprocessing

In the preprocessing stage of the dataset, various steps were taken to prepare the data for analysis. The diabetes dataset consists of 21 numerical features, 0 categorical features(all categorical features are already encoded), 7 continuous features, and 14 binary features. Additionally, there are 7 features with more than two unique values. The first step in the data cleaning process involved identifying and removing duplicate rows, with a total of 25,772 duplicate instances found and the dataset does not contain any NULL values, ensuring a clean set of data for further analysis. The dataset initially showed a high class imbalance, with the majority class (no diabetes) accounting for 218,334 instances, or 86.07% of the data, while the minority class (diabetes) had only 35,346 instances, or 13.93%. To address this, the majority class was downsampled during preprocessing to align with the size of the minority class. This method was chosen because the minority class had ample data (35,346 instances) for effective model training. Downsampling helps reduce bias toward the majority class and ensures the model can predict both classes accurately.

4.3 Exploratory Data Analysis

In the exploratory data analysis step, various visualizations have been done to understand, explore the dataset and its underlying patterns. Some of the noticeable patterns and figures are added to understand the data. First, correlation analysis was done and plotted a heatmap of all the features as shown in Fig 2. The correlation analysis showed that PhysHlth and DiffWalk are moderately positively correlated with GenHlth at 0.52 and 0.46, respectively. Additionally, DiffWalk and PhysHlth have a moderate positive correlation of 0.48, while Income and Education are correlated at 0.45. These are the highest correlations observed among the variables, but since none exceed the threshold of 0.8, none of them are removed based on the correlation analysis.

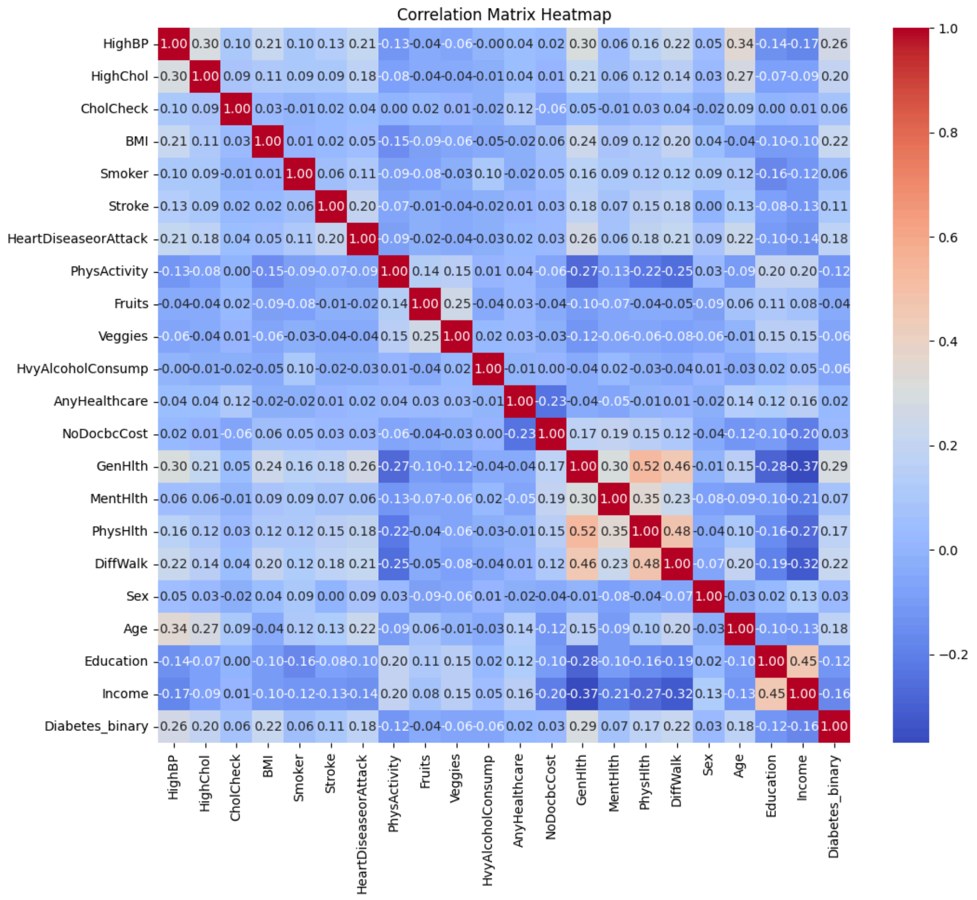


Figure 2

The fig 2 illustrates the distribution of diabetes classes (0 = No Diabetes, 1 = Diabetes) by age. It shows that the occurrence of diabetes generally increases with age, particularly in age groups 8 to 10, where the diabetes class becomes more prevalent. The imbalance between the two classes is visible, especially in younger age groups, where the "No Diabetes" class dominates. The fig 3 graph shows how diabetes is distributed between genders. It indicates that males (1) are more likely to have diabetes compared to females (0). However, for the "No Diabetes" group, the difference between males and females is smaller. This analysis helps in understanding how gender plays a role in diabetes patterns in the dataset.

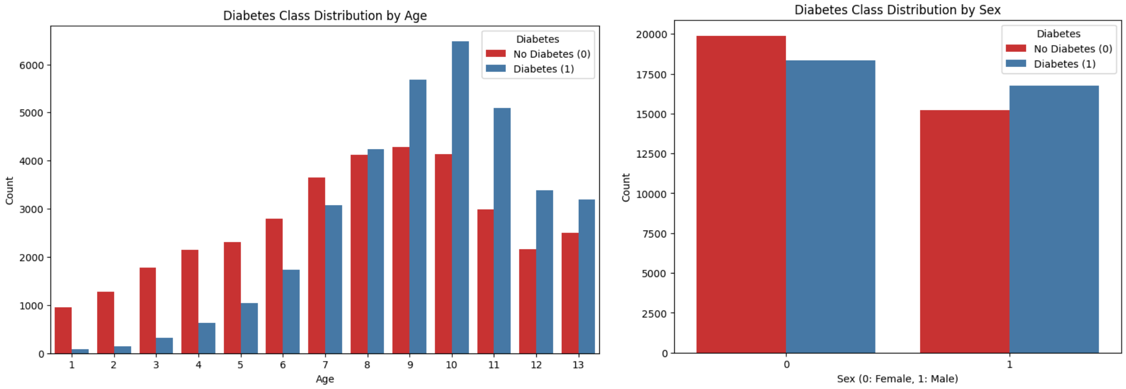


Figure 3

Inorder to understand the data distribution of the data and find the outliers, box plots and distribution plots are plotted. The box plot in fig 3 visual compares BMI values for individuals with and without diabetes. Both groups show similar ranges, but the median BMI is slightly higher for the diabetes class (1) compared to the no diabetes class (0). Outliers are present in both groups, particularly at higher BMI values, indicating variability in BMI among individuals. Those identified outliers are removed from the dataset. The graph in fig 4 shows the distribution of BMI for both classes. The diabetes class (orange) generally has a higher BMI compared to the no diabetes class (blue), with a noticeable overlap. However, individuals with diabetes are more concentrated in the higher BMI range, suggesting a relationship between elevated BMI and diabetes.

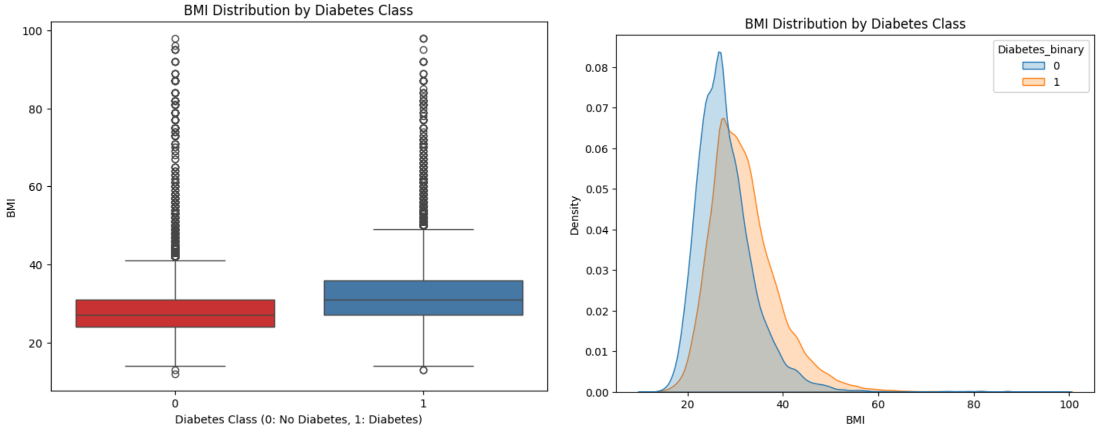


Figure 4

To understand the diabetes class distribution by income, a graph plotted to show the distribution of diabetes cases across different income levels as shown in fig 5. As income increases, the number of individuals without diabetes rises significantly, especially at the highest income level (8). In comparison, the number of individuals with diabetes is more evenly spread across lower income levels but decreases as income increases. This suggests that higher income levels may be associated with better health outcomes, possibly due to improved access to healthcare and healthier lifestyle choices.

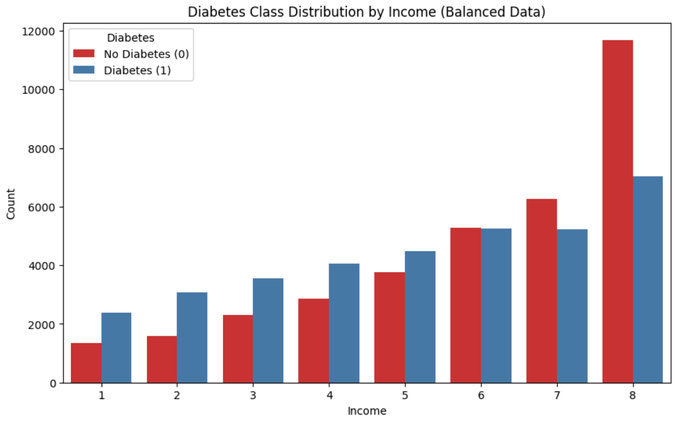
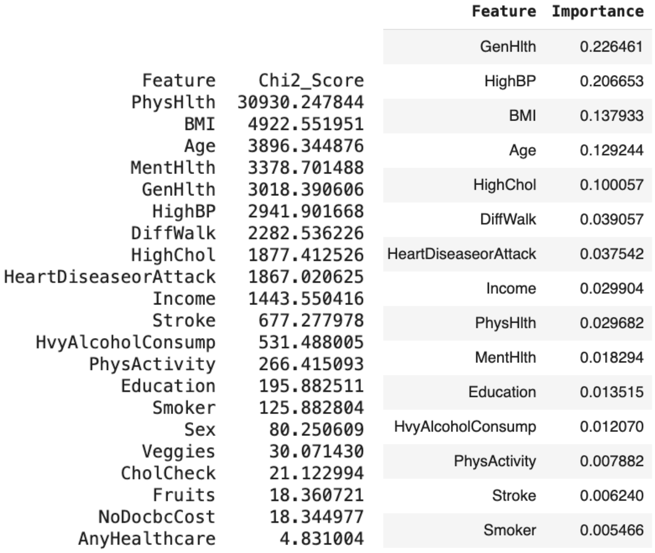


Figure 5

4.4 Feature Engineering and Selection

In the dataset, as we discussed in the preprocessing section, we have different ranges of values in each feature, like binary, numerical. So in order to train the model and don't show any bias on high valued features. Standardization has been done using StandardScaler library from scikit-learn preprocessing.

Feature Selection: In this dataset, there are 20 features, I used different feature selection methods like chi-square test and selected top 15 features with high scores and the recursive feature elimination method using Random Forest. In the below fig 6 the table shows all features and its respective chi square score only top 15 are selected from the table. The selected features are 'PhysHlth', 'BMI', 'MentHlth', 'Age', 'GenHlth', 'HighBP', 'DiffWalk', 'HighChol', 'HeartDiseaseorAttack', 'Income', 'Stroke', 'HvyAlcoholConsump', 'PhysActivity', 'Education', 'Smoker'



```
             Feature    Chi2_Score
            PhysHlth  30930.247844
                 BMI   4922.551951
                 Age   3896.344876
            MentHlth   3378.701488
             GenHlth   3018.390606
              HighBP   2941.901668
            DiffWalk   2282.536226
            HighChol   1877.412526
HeartDiseaseorAttack   1867.020625
              Income   1443.550416
              Stroke    677.277978
   HvyAlcoholConsump    531.488005
        PhysActivity    266.415093
           Education    195.882511
              Smoker    125.882804
                 Sex     80.250609
             Veggies     30.071430
           CholCheck     21.122994
              Fruits     18.360721
         NoDocbcCost     18.344977
        AnyHealthcare      4.831004
```

Feature	Importance
GenHlth	0.226461
HighBP	0.206653
BMI	0.137933
Age	0.129244
HighChol	0.100057
DiffWalk	0.039057
HeartDiseaseorAttack	0.037542
Income	0.029904
PhysHlth	0.029682
MentHlth	0.018294
Education	0.013515
HvyAlcoholConsump	0.012070
PhysActivity	0.007882
Stroke	0.006240
Smoker	0.005466

Figure 6

4.5 Model Training:

Model training is the process of teaching a machine learning algorithm to recognize patterns and make predictions by providing it with labeled data. During training, the model learns to map input features to the target variable by minimizing a defined loss function. This step is crucial to ensure the model generalizes well to unseen data for accurate predictions. The dataset is split into training and test sets in an 80:20 ratio. The 80% training data is then used to train several machine learning algorithms using cross-validation to evaluate generalizability. Additionally, GridSearchCV is applied to iterate through various hyperparameters for tuning and to identify the best model.

The machine learning classifiers trained on the diabetic dataset are Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Ensemble methods like Random Forest, Gradient Boosting, Adaboosting.

Each model is trained using cross-validation to evaluate its generalizability and ensure it performs well. The test set is then used to evaluate the final model, which is trained with the best hyperparameters, on unseen data. The performance of each model is summarized below in fig 7. To provide a comprehensive interpretation, various performance metrics such as Accuracy, Area Under the Curve (AUC) score, Precision, Recall, and F1-Score are calculated.

Algorithms	**CV accuracy**	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.819	0.822	0.812	0.791	0.803	0.810
Gradient Boosting	0.788	0.791	0.782	0.769	0.799	0.798
AdaBoosting	0.745	0.742	0.742	0.732	0.759	0.745
Logistic Regression	0.729	0.731	0.735	0.732	0.741	0.725
SVM	0.732	0.733	0.729	0.742	0.753	0.747
Decision Tree	0.721	0.727	0.727	0.704	0.777	0.739

Figure 7

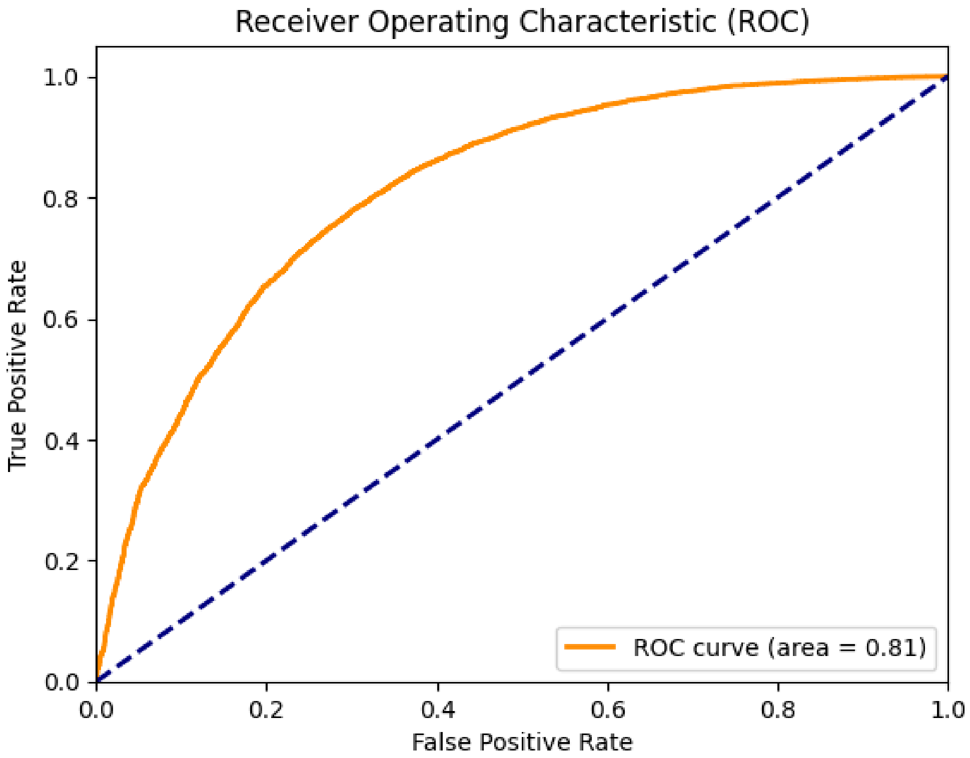


Figure 8

5. Results:

The performance of different machine learning algorithms was evaluated using several metrics, including Accuracy, AUC, Precision, Recall, and F1-Score. Among the algorithms tested, Random Forest achieved the highest overall performance, with an Accuracy of 0.822, an AUC of 0.812, a Precision of 0.791, a Recall of 0.803, and an F1-Score of 0.810. These results demonstrate that Random Forest performed well in all metrics, providing a good balance between precision and recall. The ROC curve with AUC score of 0.81 as shown in the fig 7 indicates that the model has a strong ability to distinguish between positive and negative classes. With a value closer to 1, it suggests that the model performs well in identifying true positives while minimizing false positives. AUC of 0.81 demonstrates good performance in predicting whether diabetic or not using the selected 15 features, though there is still room for improvement in the model's ability to perfectly separate the classes.

Gradient Boosting showed slightly lower performance, with an Accuracy of 0.791 and an AUC of 0.782. Its Precision and Recall were 0.769 and 0.799, respectively, resulting in an F1-Score of 0.798. While it performed well, especially in terms of Recall, it couldn't match Random Forest in overall metrics. AdaBoosting displayed a lower performance across the board, with an Accuracy of 0.742, an AUC of 0.742, and a Precision of 0.732.

In contrast, Decision Tree, SVM and Logistic Regression exhibited the lowest scores of all the algorithms tested. These algorithms struggled to match the performance of the other more complex algorithms like Random Forest, Gradient Boosting.

The top 5 most important features are used by Random Forest model are GenHlth, HighBP, BMI, Age and HighChol as shown in fig 6. The comparison between the cross-validation mean accuracy scores and the accuracy obtained from the test data clearly indicates that the models are not overfitting. When a model overfits, it typically performs well on the training data but fails on the test data. Additionally, during cross-validation, an overfitting model would yield significantly lower scores since it trains on (k-1) folds and tests on the kth fold.

6. Discussion:

The CDC diabetic health data mainly consists of categorical data, such as sex, age groups, income groups, and several binary features. The selected features include 'PhysHlth', 'BMI', 'MentHlth', 'Age', 'GenHlth', 'HighBP', 'DiffWalk', 'HighChol', 'HeartDiseaseorAttack', 'Income', 'Stroke', 'HvyAlcoholConsump', 'PhysActivity', 'Education', and 'Smoker'. Random Forest is effective for categorical data because it uses an ensemble of decision trees as base models. Decision trees are particularly good at handling categorical data, as they split nodes based on the different categories, choosing the best split to maximize information gain or reduce impurity. The other algorithms like Gradient Boosting aslo performed well on the dataset. But can be improved more inorder to use those models in real world application. The other models like SVM, Logistic Regression and Decision Tree didn’t perform well compared to the Ensemble methods. The reason could be that the dataset contains significantly more binary features than continuous numerical features. This might explain why ensemble methods, which combine basic models to enhance performance, work well for this problem. Incorporating more numerical features could potentially improve the models' performance further.

There are numerous research papers addressing the problem of diabetes prediction using supervised machine learning algorithms. Most of these studies rely on datasets that include basic demographic information such as sex, age, education, location, and income, as well as extensive laboratory test data. While this data is highly useful and predictive, I conducted my study using a dataset sourced from the CDC Diabetes Health Indicators in the U.S., available from the UCI Machine Learning Repository. This dataset includes 21 features, with over 17 of them related to daily habits and lifestyle data.

I believe that every decision we make impacts our health, either positively or negatively. Using lifestyle data rather than laboratory results could significantly help individuals make informed daily decisions to maintain their health. Additionally, I chose this dataset because of its focus on lifestyle survey information. When a model performs exceptionally well with such data, it can be deployed in real-world applications, such as mobile or web apps. These apps would allow individuals to input their lifestyle information and receive feedback on whether they are diabetic, prediabetic, or non-diabetic. Moreover, such models could guide people on which aspects of their daily lifestyle to improve, helping them avoid diabetes in the future and promoting better overall health.

7. Conclusion:

This machine learning approach to addressing the diabetes problem using lifestyle survey data made significant progress by testing various algorithms, including Random Forest, Gradient Boosting, AdaBoost, SVM, Logistic Regression, and Decision Tree. Among these, Random Forest outperformed the others, achieving an accuracy of 0.82 and an AUC of 0.81. This study has the potential to help many individuals who are unaware of their health and diabetes status, especially those who cannot afford regular checkups or who wish to improve their health. The study's focus on lifestyle survey information makes it particularly accessible and practical.

By developing an app where users can answer a series of questions based on these features, we could assess their likelihood of being diabetic. The app could also highlight the most influential factors contributing to their risk, empowering users to take proactive steps toward better health management.

This would be especially valuable for people who are often unaware of their diabetes status and only discover it when the condition becomes severe. By leveraging models that analyze daily habit data, we can increase awareness and promote early intervention, helping individuals effectively monitor and maintain their health.

8. Future Work:

The findings of this study demonstrated strong performance in predicting diabetes based on lifestyle features. I used the Chi-square test and recursive methods for feature selection; however, more advanced feature selection techniques could further improve the model's performance. The initial dataset consisted of approximately 250,000 instances, with 86.07% labeled as non-diabetic and 13.97% as diabetic, resulting in a data imbalance. To address this, I applied downsampling, reducing the dataset to 35,346 instances per class, resulting in a total of 70,692 rows, which I considered sufficient for training.

After training, the model achieved an AUC of 0.81 and an accuracy of 0.82, effectively distinguishing between diabetic and non-diabetic cases. However, there is potential to enhance the model's performance further. In the future, instead of downsampling the data from 250,000 to 70,692 instances, upsampling techniques like SMOTE could be used to generate more instances for the diabetes class. This approach could help the classification model capture more underlying patterns in the data.

Deep learning techniques are well-suited for addressing complex problems and datasets. Applying different deep learning algorithms may uncover more intricate patterns in the data, potentially boosting model performance. I also experimented with Principal Component Analysis (PCA), but the results were not satisfactory. This may be due to PCA identifying linear combinations of features that maximize variance, which assumes continuous data. Since binary features have limited variance (0 or 1), PCA may not perform well in this context. In the future, a more detailed analysis and careful implementation of PCA could be explored to improve the model's performance.

Many features related to daily habits and lifestyle data significantly contribute to individuals developing diabetes and pre-diabetes. There is substantial ongoing research and evidence that supports the creation of new features from existing ones. For instance, features such as difficulty walking, mental health, and alcohol consumption could be combined, as they are interconnected and might collectively influence diabetes risk. There are around 20 such features that can potentially be transformed into new, more informative features to enhance the model's ability to distinguish between diabetic and non-diabetic individuals using only lifestyle data.This is an area that can be explored and improved upon in future work. Therefore, these are the ideas that I believe could help improve the model's performance for this diabetes problem.

9. References

1. Qin, Yifan, Jinlong Wu, Wen Xiao, Kun Wang, Anbing Huang, Bowen Liu, Jingxuan Yu, Chuhao Li, Fengyu Yu, and Zhanbing Ren. 2022. "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type" International Journal of Environmental Research and Public Health 19, no. 22: 15027. https://doi.org/10.3390/ijerph192215027
2. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science, 165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047
3. Gupta, S., Verma, H. K., & Bhardwaj, D. (2020). Classification of Diabetes Using Naïve Bayes and Support Vector Machine as a Technique. In Lecture notes on multidisciplinary industrial engineering (pp. 365–376). https://doi.org/10.1007/978-981-15-6017-0_24
4. Kaviyaadharshani, D., Nivedhidha, M., Jeyarohini, R., Rani, J. L. E., Ramkumar, M. P., & Selvan, G. S. R. E. (2024). Diagnosing Diabetes using Machine Learning-based Predictive Models. Procedia Computer Science, 233, 288–294. https://doi.org/10.1016/j.procs.2024.03.218
5. Orabi, K.M., Kamal, Y.M., Rabah, T.M. (2016). Early Predictive System for Diabetes Mellitus Disease. In: Perner, P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2016. Lecture Notes in Computer Science(), vol 9728. Springer, Cham. https://doi.org/10.1007/978-3-319-41561-1_31
6. Ren, Xinyi. (2024). Predictions of diabetes through machine learning models based on the health indicators dataset. Applied and Computational Engineering. 32. 216-222. 10.54254/2755-2721/32/20230214.https://www.ewadirect.com/proceedings/ace/article/view/9942

7. Centers for Disease Control National Diabetes Statistics Report 2022; National Institutes of Health **https://www.cdc.gov/diabetes/php/data-research/index.html**