COSC2753 - Machine Learning

Assignment 1 - Individual

Seokyung Kim s3939114

"I declare that in submitting all work for this assessment I have read, understood, and agree to the content and expectations of the Assessment declaration"

I. Project Overview

Diabetes, as defined by the World Health Organization (WHO), is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces [1]. The dataset provided contains a diverse range of health metrics and demographic attributes of individuals, with the target variable being the presence or absence of diabetes.

This project aims to leverage machine learning techniques to predict the onset of diabetes in individuals based on a set of health metrics and demographic information. As a student undertaking this assignment, the primary objectives include proposing, implementing, and evaluating machine learning models to accurately forecast the status of diabetes development in patients.

II. Problem Statement

A. Feature Description

Feature Name	Data Type	Description	Valid Range
Id	String	Unique identifier for each patient	
Status	Integer	Diabetic status indicator	0 = No diabetes, 1 = Prediabetes or diabetes
HighBP	Integer	High blood pressure indicator	0 = No high BP, 1 = High BP
HighChol	Integer	High cholesterol indicator	0 = No high cholesterol, 1 = High cholesterol
CholCheck	Integer	Cholesterol check within 5 years indicator	0 = No check, 1 = Check
BMI	Float	Body Mass Index measurement	0 = No, 1 = Yes
Smoker	Integer	Smoking history indicator	0 = No, 1 = Yes
HeartDiseaseorAttack	Integer	Coronary heart disease or myocardial infarction indicator	0 = No, 1 = Yes
PhysActivity	Integer	Physical activity indicator	0 = No, 1 = Yes
Fruits	Integer	Daily fruit consumption indicator	0 = No, 1 = Yes
Veggies	Integer	Daily vegetable consumption indicator	0 = No, 1 = Yes
HvyAlcoholConsump	Integer	Heavy alcohol consumption indicator	0 = No, 1 = Yes
AnyHealthcare	Integer	Healthcare coverage indicator	0 = No, 1 = Yes
NoDocbcCost	Integer	Inability to see a doctor due to cost indicator	0 = No, 1 = Yes
GenHlth	Integer	General health rating	1 = Excellent, 2 = Very good, 3 = Good, 4 = Fair, 5 = Poor
MentHlth	Integer	Days of poor mental health in past 30 days	1-30 days
PhysHlth	Integer	Days of poor physical health in past 30 days	1-30 days
DiffWalk	Integer	Difficulty walking or climbing stairs indicator	0 = No, 1 = Yes
Sex	Integer	Gender indicator	0 = Female, 1 = Male
Age	Integer	Age category	1 = 18-24, 9 = 60-64, 13 = 80 or older
Education	Integer	Education level indicator	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)
Income	Integer	Income level indicator	1 = Less than $10,000, 5 = Less than $35,000, 8 = $75,000 or more
ExtraMedTest	Float	Result of an extra medical test	-100 to 100
ExtraAlcoholTest	Float	Result of an extra medical test	-100 to 100

Figure 1. Feature Description Table

B. Procedure

1. **Exploratory Data Analysis (EDA)**: We analyze the dataset to understand the distribution of features, identify correlations, and visualize patterns in the data.
2. **Data Preprocessing**: We handle missing values, outliers, and perform necessary data transformations to prepare the dataset for modeling.
3. **Model Development**: We build machine learning models using the prepared dataset to predict the onset of diabetes based on health metrics and demographic information. We evaluate multiple models and select the one that demonstrates the best performance in predicting diabetes development, considering F1-score.
4. **Model Evaluation**: We predict the status of diabetes for each individual in the preprocessed test data sets.

C. Evaluation Metrics and Criteria

In this project, we assess our machine learning model using two key metrics: accuracy and F1 score. While accuracy measures correct classifications, it might mislead with imbalanced data, like medical diagnoses. F1 Score offers a balanced view, crucial for tasks with class imbalances or differing costs of false positives and false negatives, like medical diagnoses.

By prioritizing F1 score, we aim to strike a balance between minimizing both types of errors, recognizing the criticality of accurate diabetes predictions and the potential implications of misclassifications in healthcare settings. Our performance improvement criteria aim for an **F1 score of 87~92%**.

Additionally, we will utilize ROC Curve and AUC as supplementary metrics to provide additional insights into model performance, recognizing their value in understanding the trade-offs between sensitivity and specificity, although our primary focus remains on the F1 score due to its relevance to the task and dataset characteristics.

III. EDA

A. Overview of Datasets

Data Characteristics:

- **Train Dataset**: Contains 202944 rows and 25 columns.
- **Test Dataset**: Consists of 50736 rows and 25 columns.

Missing Values:

- Both datasets exhibit no missing values across all columns.

Duplicated Rows without the 'Id' feature:

- **Train Dataset**: Identifies 208 duplicate rows.
- **Test Dataset**: Reveals 14 duplicate rows.

Out of Range Values:

- Outliers, detected during descriptive statistics, are observed in both datasets.
- Specifically, the 'ExtraMedTest' and 'ExtraAlcoholTest' features display values beyond the defined range [-100, 100].
- Implications include potential data anomalies affecting analysis or modeling, requiring careful review or treatment as missing data.

B. Data Distribution

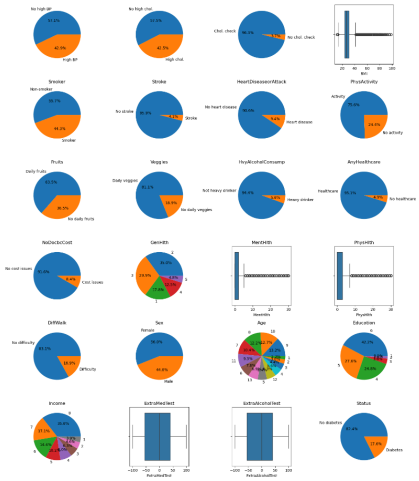
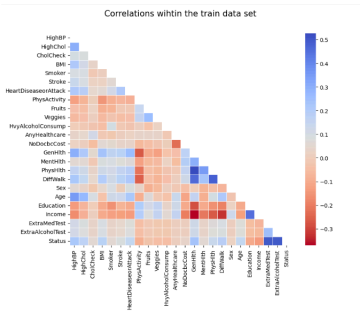


Figure 2. Data Distribution with Bar Charts **Figure 3. Data Distribution with Pie Charts and Box Plots**

The overall tendency in the datasets indicates:

- **Relationship between the train and test datasets:** There is a strong similarity in distributions and characteristics between the train and test datasets, indicating a comparable relationship between the two datasets. This consistency is valuable for modeling and making inferences about the target population based on the train dataset.
- **Health Indicators**: Majority of individuals exhibit favorable health indicators such as normal blood pressure, cholesterol levels, and BMI. Most individuals also report no history of stroke, heart disease, or heavy alcohol consumption. Access to healthcare services is prevalent among the majority, with cost not being a significant barrier.
- **Lifestyle Habits**: A considerable portion of individuals engage in regular physical activity and consume fruits and vegetables daily, reflecting positive lifestyle choices.
- **Demographic Distribution**: The dataset skews slightly towards females, with a predominant representation of college-educated individuals and those with higher income levels.
- **Age Distribution**: The dataset is multimodal, representing various age groups, with a notable concentration around middle-aged individuals.
- **Diabetes Status**: Around 80% of individuals in the dataset do not have diabetes, while approximately 20% do.

C. Feature Correlations



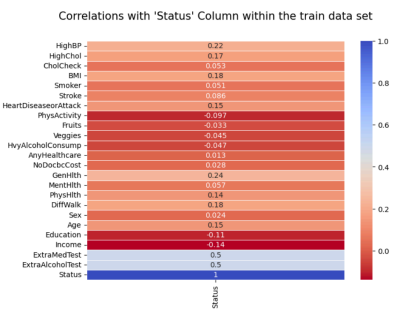


Figure 4. Correlations Between Every Feature
Figure 5. Correlations Between 'Status' and Others

The 'Status' column shows strong positive correlations (0.18 to 0.5) with 'ExtraMedTest', 'ExtraAlcoholTest', 'GenHlth', 'HighBP', 'BMI', and 'DiffWalk'. These suggest higher scores on medical and alcohol tests, better general health, high blood pressure, higher BMI, and difficulty in walking may indicate a positive 'Status', possibly indicating prediabetes or diabetes.

Conversely, 'Income' (-0.14), 'Education' (-0.11), and 'PhysActivity' (-0.097) exhibit weak negative correlations with 'Status'. This implies that factors like income, education level, and recent physical activity levels may have a weaker influence on prediabetes or diabetes status.

IV. Data Preprocessing

1. **Column Dropping**: We dropped the 'Id' column from both the train and test datasets.
2. **Remove Duplicates**: We removed duplicated rows, excluding the 'Id' column, from both datasets.
3. **Handle invalid values**: We opted for mean imputation instead of winsorization due to its ability to maintain dataset size and structure while handling extreme values, contrasting with winsorization which may introduce bias and distort the data distribution.
4. **Handle Outliers**: We utilized IQR (Interquartile Range) and Capping together to address outliers effectively. IQR is robust against extreme values and provides a measure of data spread, while capping ensures extreme values are bounded within a reasonable range, maintaining data integrity and model stability. This combined approach offers a balanced strategy for outlier treatment, enhancing the reliability of our analysis.
5. **Feature Scaling**: We chose min-max normalization over standardization for its specific characteristics. Normalization scales numerical features to a range between 0 and 1, aligning them with binary features. This ensures equitable contributions from all features during modeling, promoting balanced analysis. In contrast, standardization scales features to have a mean of 0 and a standard deviation of 1, which might not align well with binary features and could potentially skew the data distribution.
6. **Handle Imbalanced Classes**: We chose SMOTE over random oversampling and undersampling after comparing techniques. SMOTE generates synthetic samples by interpolating between minority class instances, effectively capturing the distribution and reducing the risk of overfitting. This enhances model generalization and robustness in predicting patients' 'Status'.

V. Model Development

We have experimented with 5 different machine learning models that are suitable for predicting the status of diabetes development in patients: **Logistic Regression (Linear Features)**, **Logistic Regression (Polynomial Features with degree 1)**, **Logistic Regression (Polynomial Features with degree 2)**, **Decision Tree**, **Random Forest**.

- We used **K-fold cross-validation** with n_splits 4 and random_state 0 for every model.
- **Feature selection via RFE** was applied with 6 selected features: 'HighBP', 'CholCheck', 'BMI', 'HvyAlcoholConsump', 'ExtraMedTest', and 'ExtraAlcoholTest'.
- **Logistic regression model (linear)** was developed with liblinear solver, L2 regularization, and Regularization strength parameter 'C' set to 1.
- **Logistic regression polynomial degree 1 and 2 models** were developed with selected features, liblinear solver, L2 regularization, and 'C' values of 0.58 and 3.745, respectively.
- **Decision tree model** was developed with the selected features and parameter distributions: 'max_depth': None, 'max_features': 10, 'min_samples_leaf': 6, 'min_samples_split': 45, using RandomizedSearchCV for hyperparameter tuning.
- **Random forest model** was developed with the selected features and parameter distributions: 'max_depth': 19, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 18, 'n_estimators': 41, using RandomizedSearchCV for hyperparameter tuning.

	Train F1	Val F1	Train AUC	Val AUC
Logistic Linear	0.80	0.80	0.80	0.80
Logistic Poly 1	0.79	0.78	0.79	0.78
Logistic Poly 2	0.79	0.78	0.79	0.78
Decision Tree	0.93	0.91	0.93	0.92
Random Forest	0.92	0.91	0.92	0.91

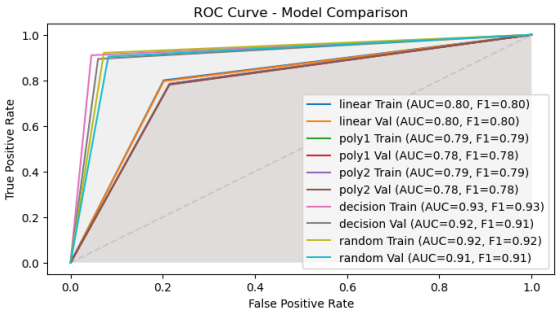


Figure 6. Table for F1-Score and AUC of Each Model
Figure 7. ROC Curve Comparison Between All Models

We went through various models, focusing on their F1 scores within the desired range of 87% to 92%. All the Logistic Regression models' performances were moderate, indicating underfitting.

The Decision Tree Model showed strong performance, especially on the training dataset, but there was a slight overfitting issue, as indicated by the difference between training and validation scores. Similarly, the Random Forest Model performed well on both datasets with minimal overfitting compared to decision trees and aligned closely with the target F1 score range.

Considering these factors, **the Random Forest model**, with F1 scores of 0.92 and 0.91 for training and validation sets respectively, was selected as the final choice due to its robust performance, minimal overfitting, and alignment with the desired F1 score range.

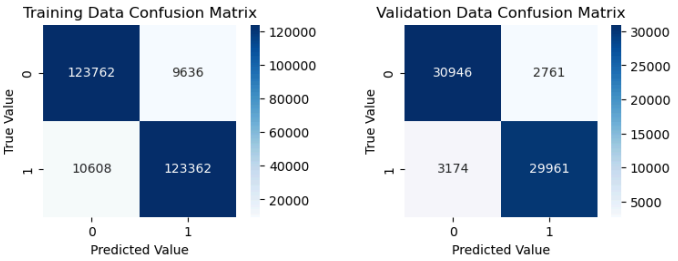


Figure 8. Confusion Matrix for Train and Validation with the Random Forest Model

VI. Conclusion (Model Evaluation)

	Id	Status
0	202944	0
1	202945	0
2	202946	0
3	202947	0
4	202948	0
...	...	...
50731	253675	1
50732	253676	0
50733	253677	1
50734	253678	0
50735	253679	0

We predicted the status of diabetes for each individual in the preprocessed test data sets with the 6 selected features and the Random Forest model.

We have successfully completed the assignment by processing 5 different models, such as Logistic Regression (Linear Features), Logistic Regression (Polynomial Features with degree 1), Logistic Regression (Polynomial Features with degree 2), Decision Tree, Random Forest, with our data sets, through the following steps: Exploratory Data Analysis (EDA), Data Preprocessing, Model Development, Model Evaluation.

Figure 9. Predicted Status Values

VII. References

[1] World Health Organization, “Diabetes,” *www.who.int*, Apr. 05, 2023. https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Diabetes%20is%20a%20chronic%20disease