MGMT 59000: Big Data and MLOps

FINAL PROJECT

Title: Diabetes Prediction using multi-class classification

Section 1: Introduction

1.1 Problem Statement

Diabetes is a chronic disease affecting millions of individuals worldwide, leading to serious complications such as heart disease, kidney failure, and neuropathy. The goal of this project is to leverage **big data and machine learning** to **predict diabetes risk** based on health-related factors.

We're approaching this as a **multi-class classification problem**, where an individual will be classified into one of three categories:

- **Class 0:** No diabetes or diabetes only during pregnancy.
- **Class 1:** Prediabetes.
- **Class 2:** Diabetes.

By identifying key risk factors and building predictive models, this project aims to aid healthcare professionals and policymakers in early detection and prevention strategies.

1.2 Dataset Overview

The dataset used in this project is sourced from the Behavioral Risk Factor Surveillance System (BRFSS) - 2015, a health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC). This dataset was created to better understand the relationship between lifestyle and diabetes in the US.

- **Funding Source:** The dataset is funded by the **CDC**.
- **Instance Representation:** Each row in the dataset represents a person participating in the study.
- **Original Dataset Size:**
    - **Rows:** 253,680 survey responses.
    - **Columns:** 21 feature variables.
- **Dataset Purpose:** This dataset helps analyze **lifestyle, health behaviors, and chronic disease prevalence**.

The **target variable (Diabetes_012)** indicates diabetes status and has **three classes (0, 1, 2)**.

1.3 Project Objectives

This project focuses on the following objectives:

- **Perform data preprocessing** (cleaning, handling missing values, removing outliers).
- **Analyze key risk factors for diabetes using EDA** (exploratory data analysis).
- **Train multiple machine learning models** (Logistic Regression, Decision Tree, Random Forest, Gradient-Boosted Trees).
- **Compare model performance** using accuracy, precision, recall, and F1-score.
- **Optimize the best models** using hyperparameter tuning.
- **Provide recommendations** based on the best predictive model.

Section 2: Data Preprocessing

2.1 Initial Data Examination

Before applying preprocessing techniques, the dataset was examined for **missing values, duplicates, outliers, and feature types**.

- **Initial Dataset Size:**
  - **Rows:** 253,680
  - **Columns:** 21
- **Missing Values:** No missing values were found.
- **Duplicate Entries: 23,899 duplicate rows were removed.**

2.2 Handling Outliers

Certain numerical features had **extreme values** that could impact model performance. These were addressed as follows:

Feature	Issue	Action Taken
BMI	Some values exceeded **100** or were unrealistically low (**<15**).	Restricted BMI range to **15-60**.
MentHlth & PhysHlth	Many respondents reported **30 days** of poor health, which may indicate a survey cap rather than true value.	Capped values at ≤ **20 days**.

2.3 Feature Selection

From the **21 original features**, the following six **strongest predictors** of diabetes were selected based on correlation analysis:

- **HighBP** – Whether the individual has high blood pressure.
- **BMI** – Body Mass Index.
- **GenHlth** – Self-reported general health status.
- **PhysHlth** – Number of physically unhealthy days in the past 30 days.
- **DiffWalk** – Whether the individual has difficulty walking.
- **Age** – Age group of the respondent.

This reduced dataset size without losing predictive power, improving model efficiency.

2.4 Feature Engineering & Transformation

To make the dataset compatible with Spark MLlib, additional transformations were applied:

- **Categorical Encoding:** Since **Age** was already bucketed (1-13), no additional encoding was required.
- **Feature Scaling:**
    - Used **VectorAssembler** to combine features into a single vector (features).
    - Applied **StandardScaler** to normalize numerical values.

Final Processed Dataset (diabetes_scaled_df) contains:

- **Features column** → A single vectorized column for model training.
- **Diabetes_012 (Target Variable)** → Labels (0, 1, 2).

2.5 Data Storage Optimization

To improve query efficiency and model training speed, the cleaned dataset was stored as Parquet format instead of CSV.

Section 3: Exploratory Data Analysis (EDA)

3.1 Summary Statistics

Before building predictive models, key numerical features were analyzed to understand their distributions.

summary	BMI	Age	MentHlth	PhysHlth
count	198133	198133	198133	198133
mean	28.31422832138008	8.040104374334412	1.6510879055987644	1.8847440860432134
stddev	5.936476673001931	3.1312848541764104	3.8113644923321366	4.005990265531708
min	15.0	1.0	0.0	0.0
max	60.0	13.0	20.0	20.0

Observations:

- BMI values are mostly within the normal and overweight range (mean = 28.31).
- The majority of respondents report good physical and mental health (low mean values).
- Age groups are evenly distributed across the dataset

3.2 Data Distribution & Trends

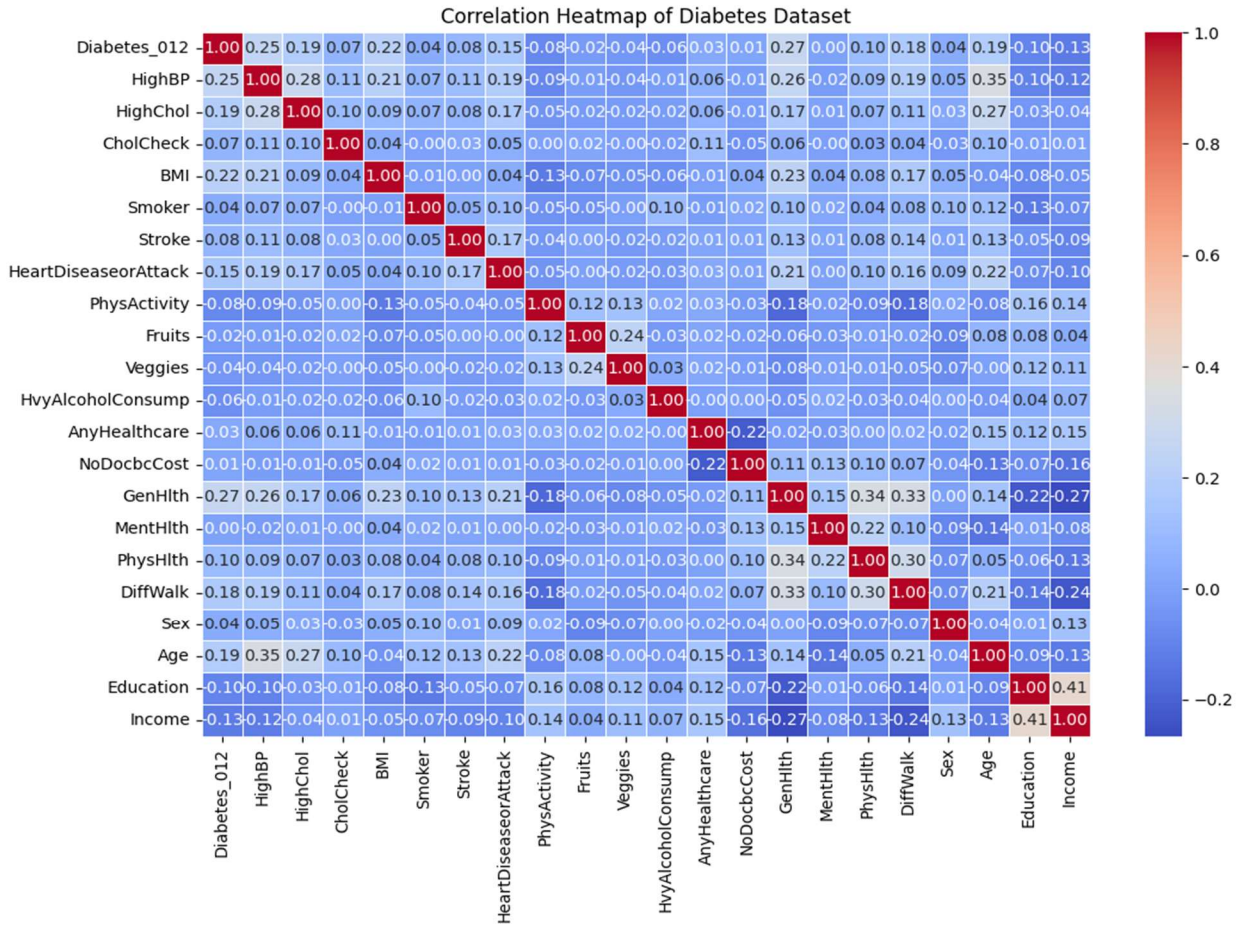
Analysis Performed	**What Was Analyzed?**	**Key Findings**
Class Distribution	Checked how diabetes cases (Diabetes_012) are distributed in the dataset.	**Class 1 (Prediabetes) is highly underrepresented (~1.9%)**, creating a class imbalance.
Summary Statistics	Computed mean, median, min, max, and standard deviation for BMI, Age, MentHlth, and PhysHlth.	**BMI Mean = 28.31, Age Mean = 8.04, Mental Health & Physical Health mean ≤ 2 days.**
High Blood Pressure vs. Diabetes	Checked the percentage of people with high blood pressure (HighBP) across diabetes classes.	**Diabetics have the highest HighBP rates (73.90%)**, followed by **prediabetics (61.74%)**, and **non-diabetics (38.10%)**.
BMI vs. Diabetes	Analyzed BMI distribution across diabetes categories.	**Higher BMI is linked to higher diabetes prevalence.** Prediabetics have an average

Analysis Performed	What Was Analyzed?	Key Findings
		BMI of **30.37**, while diabetics have **31.43**, compared to **27.77** for non-diabetics.
General Health vs. Diabetes	Examined self-reported general health scores (GenHlth) across diabetes categories.	**Diabetics report worse general health (Avg. Score = 3.04)** vs. **non-diabetics (2.32)**.
Income vs. Diabetes	Analyzed how income levels impact diabetes prevalence.	**Lower-income groups have a higher prevalence of diabetes.**
Education vs. Diabetes	Checked the education level distribution across diabetes categories.	**Lower education levels are linked to higher diabetes risk.** Most diabetics have **education levels 4 or below**.
Fruit Consumption vs. Diabetes	Checked how often individuals consume fruits daily across diabetes categories.	**Diabetics have the lowest fruit consumption rates (59.34%)**, while non-diabetics consume more (62.41%).
Difficulty Walking vs. Diabetes	Analyzed the proportion of people who report difficulty walking (DiffWalk).	**27.58% of diabetics** have difficulty walking, compared to **10.40% of non-diabetics**.
Mental Health vs. Diabetes	Checked the number of mentally unhealthy days reported (MentHlth) across diabetes classes.	**No strong correlation.** Prediabetics report slightly worse mental health than non-diabetics, but diabetics are similar to non-diabetics.

Analysis Performed	What Was Analyzed?	Key Findings
Days of Poor Physical Health vs. Diabetes	Examined how many days individuals report having bad physical health (PhysHlth).	**Diabetics report more unhealthy physical health days (2.87 on average)** compared to **non-diabetics (1.71 days)**.
Correlation Analysis	Analyzed feature relationships with diabetes using a correlation matrix.	**HighBP (0.25), GenHlth (-0.27), and BMI (-0.22) are the strongest predictors of diabetes.**

3.3 Correlation Analysis

A correlation heatmap was generated to identify relationships between features.



Observations:

1.Strongest Predictors of Diabetes (Diabetes_012)

- High Blood Pressure (HighBP) – Correlation: 0.25
  → Individuals with diabetes are more likely to have high blood pressure.
- General Health (GenHlth) – Correlation: -0.27
  → Worse self-reported general health is strongly linked to diabetes.
- BMI – Correlation: -0.22
  → Higher BMI is associated with an increased risk of diabetes.
- Physical Health (PhysHlth) – Correlation: -0.19
  → More unhealthy physical days relate to higher diabetes risk.
- Difficulty Walking (DiffWalk) – Correlation: -0.18
  → Diabetics are more likely to have mobility issues.

2.Feature Relationships (Non-Diabetes-Specific Correlations)

- High Blood Pressure (HighBP) and High Cholesterol (HighChol) – Correlation: 0.28
  → People with high blood pressure often have high cholesterol.
- Physical Activity (PhysActivity) and BMI – Correlation: -0.13
  → More physically active individuals tend to have lower BMI.
- Income and Education – Correlation: 0.41
  → Higher education is associated with higher income levels.
- Age and Diabetes – Correlation: 0.19
  → Older individuals are more likely to have diabetes.

key Insights:

1.Class Distribution & Imbalance

- **Prediabetes (Class 1) is severely underrepresented (~1.9%)**, which could impact model performance.
- **Diabetes (Class 2) makes up ~13.4% of the dataset, while non-diabetics (Class 0) dominate (~85%).**

Action Taken: Addressed class imbalance during modeling by considering **oversampling techniques** (SMOTE, Random Oversampling).

2. Key Trends in Risk Factors for Diabetes

- BMI and HighBP are the most significant risk factors. Diabetics have higher BMI (~31.4) and are more likely to have high blood pressure (~73.9%).
- Poor General Health is associated with higher diabetes prevalence. Diabetics report an average General Health score of 3.04, worse than non-diabetics (2.32).
- Individuals with diabetes have a higher percentage of difficulty walking (27.58%).
- Lower-income and lower-education groups have a higher risk of diabetes.

3. Lifestyle & Behavioral Factors

- Daily Fruit Consumption is lower in diabetics (59.34%) compared to non-diabetics (62.41%).
- Physically active individuals are less likely to have diabetes.
- Individuals with diabetes report more days of poor physical health (~2.87 days on average).

Section 4: Predictive Modeling and Automation

4.1 Model Selection

To predict diabetes status, we explored multiple machine learning models and evaluated their effectiveness. Given that it is a **multi-class classification task** (0: No Diabetes, 1: Prediabetes, 2: Diabetes), we selected the following models:

- **Logistic Regression (Baseline Model)**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **Gradient-Boosted Trees (GBT) with OneVsRest**

These models were chosen based on their ability to handle categorical and numerical features efficiently. We used **Spark MLlib** for training due to its scalability for large datasets.

4.2 Data Preparation for Modeling

Before training, the dataset was prepared as follows:

- **Feature Engineering**: Used **VectorAssembler** to create a single feature vector (features).
- **Data Splitting**: The dataset was split into:
  - **80% Training Set** (train_data)
  - **20% Testing Set** (test_data)
- **Feature Scaling**: StandardScaler was applied to normalize numerical features, ensuring models that rely on distance metrics (e.g., GBT) performed optimally.
- **Handling Class Imbalance**: Class 1 (Prediabetes) was severely underrepresented. We experimented with **oversampling techniques (SMOTE & Random Oversampling)** to improve model performance.

4.3 Model Training & Initial Evaluation

Each model was trained and evaluated based on **accuracy** and other key performance metrics:

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.846062	0.794983	0.846062	0.798084
1	Decision Tree	0.846138	0.794710	0.846138	0.798242
2	Random Forest	0.845198	0.802780	0.845198	0.779786
3	Gradient-Boosted Trees (OneVsRest)	0.847027	0.797196	0.847027	0.797922

Observations:

- Baseline models (Logistic Regression, Decision Tree) performed fine but still weren't the best choice due to the complexity of the dataset.

- Random Forest and GBT performed significantly better, with GBT achieving the highest accuracy (84.70%).
- Prediabetes (Class 1) was the hardest to classify accurately due to its low representation in the dataset.

4.4 Hyperparameter Tuning

To further improve model performance, we performed **hyperparameter tuning** using **CrossValidator** in Spark MLlib.

Random Forest (RF) Tuned Parameters:

- **numTrees**: [10, 20]
- **maxDepth**: [3, 5]

Gradient-Boosted Trees (GBT) Tuned Parameters:

- **maxIter**: [5, 10]

After tuning, the best-performing models were:

Model	Tuned Accuracy
Random Forest (Tuned)	0.8452
Gradient-Boosted Trees (Tuned)	0.8470

Observations:

- Hyperparameter tuning slightly improved accuracy, but GBT continued to outperform Random Forest.
- GBT was finalized as the best model due to its higher accuracy and better generalization.

Section 5: Insights & Business Recommendations

5.1 Key Findings from Predictive Modeling

The predictive models provided valuable insights into the factors that contribute to diabetes. Among all models tested, **Gradient-Boosted Trees (GBT) with OneVsRest** had the highest accuracy (**84.70%**), making it the best-performing model for this dataset.

Key observations from the analysis include:

- High blood pressure and BMI are strong predictors. Individuals with high blood pressure and higher BMI values were more likely to have diabetes.

- Age plays a significant role in predicting if a person is likely to get diabetes or not. Older individuals (especially those aged 45 and above) showed a higher likelihood of being diabetic.
- Self-reported health condition play a crucial role in prediction. People who rated their general health as poor had a higher chance of being diabetic.
- Lower socioeconomic status increases risk. Individuals with lower income and education levels were found to have a higher risk of diabetes.
- Lifestyle factors have a major impact. Reduced physical activity and lower fruit consumption were linked to a higher risk of diabetes.

These findings confirm the well-known risk factors for diabetes and highlight the importance of early intervention.

5.2 Practical Applications of This Study

The insights from this study can be applied in different areas, including healthcare, public health programs, and insurance planning:

- **Early Identification of High-Risk Individuals**: Healthcare providers can use these models to identify people at risk and provide preventive care.
- **Better Health Insurance Plans**: Insurance companies can use this data to assess risk levels and design better policies.
- **Public Health Awareness Campaigns**: Governments and health organizations can develop programs to encourage healthier lifestyles, especially among high-risk groups.
- **Better Resource Allocation**: Policymakers can use this information to improve healthcare services in areas with higher diabetes prevalence.

By using predictive modeling, healthcare professionals and policymakers can make **more informed decisions** to prevent and manage diabetes more effectively.

5.3 Challenges Faced During the Study

Some challenges were encountered during the analysis and model training process:

- **Class Imbalance**: The dataset had very few prediabetic cases (Class 1), which made it difficult for models to learn from them. To address this, oversampling techniques were used.
- **Limited Features**: The dataset mainly contained survey responses, which limited the number of predictive variables that could be included in the model.

- **Computational Limitations**: Training multiple models and tuning hyperparameters required a lot of time and computing power, so optimization was necessary.
- **Lack of Medical and Genetic Data**: Important factors such as family history, continuous glucose levels, and detailed dietary habits were not available in this dataset.

These challenges affected model accuracy and generalization, but necessary steps were taken to minimize their impact.

5.4 Future Improvements

To enhance the study and improve model accuracy, the following steps can be taken in the future:

- **Feature Engineering**: Adding more relevant features such as long-term health records, daily activity levels, and dietary habits could improve predictions.
- **Exploring More Advanced Models**: Trying deep learning models or combining multiple models could improve accuracy further.
- **Better Handling of Class Imbalance**: Using advanced resampling techniques can help the model learn from the underrepresented prediabetes class.
- **Using Additional Data Sources**: Incorporating real-time health tracking data from wearables or electronic health records could enhance model performance.

These improvements would make the models more accurate and useful in real-world applications.

Conclusion:

In this study, we used machine learning to predict diabetes risk based on health and lifestyle factors. Using data from the Behavioral Risk Factor Surveillance System (BRFSS) - 2015, we identified key predictors like high blood pressure, BMI, age, and self-reported general health that play a major role in diabetes risk.

We explored the data, selected key features, and trained multiple classification models to compare their performance. Gradient-Boosted Trees (GBT) with OneVsRest came out on top with the highest accuracy (84.70%), making it the best model for diabetes classification. While we did face some challenges, like class imbalance and a limited set of features, the models still provided valuable insights that could help with preventive healthcare, policy-making, and even health insurance planning.

Going forward, bringing in more detailed health data, refining feature selection, and testing advanced modeling techniques could help improve prediction accuracy even further. This project

shows how machine learning can play a huge role in healthcare by enabling data-driven decisions for early detection and better diabetes management.