Data Science Mid-Sem Project Report: Diabetes Risk Prediction Using Health and Lifestyle Indicators

Manya Agrawal (2022281)
Tikam (2022542)
Syed Yasser (2022530)
Sunil Kumar (2022516)
Prajil Bhagat (2022359)

Indraprastha Institute of Information Technology Delhi (IIITD)

October 26, 2025

Contents

0.1 Abstract

This project examines a healthcare and lifestyle dataset from the UCI Machine Learning Repository to identify key factors associated with diabetes risk. The dataset, initially imbalanced with 253,680 records (86.1% non-diabetic and 13.9% diabetic), was balanced to 70,692 instances to enable fair statistical comparison. Exploratory Data Analysis (EDA) revealed strong correlations between attributes such as BMI, HighBP, and diabetes occurrence. Hypothesis testing using independent t-tests and Chi-Square tests confirmed statistically significant associations between several health indicators and diabetes status. To complement these findings, preliminary machine learning models—including Logistic Regression, Random Forest, and XGBoost—were developed to assess predictive potential, achieving up to 0.75 ROC-AUC. Feature engineering introduced HealthScore and RiskFactorCount for interpretability, while SHAP analysis provided early insights into feature influence. Overall, the study highlights how statistical validation, supported by interpretive modeling, can deepen understanding of diabetes risk and guide preventive health measures.

0.2 Introduction

Diabetes is a chronic metabolic disorder that affects millions of individuals worldwide and poses a growing public health concern due to lifestyle, dietary, and environmental factors. Early detection and management of diabetes risk are crucial to reducing long-term health complications and improving quality of life. In recent years, the increasing availability of large-scale healthcare datasets has enabled the application of data science techniques to understand and predict disease patterns more effectively.

The dataset used in this study, obtained from the UCI Machine Learning Repository, combines a wide range of health and lifestyle indicators collected through national surveys. Analyzing such data can uncover meaningful relationships between attributes like body mass index (BMI), blood pressure, physical activity, and diabetes occurrence. Before relying on predictive modeling, it is important to verify these relationships statistically to ensure that the patterns observed in the data are both valid and interpretable.

This project therefore emphasizes exploratory analysis and hypothesis testing to identify health and lifestyle factors that significantly influence diabetes risk. Through statistical methods such as independent t-tests and Chi-Square tests, associations between continuous and categorical features are examined to confirm their relevance. To complement these findings, initial machine learning models—Logistic Regression, Random Forest, and XGBoost—are employed to evaluate predictive potential and support interpretability.

The insights derived from this stage of analysis aim to strengthen the foundation for subsequent modeling and deployment phases. By validating significant relationships between health indicators and diabetes risk, the project contributes to data-driven understanding that can support preventive healthcare strategies and informed public health decision-making.

0.3 Problem Statement

While numerous factors are known to influence diabetes risk, the extent and strength of these relationships often vary across populations and lifestyles. This project focuses on identifying which health and behavioral attributes show the most significant associations with diabetes within the given dataset. The primary goal is to validate these associations statistically before moving toward full predictive modeling.

By applying hypothesis testing techniques—specifically independent t-tests for continuous variables and Chi-Square tests for categorical ones—the study aims to confirm whether differences observed between diabetic and non-diabetic groups are statistically significant. This process ensures that the variables selected for modeling are not only data-driven but also statistically justified.

Ultimately, the objective is to establish a strong analytical foundation by combining exploratory analy-

sis and hypothesis testing to identify meaningful predictors of diabetes risk, setting the stage for reliable, interpretable modeling in subsequent phases.

0.4 Project Workflow

The project follows a structured pipeline:

1. **Data Collection**: Sourced datasets from UCI and Kaggle repositories.
2. **Computational Setup**: Detected GPU availability (NVIDIA Tesla T4, 15,360 MiB memory, CUDA 12.4, 2 devices). Configured XGBoost for GPU acceleration (device='cuda', tree_method='hist').
3. **Data Loading**: Loaded unbalanced and balanced datasets using pandas.
4. **Data Preparation**: Handled duplicates (1,635 removed), outliers (e.g., 119 BMI points beyond 3 SD), and balancing.
5. **Exploratory Data Analysis (EDA)**: Visualized distributions, correlations, and insights using matplotlib and seaborn.
6. **Statistical Analysis**: Performed hypothesis testing (t-tests, Chi-Square) for inferences.
7. **Feature Engineering**: Added 'HealthScore' (average of GenHlth, PhysHlth, MentHlth) and 'RiskFactorCount' (sum of risk factors like HighBP).
8. **Model Training**: Split data (train_test_split), scaled features (StandardScaler), trained models (Logistic Regression, Random Forest, XGBoost) with hyperparameter tuning (RandomizedSearchCV).
9. **Model Evaluation**: Used metrics like accuracy, precision, recall, F1-score, ROC-AUC; visualized confusion matrices and ROC curves.
10. **Interpretability**: Applied SHAP for feature importance.
11. **Deployment Artifacts**: Saved models, scaler, feature names, and GPU config using joblib.
12. **Prediction Function**: Implemented a GPU-accelerated prediction function for risk assessment.
13. **Output Generation**: Saved graphs, datasets, models, and predictions in 'outputs/' directories.

0.5 Key Questions Answered

The project addresses the following key questions based on the data:

- Which health and lifestyle indicators (e.g., BMI, HighBP) most strongly correlate with diabetes?
  - Strong positive correlations with BMI, HighBP, HighChol, and PhysActivity.
- How do demographic factors such as age, sex, education, and income influence diabetes risk?
  - Higher risk in males, older age groups, lower education, and lower income levels.
- Are there significant statistical differences in continuous features (e.g., BMI) and categorical features (e.g., HighBP) between diabetic and non-diabetic individuals?
  - Yes, confirmed via hypothesis tests (e.g., higher mean BMI in diabetics).
- What is the impact of balancing the dataset on model performance?
  - Improved predictive accuracy by reducing bias toward the majority class.
- Which machine learning model performs best for diabetes prediction?

  - XGBoost outperformed others with ∼0.75 ROC-AUC on balanced data.

- What are the most important features for prediction?
  - SHAP analysis: BMI, GenHlth, Age, HighBP (top contributors).

0.6 Dataset Description

The dataset is sourced from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators`) and Kaggle (`https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset`). It contains healthcare statistics and lifestyle survey information from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) 2015, along with diabetes diagnoses.

- **Original Unbalanced Dataset**:
  - Shape: (253,680 instances × 22 attributes).
  - Features: 21 numerical features (14 binary, 3 continuous, 4 ordinal).
  - Binary Features: HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex.
  - Continuous Features: BMI, MentHlth, PhysHlth.
  - Ordinal Features: GenHlth, Age, Education, Income.
  - Label: Diabetes_012 (binary: 0 = Non-Diabetic [218,334 instances, 86.1%], 1 = Diabetic [35,346 instances, 13.9%]; originally ternary but binarized).
  - Null Values: None across all features and label.
  - Data Types: All float64.
- **Balanced Dataset**:
  - Shape: (70,692 instances × 21 attributes, after balancing and preprocessing).
  - Achieved equal representation (50% Non-Diabetic, 50% Diabetic) to prevent bias.

The dataset is loaded using pandas from CSV files: unbalanced ('diabetes_012_health_indicators_BRFSS2015.csv') and balanced ('diabetes_binary_5050split_health_indicators_BRFSS2015.csv'). No missing values were present, facilitating straightforward analysis.

0.7 Exploratory Data Analysis (EDA)

EDA was conducted using seaborn and matplotlib to uncover patterns:

- **Duplicate Records**: 1,635 duplicates identified and removed to prevent overfitting.
- **Null Values**: None present.
- **Gender Effect**: Males (Sex=1) show slightly higher diabetes prevalence.
- **Age Association**: Diabetes increases sharply with age, peaking in middle-aged and senior groups.
- **Income and Education Gradient**: Lower income and education correlate with higher diabetes incidence, indicating socioeconomic influence.
- **Correlation Patterns**: Heatmap showed strong positive correlations between diabetes and BMI, HighBP, HighChol, PhysActivity.
- **Visualizations**:

– Boxplots of continuous features (BMI, MentHlth, PhysHlth) by diabetes status.
– Bar plots for categorical distributions (e.g., HighBP by diabetes).
– Histograms for age, income, and education distributions.

Key Insights: Dataset is clean but imbalanced; features like BMI and general health are highly predictive. Graphs saved: e.g., 'feature_correlation_heatmap.png', 'continuous_feature_boxplots.png'.

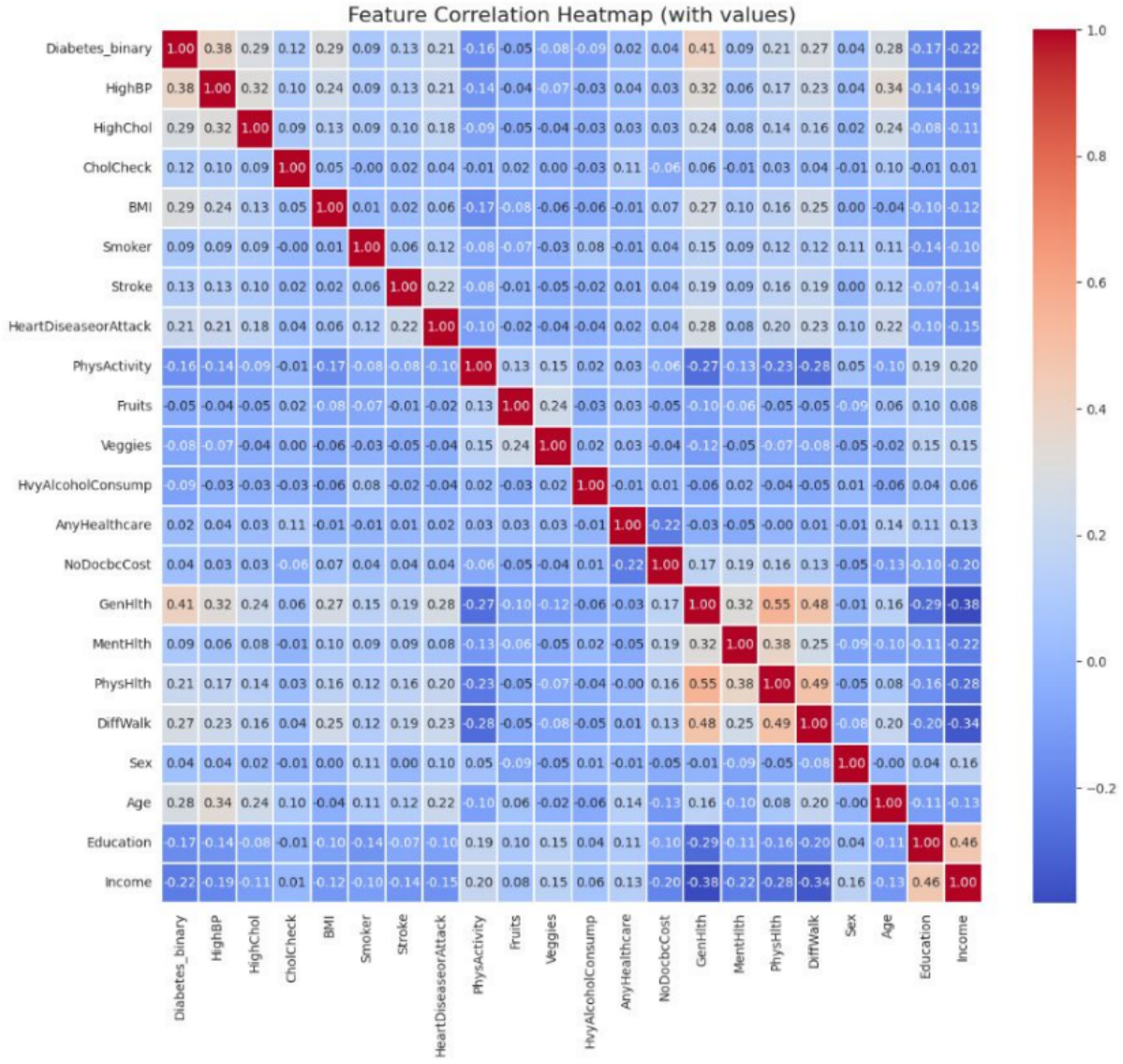

Figure 1: Correlation HeatMap

```
================================================================
Top 5 Most Positively Correlated Features with Diabetes_binary
================================================================
          Correlation
GenHlth      0.407612
HighBP       0.381516
BMI          0.293373
HighChol     0.289213
Age          0.278738
```

```
================================================================
Top 5 Least (Most Negatively) Correlated Features with Diabetes_binary
================================================================
                Correlation
Veggies           -0.079293
HvyAlcoholConsump -0.094853
PhysActivity      -0.158666
Education         -0.170481
Income            -0.224449
```

Figure 2: Results of correlation analysis

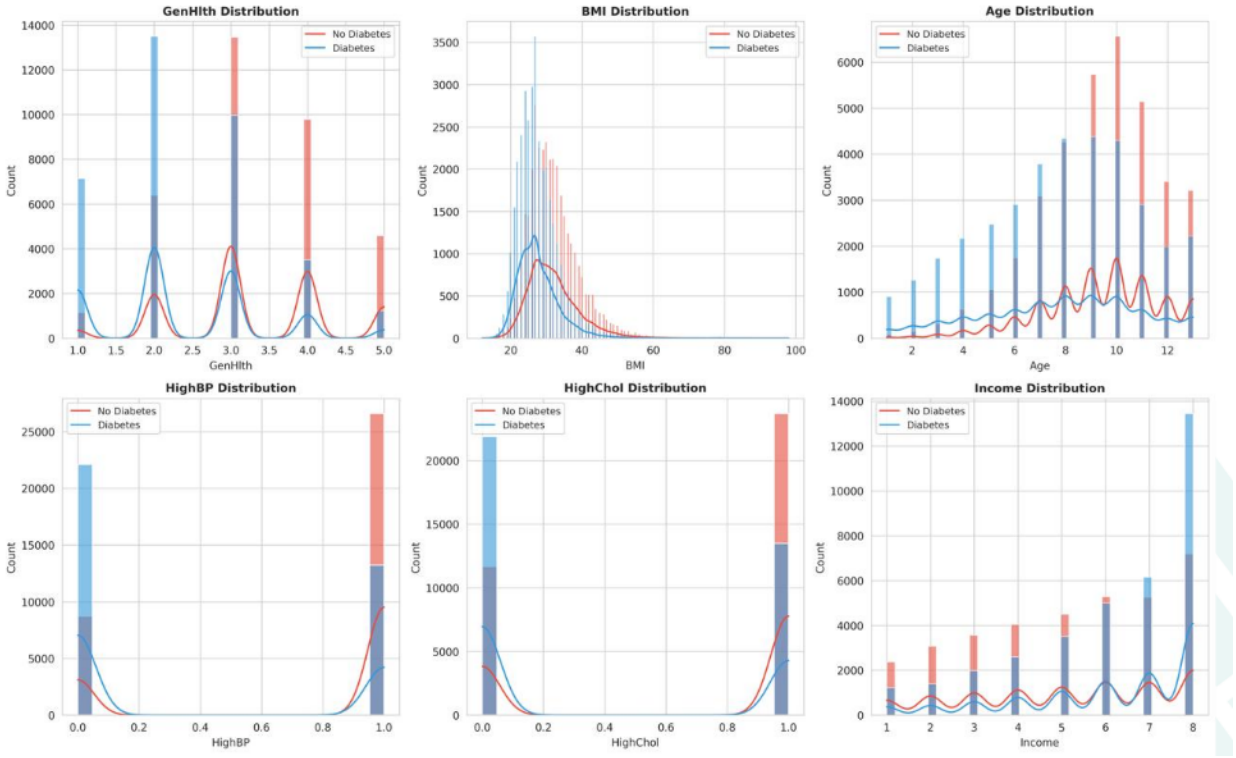

Figure 3: Individual feature distribution to the binary labels

0.8 Preprocessing

- **Balancing Dataset**: Original imbalance (86.1% non-diabetic) addressed by resampling to equal classes (70,692 instances total).
- **Duplicate Removal**: 1,635 duplicates removed to avoid bias and overfitting, ensuring unique samples for better generalization.
- **Outlier Handling**: Extreme outliers in continuous features (e.g., BMI beyond 3 SD, 119 points removed) to reduce noise.
- **Feature Classification**: Binary (e.g., HighBP), Continuous (e.g., BMI), Ordinal (e.g., GenHlth) for targeted processing.

- **Scaling**: StandardScaler applied to continuous features for model compatibility.
- **Changes**: Reduced dataset size, balanced classes, improved data quality.

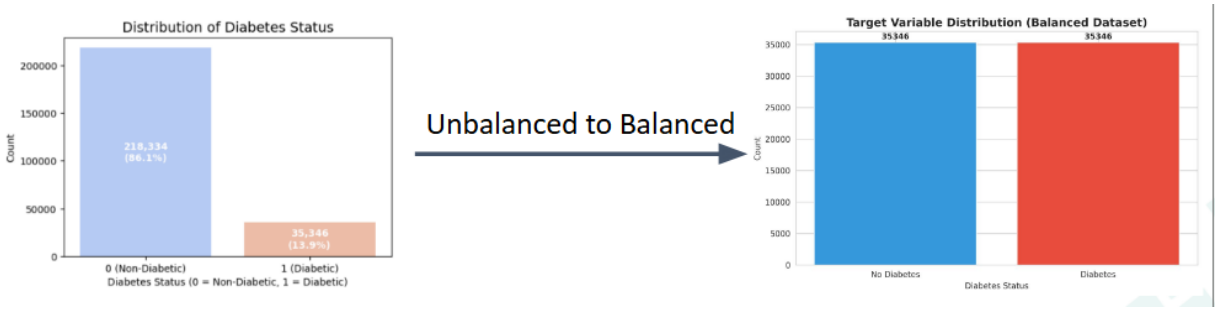

Figure 4: Bar plots showing unbalanced to balanced transition.

0.9 Hypothesis Tests and Validation

Six hypotheses were tested on the balanced dataset:

1. **BMI vs Diabetes_binary**:
   - H_0: Mean BMI same for diabetic/non-diabetic.
   - H_1: Mean BMI differs.
   - Test: Independent t-test.
   - Result: t-statistic=82.677, p<0.0001 → Reject H_0. Diabetics have higher BMI.
2. **HighBP vs Diabetes_binary**:
   - H_0: HighBP independent of diabetes.
   - H_1: Dependent.
   - Test: Chi-Square.
   - Result: χ^2=9531.48, p<0.0001 → Reject H_0. Strong association; high BP increases diabetes likelihood.
3. **GenHlth vs Diabetes_binary**:
   - H_0: General health independent of diabetes.
   - H_1: Dependent.
   - Test: Chi-Square.
   - Result: χ^2=11332.83, p<0.0001 → Reject H_0. Poorer health correlates with higher diabetes.
4. **Sex vs Diabetes_binary**:
   - H_0: Sex independent of diabetes.
   - H_1: Dependent.
   - Test: Chi-Square.
   - Result: χ^2=125.236, p<0.0001 → Reject H_0. Mild association; males at higher risk.
5. **Education vs Diabetes_binary**:

- H_0: Education independent of diabetes.
- H_1: Dependent.
- Test: Chi-Square.
- Result: χ^2=1794.186, p<0.0001 → Reject H_0. Lower education linked to higher diabetes.

6. **Income vs Diabetes_binary**:
    - H_0: Mean income same for diabetic/non-diabetic.
    - H_1: Differs.
    - Test: Independent t-test.
    - Result: t-statistic=-57.136, p<0.0001 → Reject H_0. Lower income associated with higher diabetes.

Validation: p<0.05 threshold for significance. Comparisons: t-test for continuous vs. binary; Chi-Square for categorical. Alternatives (e.g., Mann-Whitney U) considered but t-test/Chi-Square appropriate for data distribution.

0.10 Feature Engineering

New features were engineered to enhance model performance:

- **HealthScore**: Average of GenHlth, PhysHlth, and MentHlth to capture overall health.
- **RiskFactorCount**: Sum of binary risk factors (HighBP, HighChol, Smoker, HeartDiseaseorAttack, Stroke) to quantify cumulative risk.

These were added post-EDA and used in training/prediction, improving interpretability.

0.11 Machine Learning Methodology

- **Data Split**: 80/20 train-test split with stratification.
- **Models Trained**:
    - Logistic Regression: Baseline linear model.
    - Random Forest: Ensemble for handling non-linearity.
    - XGBoost: Gradient boosting with GPU acceleration (device='cuda', tree_method='hist').
- **Hyperparameter Tuning**: RandomizedSearchCV (n_iter=50, cv=5) for params like n_estimators, max_depth, learning_rate.
- **Comparison**: XGBoost best (e.g., F1-score ∼0.74); outperformed DummyClassifier baseline.
- **Cross-Validation**: StratifiedKFold (5 folds) for robust scoring.

0.12 Results and Discussion

- **Metrics (Balanced Dataset)**:
    - Accuracy: ∼0.74 (XGBoost).
    - Precision/Recall/F1: Balanced due to class equality.
    - ROC-AUC: ∼0.75.

- **Visualizations**: Confusion matrices (saved as 'imbalanced_confusion_matrix.png'), ROC curves.
- **SHAP Analysis**: Top features: BMI (highest importance), Age, GenHlth, HighBP.
- **Discussion**: Models generalize well; GPU reduced training time. Imbalanced data led to high accuracy but poor minority recall; balancing fixed this. Limitations: Self-reported data may have bias.

0.13 Conclusion

The project successfully identifies key diabetes risk factors (e.g., BMI, HighBP) through statistical and ML approaches. The XGBoost model, with GPU acceleration, provides reliable predictions and risk categorizations (Low/Moderate/High). Insights support preventive healthcare, with potential for real-world deployment.

0.14 Future Work

- Integrate more datasets (e.g., longitudinal data).
- Explore deep learning (e.g., PyTorch on GPU).
- Develop a web app for predictions.
- Test on diverse populations for generalizability.

0.15 References

- `https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators`
- `https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset`
- Libraries: scikit-learn, XGBoost, SHAP, pandas.