Predicting Diabetes Risk

A Machine Learning Approach Using Health Indicators

By Steven Joel Ochanda

Project Overview: Predicting Diabetes Risk

- This presentation outlines a machine learning project aimed at developing a predictive model for diabetes risk.
- My goal is to leverage readily available health indicator data to identify individuals who may be at a higher risk of diabetes.
- This project demonstrates how data science and machine learning can be a powerful tool in public health initiatives and preventative medicine.

Business Understanding: Why Early Diabetes Prediction?

- Diabetes is a chronic disease affecting millions globally, leading to serious health issues like heart disease, vision loss, and kidney disease.
- Early identification of diabetes is crucial for preventative care, enabling lifestyle changes, and potentially reducing severe health complications and healthcare costs associated with the disease.
- My primary audience for this project is **business stakeholders** in the healthcare sector, public health organizations, or insurance companies.

Business Understanding: Why Early Diabetes Prediction?

A reliable prediction model can help us:

1. **Proactively Identify At-Risk Individuals:** Flag individuals for early screening or targeted interventions, potentially before severe symptoms appear.
2. **Optimize Resource Allocation:** Direct healthcare resources more efficiently to those who need them most.
3. **Inform Public Health Campaigns:** Provide data-driven insights to develop more effective prevention programs.

Machine learning is ideal for this task because:

1. It can analyze large datasets with many variables to find complex, non-obvious patterns in health indicators that may be indicative of diabetes risk.
2. It allows us to build a predictive tool that can adapt and improve with more data.

Data Understanding

The CDC Diabetes Health Indicators Dataset

- I chose this dataset because it's a comprehensive, publicly available resource from the Centers for Disease Control and Prevention (CDC).
- It contains **253,680 records** of healthcare statistics and lifestyle survey information from adults across the U.S..
- It includes 21 features (or variables) related to general health, lifestyle, and existing conditions, which are highly relevant to diabetes risk.
- The target variable, Diabetes_binary, indicates whether an individual has diabetes (1) or not (0)."

Initial Data Snapshot

```
--- First 5 rows of the dataset ---
   Diabetes_binary  HighBP  HighChol  CholCheck   BMI  Smoker  Stroke  \
0              0.0     1.0       1.0        1.0  40.0     1.0     0.0
1              0.0     0.0       0.0        0.0  25.0     1.0     0.0
2              0.0     1.0       1.0        1.0  28.0     0.0     0.0
3              0.0     1.0       0.0        1.0  27.0     0.0     0.0
4              0.0     1.0       1.0        1.0  24.0     0.0     0.0

   HeartDiseaseorAttack  PhysActivity  Fruits  ...  AnyHealthcare  \
0                   0.0           0.0     0.0  ...            1.0
1                   0.0           1.0     0.0  ...            0.0
2                   0.0           0.0     1.0  ...            1.0
3                   0.0           1.0     1.0  ...            1.0
4                   0.0           1.0     1.0  ...            1.0

   NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk  Sex   Age  Education  \
0          0.0      5.0      18.0      15.0       1.0  0.0   9.0        4.0
1          1.0      3.0       0.0       0.0       0.0  0.0   7.0        6.0
2          1.0      5.0      30.0      30.0       1.0  0.0   9.0        4.0
3          0.0      2.0       0.0       0.0       0.0  0.0  11.0        3.0
4          0.0      2.0       3.0       0.0       0.0  0.0  11.0        5.0

   Income
0     3.0
1     1.0
2     8.0
3     6.0
4     4.0

[5 rows x 22 columns]
```

```
--- Dataset Info (data types, non-null counts) ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
None
```

A Critical Challenge: Class Imbalance

A significant challenge we immediately identified was the severe imbalance in our target variable:

- Approximately **84% of individuals do NOT have diabetes** (Class 0).
- Approximately **15% of individuals DO have diabetes** (Class 1).

If a model simply predicts 'No Diabetes' for everyone, it will achieve 84% accuracy. However, it would completely miss identifying actual diabetes cases, which is unacceptable for our goal of early risk identification.

```
--- Target Variable Distribution ('Diabetes_binary') ---
0.0    218334
1.0     35346
Name: Diabetes_binary, dtype: int64
0.0    0.860667
1.0    0.139333
Name: Diabetes_binary, dtype: float64
```

Data Preparation & Preprocessing

Step 1: Handling Duplicate Records

- Upon initial inspection, i identified **24,206 duplicate rows** within the dataset. I removed them, leaving me with **229,474 unique records** for my analysis.

Step 2: Separating Features and Target Variable

- My independent variables (features, denoted as 'X') – which are the 21 health and lifestyle indicators – and my dependent variable (target, denoted as 'y') – Diabetes_binary."

Data Preparation & Preprocessing

Step 3: Splitting Data into Training and Testing Sets

I divided my dataset into two parts:

1. **Training Set (80%):** Used to teach the model patterns in the data."
2. **Testing Set (20%):** Used to evaluate how well our trained model performs on data it has never seen before, simulating real-world application.

Preventing Data Leakage: I used a technique called stratified sampling (stratify=y) during the split. This ensures that the proportion of 'Diabetes' (Class 1) and 'No Diabetes' (Class 0) cases is maintained equally in both the training and testing sets.

Data Split Summary

```
--- Data Split Summary ---
X_train shape: (183579, 21)
X_test shape: (45895, 21)
y_train shape: (183579,)
y_test shape: (45895,)

--- Class distribution in y_train ---
0.0    0.847052
1.0    0.152948
Name: Diabetes_binary, dtype: float64

--- Class distribution in y_test ---
0.0    0.847064
1.0    0.152936
Name: Diabetes_binary, dtype: float64
```

Data Preparation & Preprocessing

Step 4: Feature Scaling

I applied StandardScaler to transform our features. This scales the data so that it has a mean of 0 and a standard deviation of 1, effectively standardizing the range of all features.

Preventing Data Leakage :

- Fitting the StandardScaler ONLY on the training data (X_train). This means the scaler learns the mean and standard deviation exclusively from the training set.
- Transforming BOTH the training data (X_train) and the testing data (X_test) using the scaler fitted on the training data.

Addressing Class Imbalance: Ensuring Fair Prediction

As highlighted earlier, our dataset is severely imbalanced, with roughly 84% 'No Diabetes' cases and only 15% 'Diabetes' cases.

- Strategy 1: Model-Level Adjustment (class_weight='balanced’)
  This parameter automatically adjusts the weights of the classes during training.
- Strategy 2: Data Resampling with SMOTE (Synthetic Minority Over-sampling Technique)
  SMOTE generates new, synthetic data points that are similar to existing minority class samples but are not exact copies.
  By creating a more balanced training environment, SMOTE enables my models to learn robust patterns for both classes, significantly improving their ability to detect diabetes cases.

Model Selection & Iterative Evaluation (Results)

Key Performance Metrics Explained

- **Accuracy:** The overall proportion of correct predictions.
- **Confusion Matrix:** A table that shows where our model made correct and incorrect predictions.
  - **True Negative (TN):** Correctly predicted No Diabetes.
  - **False Positive (FP):** Incorrectly predicted Diabetes (false alarm).
  - **False Negative (FN):** Incorrectly predicted No Diabetes (missed diagnosis).
  - **True Positive (TP):** Correctly predicted Diabetes.
- **Precision (for Diabetes):** Out of all cases predicted as Diabetes, how many were actually Diabetes. High precision reduces false alarms.
- **Recall (for Diabetes - *Critical for our goal*):** Out of all actual Diabetes cases, how many did our model correctly identify. High recall minimizes missed diagnoses.
- **F1-Score (for Diabetes):** A balance between Precision and Recall.
- **AUC (Area Under the ROC Curve):** Measures the model's ability to distinguish between classes. A higher AUC (closer to 1) indicates better discrimination.

Model 1: Logistic Regression

- **Results on Test Set: Overall Accuracy:** 71.42%
- **Confusion Matrix:**
  - True Negatives (No Diabetes identified correctly): 27,445
  - False Positives (No Diabetes incorrectly identified as Diabetes): 11,431
  - False Negatives (Actual Diabetes missed): 1,686
  - True Positives (Actual Diabetes identified correctly): 5,333
- **Key Metrics for Diabetes (Class 1):**
  - Precision: 0.32
  - **Recall: 0.76**
  - F1-Score: 0.45
- **AUC Score:** 0.81

Model 2: Logistic Regression with SMOTE

- **Results on Test Set: Overall Accuracy:** 71.96%
- **Confusion Matrix:**
  - True Negatives: 27,520
  - False Positives: 11,356
  - False Negatives: 1,707
  - True Positives: 5,312
- **Key Metrics for Diabetes (Class 1):**
  - Precision: 0.32
  - **Recall: 0.71**
  - F1-Score: 0.49
- **AUC Score:** 0.81

Model 3: Decision Tree Classifier with SMOTE

- **Results on Test Set: Overall Accuracy:** 78.87%
- **Confusion Matrix:**
  - True Negatives: 31,771
  - False Positives: 7,105
  - False Negatives: 3,052
  - True Positives: 3,967
- **Key Metrics for Diabetes (Class 1):**
  - Precision: 0.36
  - **Recall: 0.57**
  - F1-Score: 0.44
- **AUC Score:** 0.79

Model 4: Random Forest Classifier

- **Results (without SMOTE): Overall Accuracy:** 98.88%
- **Confusion Matrix:**
  - True Negatives: 37,669
  - False Positives: 1,207
  - False Negatives: 5,977
  - True Positives: 1,042
- **Key Metrics for Diabetes (Class 1):**
  - Precision: 0.46
  - **Recall: 0.15**
  - F1-Score: 0.26
- **AUC Score:** 0.78

Model 5: Random Forest Classifier with SMOTE

- **Results on Test Set: Overall Accuracy:** 83.50%
- **Confusion Matrix:**
  - True Negatives: 36,398
  - False Positives: 2,478
  - False Negatives: 5,094
  - True Positives: 1,925
- **Key Metrics for Diabetes (Class 1):**
  - Precision: 0.44
  - **Recall: 0.27**
  - F1-Score: 0.34
- **AUC Score:** 0.78

Best Model Analysis, Limitations, & Recommendations

The **Decision Tree Classifier trained with SMOTE-resampled data** stands out as our best performing model for this project, particularly when considering the crucial goal of identifying actual diabetes cases.

While its overall accuracy of 78.87% is slightly lower than the Random Forest's (without SMOTE), its performance on the minority 'Diabetes' class is significantly better and more balanced.

Limitations of my Model and Approach

False Negatives Exist: Despite my efforts, the model still generates **3,052 False Negatives**.

False Positives Impact: The **7,105 False Positives** could lead to unnecessary follow-up tests, anxiety, and healthcare costs for individuals who are not diabetic.

Stakeholders need to consider the cost-benefit of these 'false alarms' versus the benefit of identifying more true cases."

Recommendations

For Stakeholders:

- This model should be deployed as an **initial screening tool**. It can identify individuals who exhibit a higher risk profile for diabetes based on their health indicators. These flagged individuals can then be prioritized for further, more definitive medical testing and consultation, optimizing healthcare resource allocation.