Image Source: Canva

Diabetes Prediction Project

12.12.2025

Sheikh Abu Nasher
Suffolk University
ISOM 835 Predictive Analytics and Machine Learning
Professor Hasan Arslan

SUFFOLK UNIVERSITY
SAWYER BUSINESS SCHOOL

Overview

This project analyzes a large medical dataset to develop a predictive model that predicts the likelihood of an individual having diabetes.

Using a predictive analytics workflow, the project examines factors such as age, BMI, blood glucose level, and Hemoglobin A1 that contribute to diabetes risk. The project also highlights the importance of model interpretability, ethical considerations, and responsible machine learning use in health-related analytics.

Outline

1. Executive Summary
2. Introduction and Business Context
3. Exploratory Data Analysis
4. Methodology
5. Results & Model Comparison
6. Business Insights
7. Ethics and Responsible AI
8. Conclusion and Future Work
9. References and AI Acknowledgment

Executive Summary

Diabetes is a growing and important public health concern not only in the United States of America but also worldwide. In healthcare analytics, early detection of diabetes is an effective way to reduce long-term health complications and healthcare costs, especially with Americans already struggling with their daily costs and high health insurance premiums.

Many Americans and individuals globally who are at-risk of diabetes go undiagnosed until mid or severe symptoms appear. The purpose of this project is to develop a predictive model that can help identify individuals who may be at higher risk of diabetes based on commonly gathered medical data.

The dataset used for this project was a diabetes prediction dataset containing over 100,000 patient records obtained from Kaggle. Each record included features such as age, gender, BMI, smoking history, blood glucose level, HbA1c level, hypertension, and heart disease status. The target variable indicated whether a patient had been diagnosed with diabetes. The dataset was highly imbalanced, with only about 8% of patients having diabetes, which required special handling with ***SMOTE*** during the modeling phase.

The project consisted of six phases, which included exploratory initial checks, data analysis, data cleaning and preprocessing, class imbalance handling, model development, and evaluation. Key preprocessing steps included imputing missing values, encoding categorical features, scaling numerical variables, and stratifying the training and testing splits.

During modeling, three different models were evaluated:

- Logistic Regression *(baseline model)*
- Decision Tree
- Random Forest

(each combined with SMOTE to balance the minority class)

The Decision Tree Model was chosen to be the final model due to balanced scores on accuracy, precision, and recall compared to Logistic Regression and Random Forest. A feature analysis on the final model revealed that blood glucose and HbA1c levels were the strongest

predictors of diabetes, which is already known in medical research. Additionally, patient BMI and age contributed on a small scale.

The final model of this project would be a huge supportive tool for healthcare professionals to identify patients who are most likely to have diabetes in the near future and need more support and additional medical tests. However, this model should not be a standalone diabetes diagnostic tool replacing medical professionals and standard testing. It should be a supportive tool while making sure patient privacy and ethical guidelines are followed.

Introduction and Business Context

Diabetes is one of the most serious health problems the United States of America has ever faced, with more than 38 million adults having diabetes, and 98 million having prediabetes, according to the ***Preventing Chronic Disease*** publication of the ***Centers for Disease Control***. A diabetes prediction model can offer healthcare providers an opportunity to identify at-risk patients and help them live healthier lifestyles, prevent long-term health concerns, and reduce healthcare costs.

The main purpose of this project is to build a machine-learning model that can predict whether a patient is likely to have diabetes based on commonly gathered medical data. By using the model results, healthcare providers can facilitate follow-up testing, improve early detection, and allocate medical resources more efficiently. The model built in this project is not intended to replace professional medical diagnosis backed heavily by science. Instead, the predictive model should serve as a supportive tool that highlights individuals who may benefit from additional screening.

This problem matters because diabetes is often silent in its early stages. Undiagnosed diabetes can lead to severe complications such as kidney failure, blindness, cardiovascular disease, etc. Many of these outcomes are preventable with early identification and lifestyle intervention.

To guide this project, several research questions were established:

1. Which clinical and demographic features are most strongly associated with diabetes in this population?
2. Can a predictive model accurately classify patients as diabetic or non-diabetic despite class imbalance?
3. Which machine-learning approach provides the best balance of accuracy, recall, and interpretability?
4. How can the insights gained from the model support healthcare decision-making and early intervention strategies?

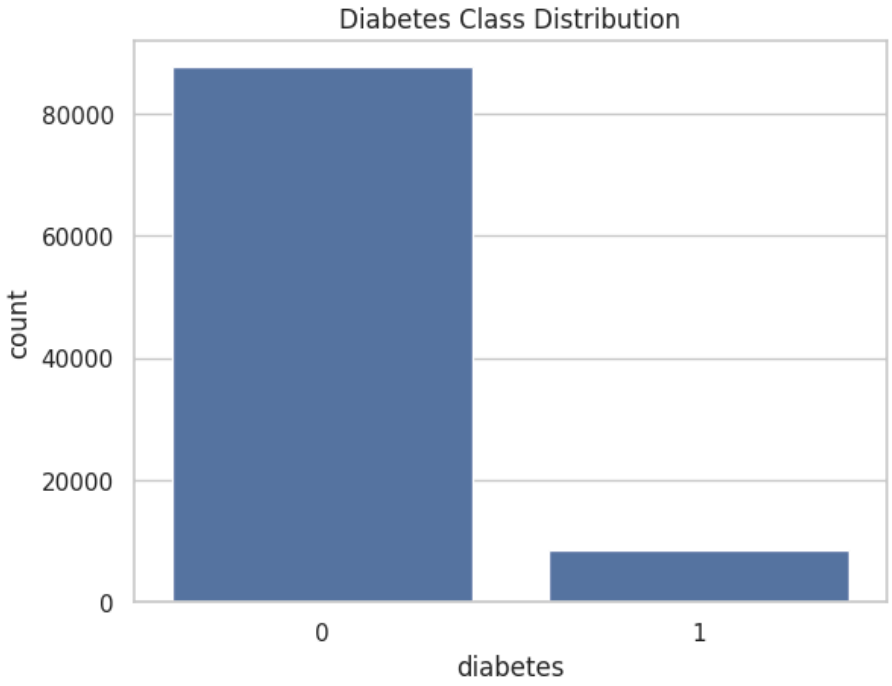
The dataset used for this project from Kaggle contains around 100,000 patient records, including variables such as gender, age, BMI, hypertension status, heart disease, smoking history, HbA1c level, and blood glucose level. The target variable, diabetes, indicates whether the patient has been diagnosed with diabetes. The dataset was selected because it is large, diverse, and reflects real-world clinical information.

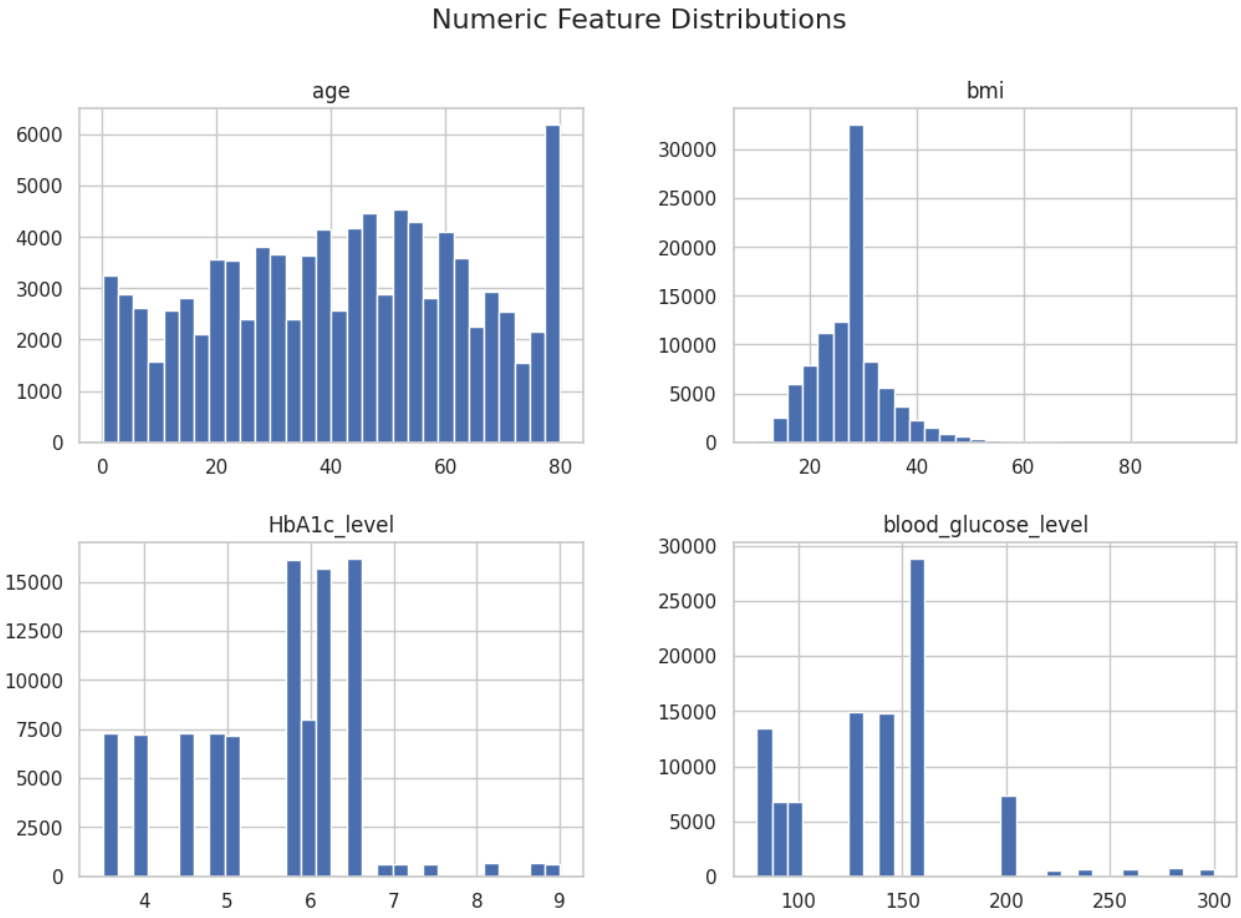
Note: The First Draft of the project has a smaller, identical dataset from Kaggle that has fewer than 1,000 rows, which did not produce standard model results and mid accuracy. Therefore, a larger Diabetes dataset was chosen for better accuracy.

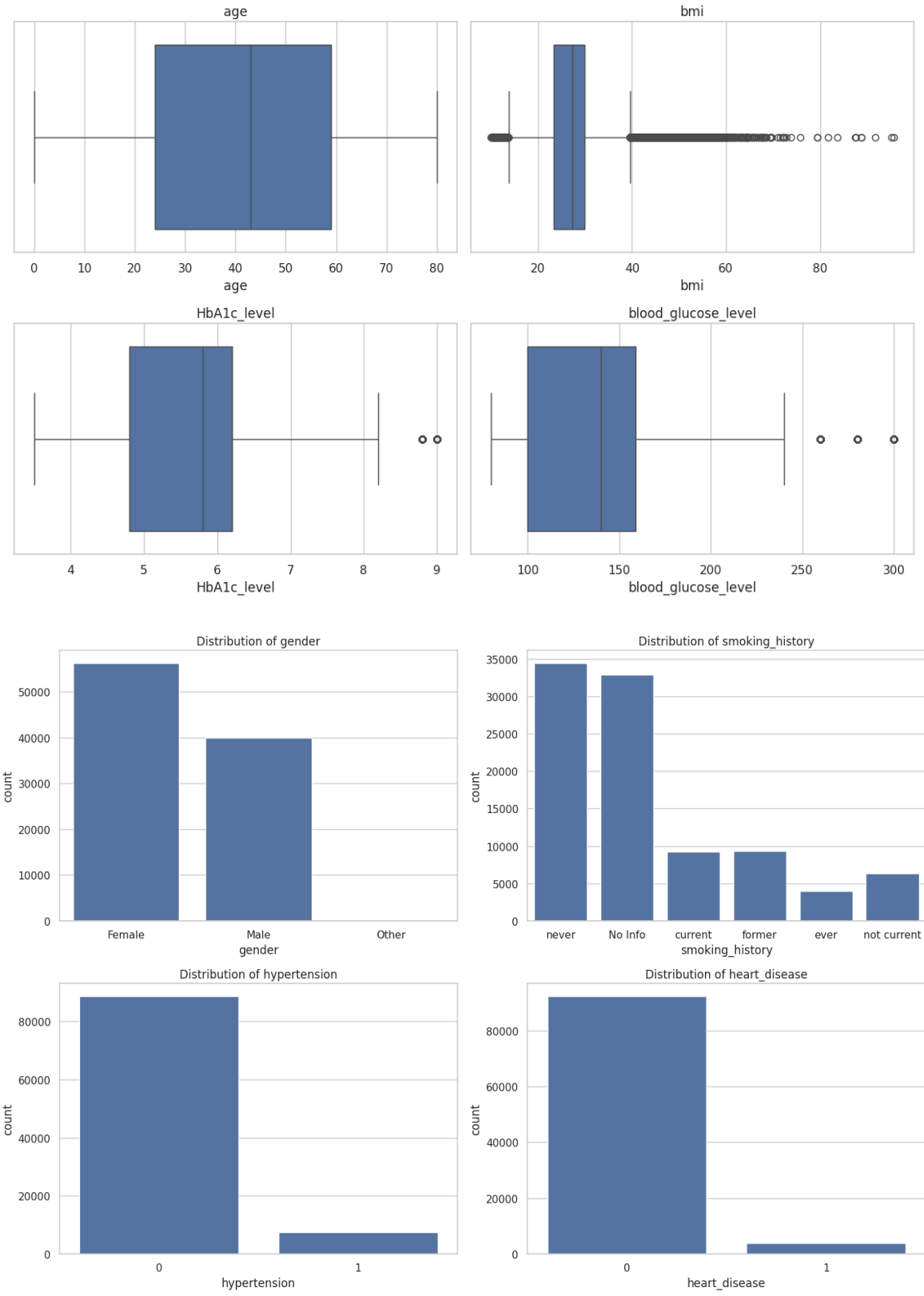
Exploratory Data Analysis

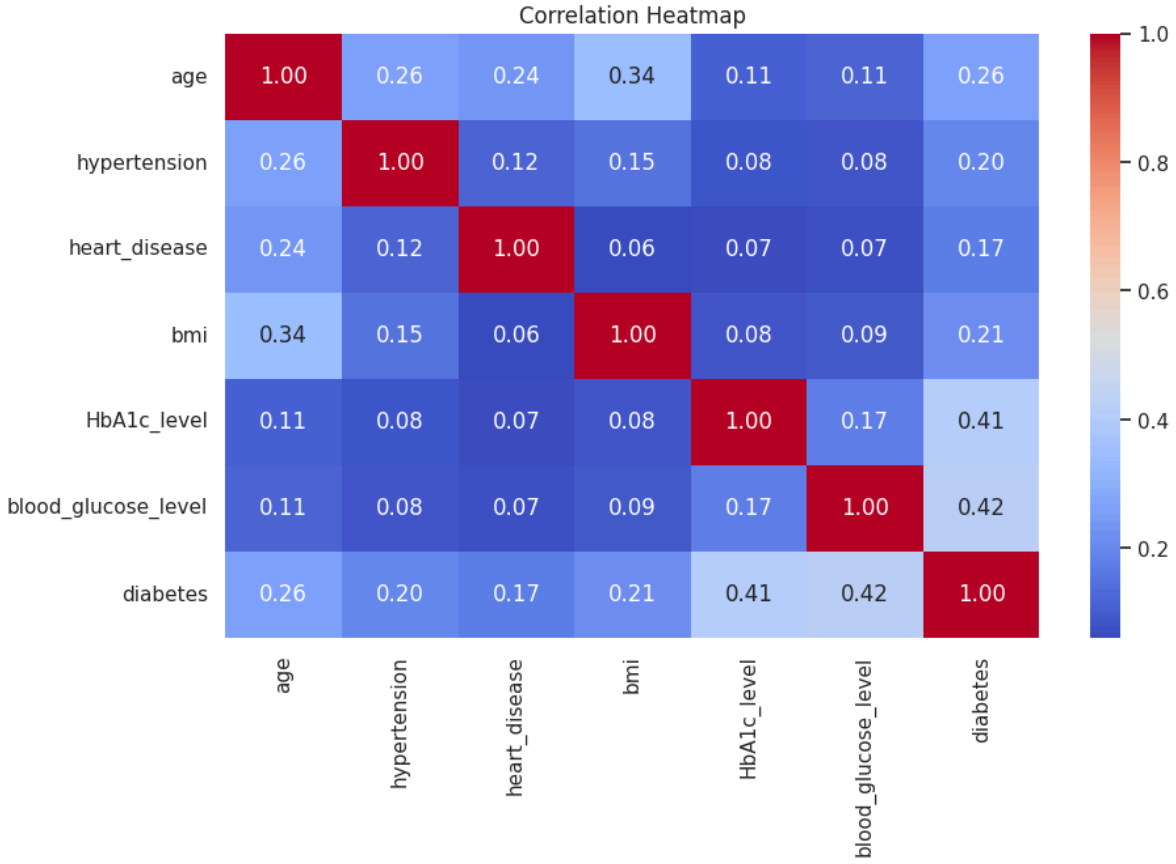
An exploratory data analysis (EDA) was conducted to understand the numerical and categorical features in the dataset, their distribution, and their relationship with each other before the modeling phase, to determine if any special handling was needed to build an unbiased model.

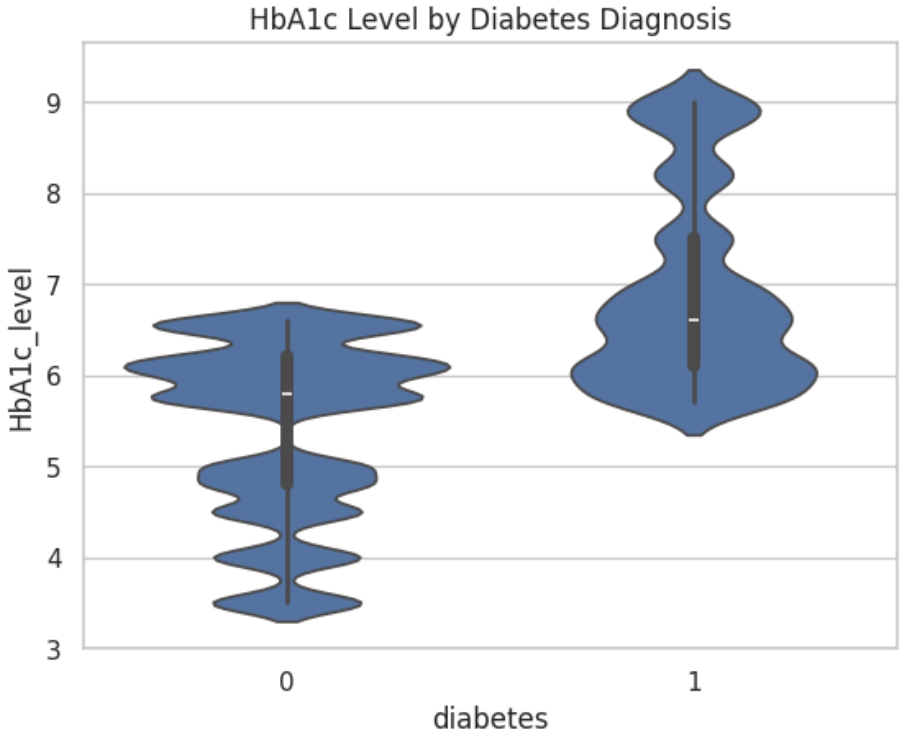
- The dataset is highly imbalanced: 91.18% of patients do NOT have diabetes, while only 8.82% do. Hence, SMOTE is needed to address this.
- Age, HbA1c level, and blood glucose level show relatively normal distributions.
- BMI is right-skewed, indicating the dataset contains a large proportion of overweight individuals.
- Gender distribution is predominantly female.
- Smoking history is dominated by “No Info” and “Never Smoked”.
- Most patients do not have hypertension or heart disease, suggesting these conditions are less represented in the dataset.
- Blood glucose level (0.42) and HbA1c level (0.41) show the strongest positive correlations with diabetes diagnosis. This aligns with already known medical research. The correlations are moderate.
- Age, hypertension, heart disease, and BMI show weaker correlations. They contribute risk, but in the context of linear correlation, they have a small contribution to diabetes diagnosis.
- Violin plot of HbA1c by diabetes diagnosis: Diabetic patients have higher HbA1c values.
- Blood glucose boxplot by diagnosis: A clear upward shift for diabetic patients, confirming glucose is a major predictor.
- Age distribution by diagnosis: The likelihood of diabetes increases with age.
- Smoking history vs. diabetes: No strong visual or statistical pattern noticed, indicating smoking history may not be a strong predictor in this dataset.

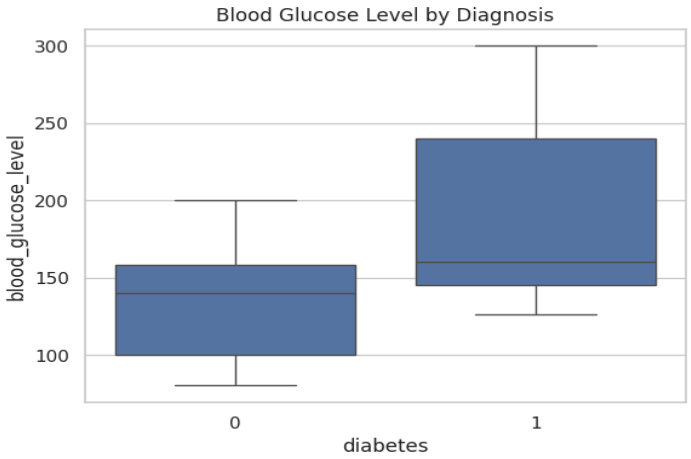

Diabetes Class Distribution
count
80000
60000
40000
20000
0
0
1
diabetes

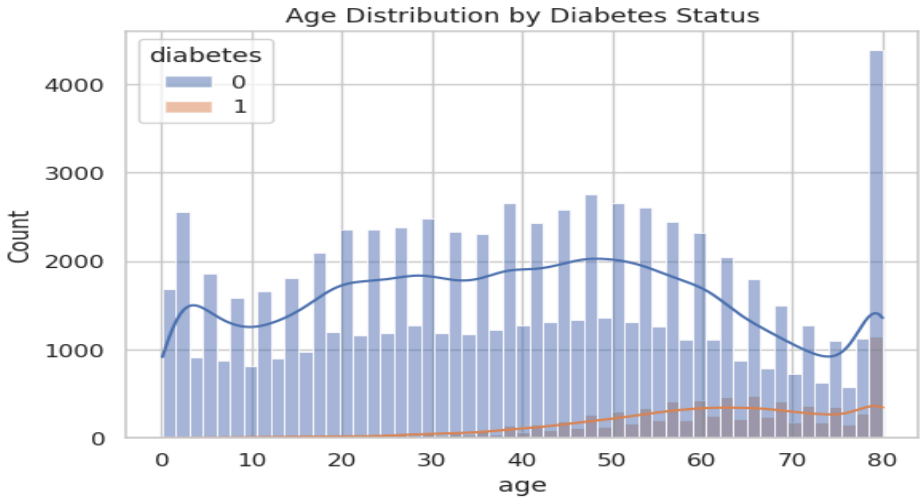

Numeric Feature Distributions
age
bmi
HbA1c_level
blood_glucose_level

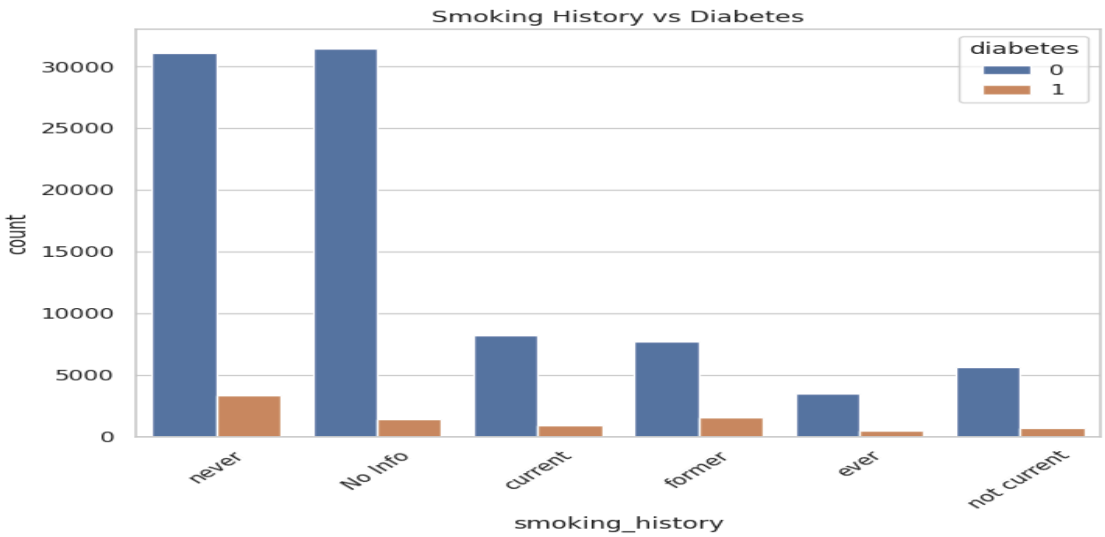

age
0 10 20 30 40 50 60 70 80
age
bmi
20 40 60 80
bmi
HbA1c_level
4 5 6 7 8 9
HbA1c_level
blood_glucose_level
100 150 200 250 300
blood_glucose_level
Distribution of gender
count
0 10000 20000 30000 40000 50000
Female Male Other
gender
Distribution of smoking_history
count
0 5000 10000 15000 20000 25000 30000 35000
never No Info current former ever not current
smoking_history
Distribution of hypertension
count
0 20000 40000 60000 80000
0 1
hypertension
Distribution of heart_disease
count
0 20000 40000 60000 80000
0 1
heart_disease


Correlation Heatmap
age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes
1.00 0.26 0.24 0.34 0.11 0.11 0.26
0.26 1.00 0.12 0.15 0.08 0.08 0.20
0.24 0.12 1.00 0.06 0.07 0.07 0.17
0.34 0.15 0.06 1.00 0.08 0.09 0.21
0.11 0.08 0.07 0.08 1.00 0.17 0.41
0.11 0.08 0.07 0.09 0.17 1.00 0.42
0.26 0.20 0.17 0.21 0.41 0.42 1.00
1.0
0.8
0.6
0.4
0.2


HbA1c Level by Diabetes Diagnosis
HbA1c_level
9
8
7
6
5
4
3
0
1
diabetes


Blood Glucose Level by Diagnosis
blood_glucose_level
300
250
200
150
100
0
1
diabetes


Age Distribution by Diabetes Status
diabetes
0
1
Count
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
age


Smoking History vs Diabetes
diabetes
0
1
count
30000
25000
20000
15000
10000
5000
0
never
No Info
current
former
ever
not current
smoking_history

Methodology

Data Preprocessing

The first step was to clean and prepare the dataset for modeling. To ensure that the large dataset chosen for this project does not contain missing values before the modeling begins, missing numerical values were filled using the median, which prevents extreme values from influencing the results. Missing categorical values were filled using the most frequent category to avoid introducing new groups during modeling.

Next, categorical variables such as gender and smoking history were converted into a numerical format through one-hot encoding, allowing machine-learning models to interpret these fields properly. All numerical columns, including age, BMI, HbA1c level, blood glucose level, hypertension, and heart disease, were standardized using a StandardScaler. Scaling places all numerical variables on a similar range, which helps models like Logistic Regression perform more effectively.

Then the data was split into 80% training and 20% testing using stratified sampling. Stratification was meant to maintain the natural class imbalance, where only about 8% of patients had diabetes. Class imbalance was handled later with SMOTE.

Feature Engineering

The goal was to keep the integrity of the medical data, clean it, address the class imbalance, and prepare it for model development. Hence, feature engineering was minimal, and no synthetic features were created. Age, BMI, blood glucose, and HbA1c, which are known to highly contribute to diabetes, were already present in the dataset. Preprocessing steps like encoding and scaling effectively transformed the raw dataset into a machine-learning-appropriate dataframe.

Models Selected

- Logistic Regression ***(baseline model)***
    - Useful for understanding linear relationships.

    - Helps us understand where we have started and where we should go.
- Decision Tree Classifier
    - Captures non-linear relationships.
    - Performs well even when features interact with each other.
- Random Forest Classifier
    - An ensemble of decision trees.
    - Reduces overfitting and improves stability.

Evaluation Metrics

The following metrics were used to find the best-balanced model:

- Precision: How many predicted diabetics were truly diabetic?
- Recall: How many actual diabetics were correctly identified?
- F1-score: A balance between precision and recall.
- Support: The number of samples in each class.

Recall and F1-score were especially important because missing a true diabetic case (a false negative) is more harmful in a real healthcare setting than a false positive.

Hyperparameter Tuning Approach

The hypermeter tuning approach for this project was simple, without making it too complex, as default settings were tested first, followed by small adjustments to improve performance.

For the Decision Tree and Random Forest models, parameters such as max_depth, min_samples_split, and min_samples_leaf were adjusted to improve generalization and reduce overfitting, but were not increased significantly due to performance delays in Google Colab. SMOTE was applied within each pipeline to handle class imbalance.

Results & Model Comparison

Model Comparison Table (Precision, Recall, F1-Score, Accuracy)

Class-Level Metrics

0 = Non-diabetic, 1 = Diabetic

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression (SMOTE)	0	0.99	0.89	0.93	17,534
	1	0.43	0.88	0.57	1,696
Decision Tree (SMOTE)	0	0.98	0.94	0.96	17,534
	1	0.56	0.83	0.67	1,696
Random Forest (SMOTE)	0	0.99	0.91	0.95	17,534
	1	0.49	0.87	0.63	1,696

Overall Model Performance

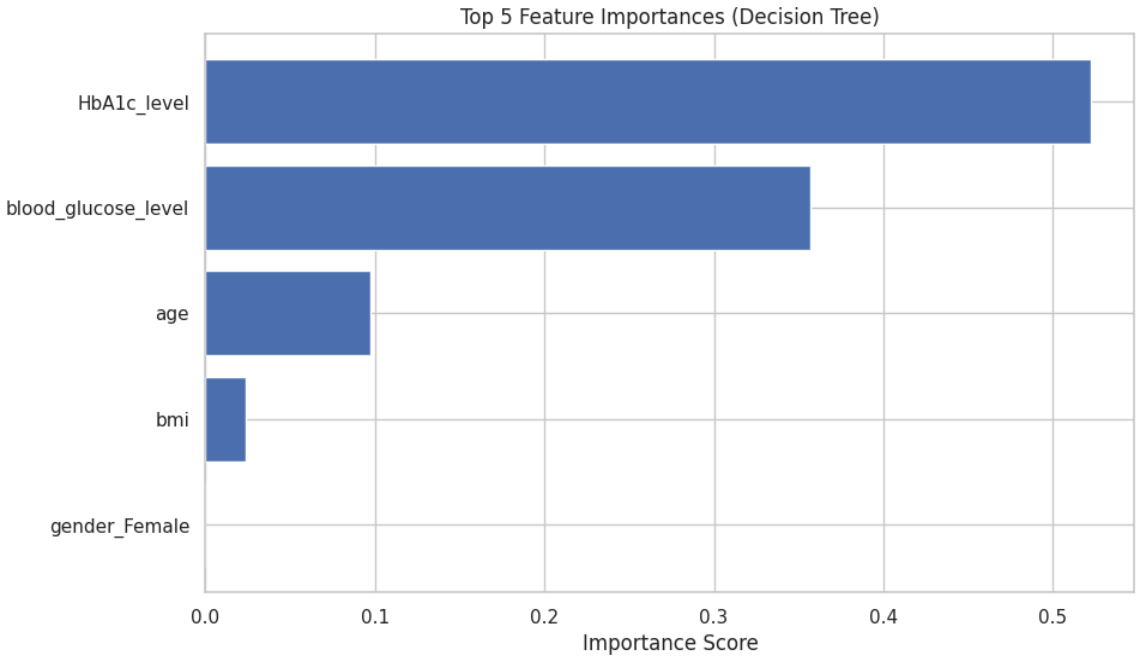
Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic Regression (SMOTE)	0.88	0.71	0.88	0.75	0.94	0.88	0.90
Decision Tree (SMOTE)	0.93	0.77	0.88	0.81	0.95	0.93	0.93
Random Forest (SMOTE)	0.91	0.74	0.89	0.79	0.94	0.91	0.92

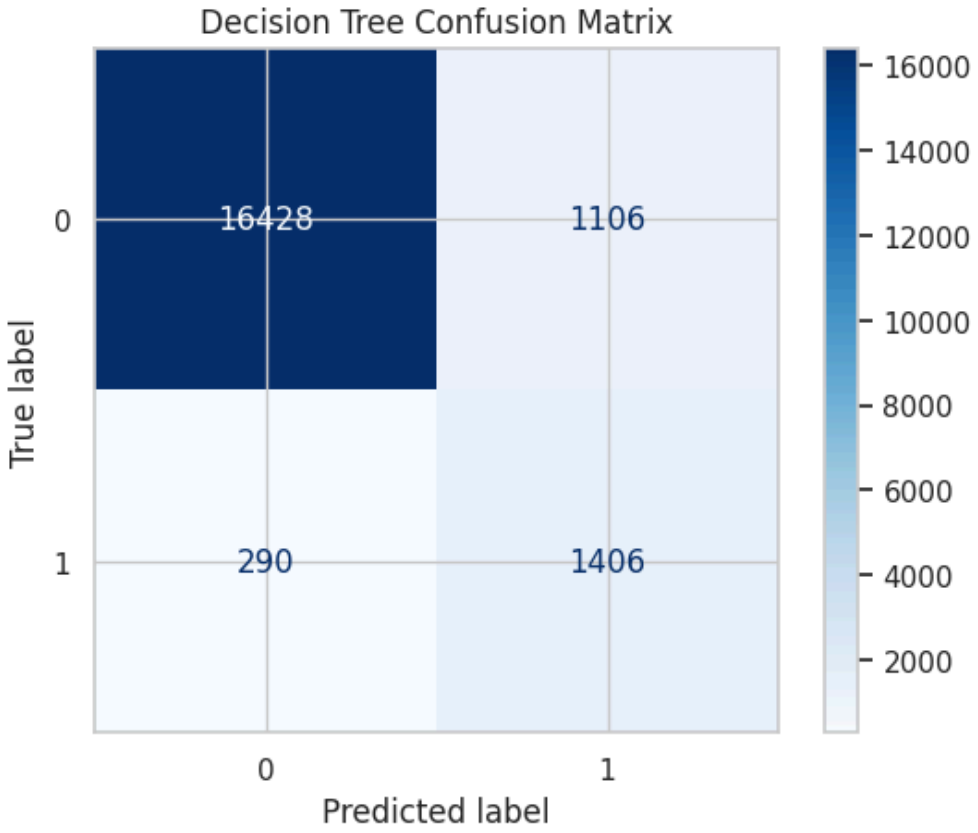
Discussion on Model Comparison & Final Model Choice

- Before training each model, the class imbalance identified in earlier phases was addressed using SMOTE (Synthetic Minority Oversampling Technique). SMOTE was applied only on the training data within each model pipeline to create synthetic samples of diabetic patients, ensuring the model learned from a more balanced dataset.
- Three models trained and compared: Logistic Regression, Decision Tree, and Random Forest.
- All models used the same preprocessing pipeline, where numerical features were scaled, and categorical features were properly encoded.
- Logistic Regression was good at catching diabetic patients, but it also made a lot of false alarms. It had a strong recall for both classes, but the precision for diabetics was low.
- The Decision Tree model showed a stronger overall balance, with higher precision and recall for both classes, being more efficient than the linear model.
- The Random Forest model also performed well. However, when comparing precision, recall, F1-score, and overall accuracy, the Decision Tree had the most balanced performance between predicting diabetic and non-diabetic patients.

Hence, based on performance metrics and interpretability, the Decision Tree model was selected as the final model because it did the best job predicting both diabetic and non-diabetic patients.

Feature Importance & Confusion Matrix





Analysis of Final Model

Feature Importance

From our Decision Tree Model, feature importance scores were used to determine which factors contributed most to predicting diabetes.

- The feature importance results showed that blood glucose level and HbA1c level were by far the most influential predictors in the model. This aligns strongly with medical research.
- Age and BMI had smaller but still meaningful contributions, indicating that lifestyle and health conditions also play supporting roles in diabetes risk.
- Female patients were the 5th in the feature importance; even if the score was extremely small, it should be noted. The dataset contained more female patients than men.

Confusion Matriz

To better understand the model's strengths and weaknesses, the confusion matrix was examined.

- The Decision Tree performed well at correctly identifying diabetic patients, meaning it successfully captured individuals who are at risk. However, because diabetes is relatively rare in the dataset, some false positives occurred when patients were predicted as diabetic when they were not. While it is not ideal, we can accept it in a healthcare context, where missing a true diabetic case would be far more harmful than issuing an extra precautionary follow-up.

Business Insights

- The final decision tree model can support early diabetes screening by flagging individuals with higher glucose or HbA1c levels for further testing and support. Healthcare providers could use the model to prioritize outreach for older patients or those with high BMI, even if glucose and HbA1c levels are borderline.
- Public health departments could also use these predictions to promote educational campaigns toward populations at higher risk. It will be a great way to keep the communities informed with recent findings and encourage healthier lifestyles.
- The final model demonstrates how patient data can help identify at-risk individuals earlier, potentially reducing long-term healthcare costs and improving patient outcomes. In the current climate in the United States with skyrocketing health insurance premiums, a model is needed to ensure that healthcare costs can be reduced.

Ethics and Responsible AI

- Privacy and security are extremely important in this project because it involves working with medical data.
- In real-world clinical settings, patient information must follow strict regulations such as HIPAA. A healthcare analytics worker should be analyzing data that has patient-identifying details removed. It also should be stored securely and only shared when proper permissions have been granted.
- A major consideration should be how the model is used in practice. A machine-learning model should never act as the final decision-maker. Even though the model identified important predictors like blood glucose and HbA1c levels, it is not a replacement for a medical professional. Instead, it should be treated as a supportive screening tool that helps highlight patients who may need additional testing or support.
- Any deployment of this model would need clear disclaimers and usage guidelines approved by senior healthcare analytics and medical professionals.

Conclusion and Future Work

This project successfully applied the full predictive analytics workflow to build a diabetes prediction model using medical data. The Decision Tree emerged as the most balanced and reliable model for identifying both diabetic and non-diabetic patients.

The model results were supported well by established medical knowledge, especially in recognizing blood glucose and HbA1c levels as the strongest predictors. Overall, the project demonstrated how machine learning can support early diabetes screening and help healthcare providers prioritize patients who may need additional testing. Even if the model performed well, there are several limitations. The dataset is highly imbalanced, with only about 8% of patients having diabetes, which makes prediction more challenging even after applying SMOTE.

In addition, another limitation is the lack of important medical features that could strengthen the model, such as family history of diabetes, physical activity levels, diet, medication use, and genetic factors. Additionally, the project used only a few machine-learning algorithms, and deeper hyperparameter tuning was limited due to computational constraints in Google Colab.

Future work could focus on gathering more diverse and comprehensive patient data, especially features known to influence diabetes risk, and having a gender-balanced dataset. Hence, future research could delve into whether diabetes diagnosis-influencing features vary by gender. Also, trying advanced models such as Gradient Boosting, XGBoost, etc., may also improve performance. Conducting more thorough hyperparameter tuning and exploring other approaches to handle class imbalance could lead to stronger results, especially on powerful Python software like PyCharm and Spyder that will support deeper hyperparameter tuning.

This project provided important lessons in working with imbalanced datasets, interpreting model performance, and translating technical findings into meaningful healthcare insights. The project also reinforced the importance of ethical considerations when building models that may impact real people and medical decisions.

References and AI Acknowledgment

Dataset

Diabetes Prediction Dataset on **Kaggle**.

Code Notebook

Project Code Notebook detailing Phases 1-6 on **Google Collab**.

Research Articles

EIT Health. (2024, September 18). Machine learning in healthcare: Uses, benefits, and pioneers in the field. https://eithealth.eu/news-article/machine-learning-in-healthcare-uses-benefits-and-pioneers-in-the-field/

Holliday, C. S., & Gabbay, R. A. (2025). Breaking barriers: CDC and American Diabetes Associations unite to combat diabetes. Preventing Chronic Disease, 22, 240273. https://doi.org/10.5888/pcd22.240273

World Health Organization. (2024). Diabetes [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/diabetes

Tutorial Used

Build a predictive model in Python: https://365datascience.com/tutorials/python-tutorials/predictive-model-python/

GeeksforGeeks: Evaluation Metrics in Machine Learning:https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

Imbalance Learn: User Guide:https://imbalanced-learn.org/stable/user_guide.html#user-guide

Scikit-learn: Machine Learning in Python: https://scikit-learn.org/stable/

Code References

Homework Assignment, California Housing Price Prediction:https://harslan.github.io/predictive-analytics/assignments/california-housing/

Imbalance Learn: SMOTE: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

Logistic Regression Model: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/#

Decision Tree Model:
https://www.geeksforgeeks.org/machine-learning/decision-tree-implementation-python/

Random Forest Model:
https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

Other References & AI Assistance Acknowledgment

Students completed DataCamp (Intro to Python & Intermediate Python courses), and ISOM 631 (Spring 2025) before taking these courses.

All of the code written for this project was heavily influenced by ISOM 631 and ISOM 835 coursework. However, constant Google searches were conducted to understand various concepts and fix errors. Also, Google Collab's Gemini assistance and coding suggestions were used to write the code and fix errors.

For the purpose of learning Class imbalance handling, learning how to use SMOTE and Streamlit deployment, Google AI Mode, and ChatGPT were also used.