

#23: Predicting Early-Stage Diabetes in Individuals Using Machine Learning

Executive Summary

With roughly 12% of Americans living with diabetes, it has grown to be a dangerous contributor to many life-threatening conditions [1]. The utilization of machine learning techniques allows for better detection of this disease, which increases the ability to make more informed healthcare choices for patients. The model constructed considers basic human data such as age and gender, along with symptoms experienced by each patient to make predictions on if the patient has or does not have diabetes. Comparing three popular discrete machine learning models — **Random Forest Classification**, **Logistic Regression**, and **XG Boost Classification** — the model's predictions had an accuracy of 99%, 92%, and 97%, respectively.

Background

With the level of Americans being diagnosed on the rise, up almost 10 million over the past decade[1], it is imperative to give clinicians a quick and easy screening tool to determine if further testing is necessary.

By using machine learning techniques to act as an early indicator, both patients and clinicians can save significant money and time. By doing so, patients can fill out the short questionnaire provided, and clinicians can receive an instantaneous prediction. After analyzing this prediction, they can then decide whether further testing is needed.

This technique was performed using Jupyter Notebook (Opensource Python), with a highlight in the Numpy and Seaborn libraries. Through a series of testing, we were able to determine that using the Random Forest Classification model returned the most accurate values for our given dataset.

Goal

Our group's mission was to provide healthcare professionals a **quick** and **affordable** tool to pre-screen patients for diabetes in order for hospitals save time and effort on unnecessary testing. By doing this, funds and manhours could be allocated to other lifesaving areas.

Experimental Setup

To begin the design of our predictive system, we examined the *Early Stage Diabetes Risk Prediction Dataset* from the UCI Machine Learning Repository [2], which contains 520 patient records with demographic information and symptom-based indicators. The **dependent variable** is class, a binary label identifying whether a patient has early-stage diabetes ($1 = \text{Positive}$, $0 = \text{Negative}$). **Independent variables** include age, gender, and fourteen symptom-based predictors such as polyuria, polydipsia, sudden weight loss, weakness, visual blurring, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. **Preprocessing steps** involved encoding all categorical variables into numerical values, scaling the features for consistency across models, and applying a 75/25 train-test split; no missing values were present. With these specifications and constraints defined, our conceptual design focused on selecting three machine learning approaches—Logistic Regression, Random Forest Classification, and XGBoost—to compare linear and nonlinear decision boundaries. These models were chosen for their proven effectiveness in medical classification tasks and their compatibility with the dataset's structure.

Experimental Test Plan

To understand the correlation between physical factors and diabetes for at-risk patients, **three** different **categorical** machine learning techniques were defined and identified. To implement these three machine learning techniques, they were all run in a **Python** environment. From there, plot diagrams and statistical information were extracted from the data and analysis of the results was performed to find the best prediction method.

Random Forest Classification: Machine learning algorithm that uses a collection of decision trees to make a final prediction using a majority vote. It is best suited for nonlinear relationships, moderate dataset sizes, and datasets with a mix of categorical and numerical features. The current dataset fits this description, which is why it is best suited for the situation.

XG Boost Classification: Machine learning method that attempts to predict data using sequential learning with shallow decision trees that removes errors in each iteration. It finalizes its prediction using the weighted sum of all its decisions. It works best with nonlinear relationships and mixed data types. Due to its sequential nature, it has a high predictive accuracy. However, the sequential learning of this algorithm results in all decision trees being additive and not independent of one another.

Logistic Regression: Machine learning algorithm that works best for dataset classifications that are linearly separable. Logistic regression has the strengths of being fast to train, working well with larger datasets, and having probabilistic outputs based on a sigmoid function. It has 2 types: one for binomial logistic regression and multiclass logistic regression. Since this dataset only requires "yes" and "no" classifications, the binomial system is more appropriate.

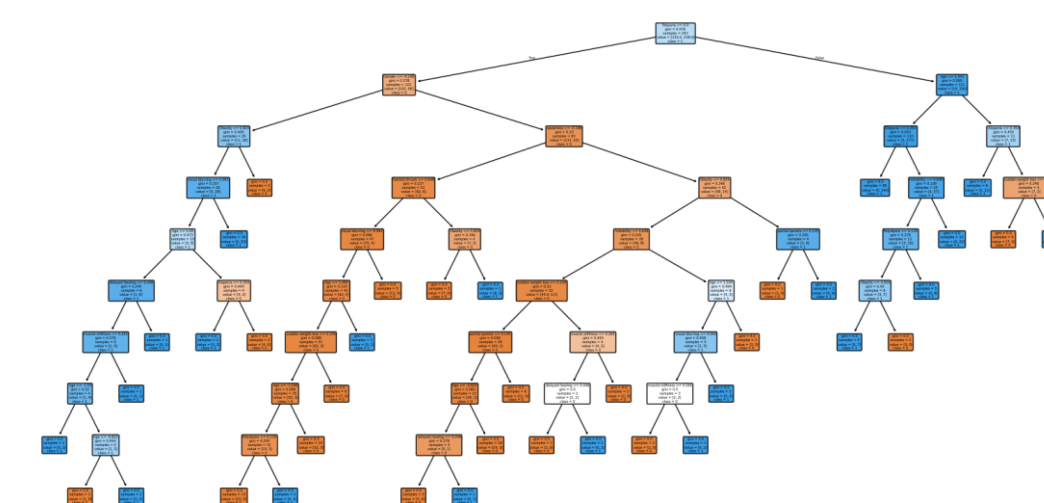


Figure 1. Random Forest Classification Example Decision Tree

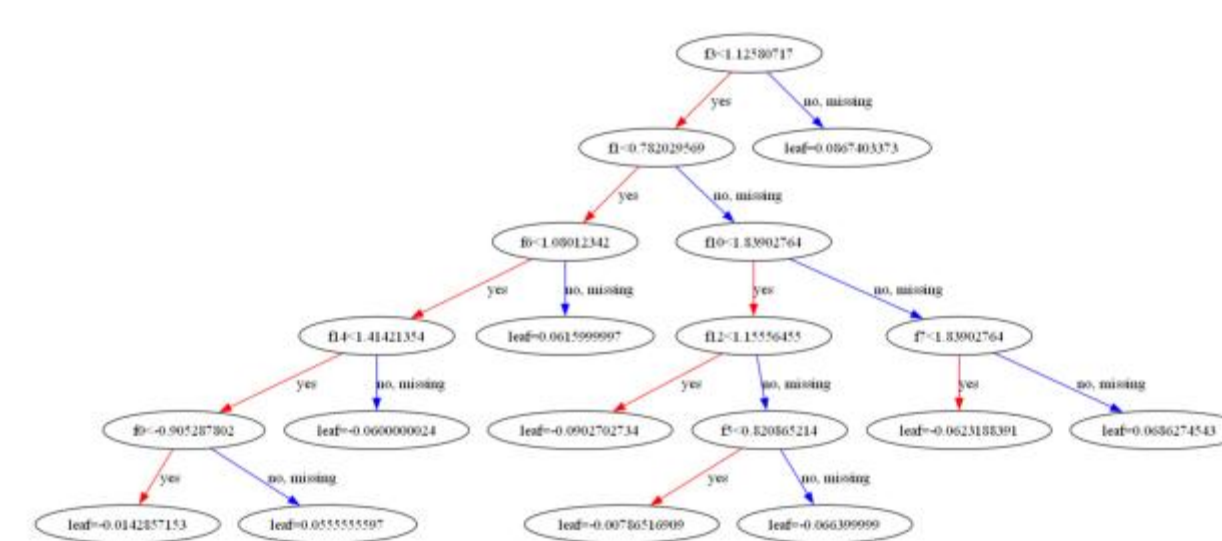


Figure 2. XGBoost Example Decision Tree

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Equation 1. Logistic Regression Sigmoid Function

Experimental Results

Final Model Scores

Model	Accuracy	F1
Random Forest	0.992	0.994
Logistic Regression	0.915	0.928
XGBoost	0.969	0.975

Best Parameters

Model	Parameter	Value
Random Forest	n_estimators	200
Logistic Regression	C	1
XGBoost	n_estimators	300

Comparing Receiver Operating Characteristic (ROC) Curves

- Random Forest Classification was identified as the **best predictive model** with an area under the curve (AUC) of 0.999, with XGBoost close behind at 0.998. Logistic Regression was also determined as a reliable predictive model at 0.961.
- This indicates that all three classification models performed strongly, but Random Forest provided the **best** overall balance of **predictive accuracy** and **class discrimination**.

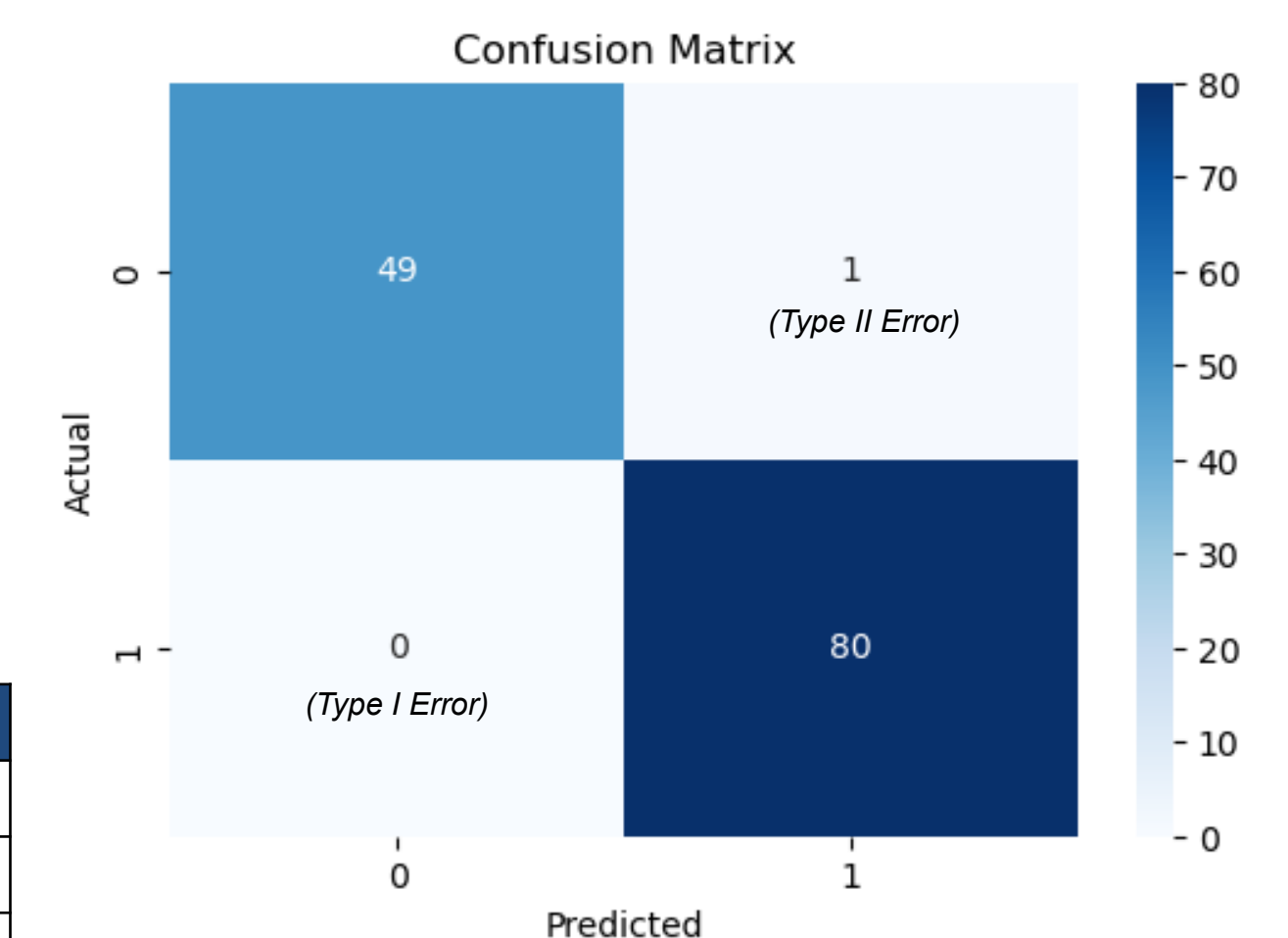


Figure 3. Random Forest Confusion Matrix

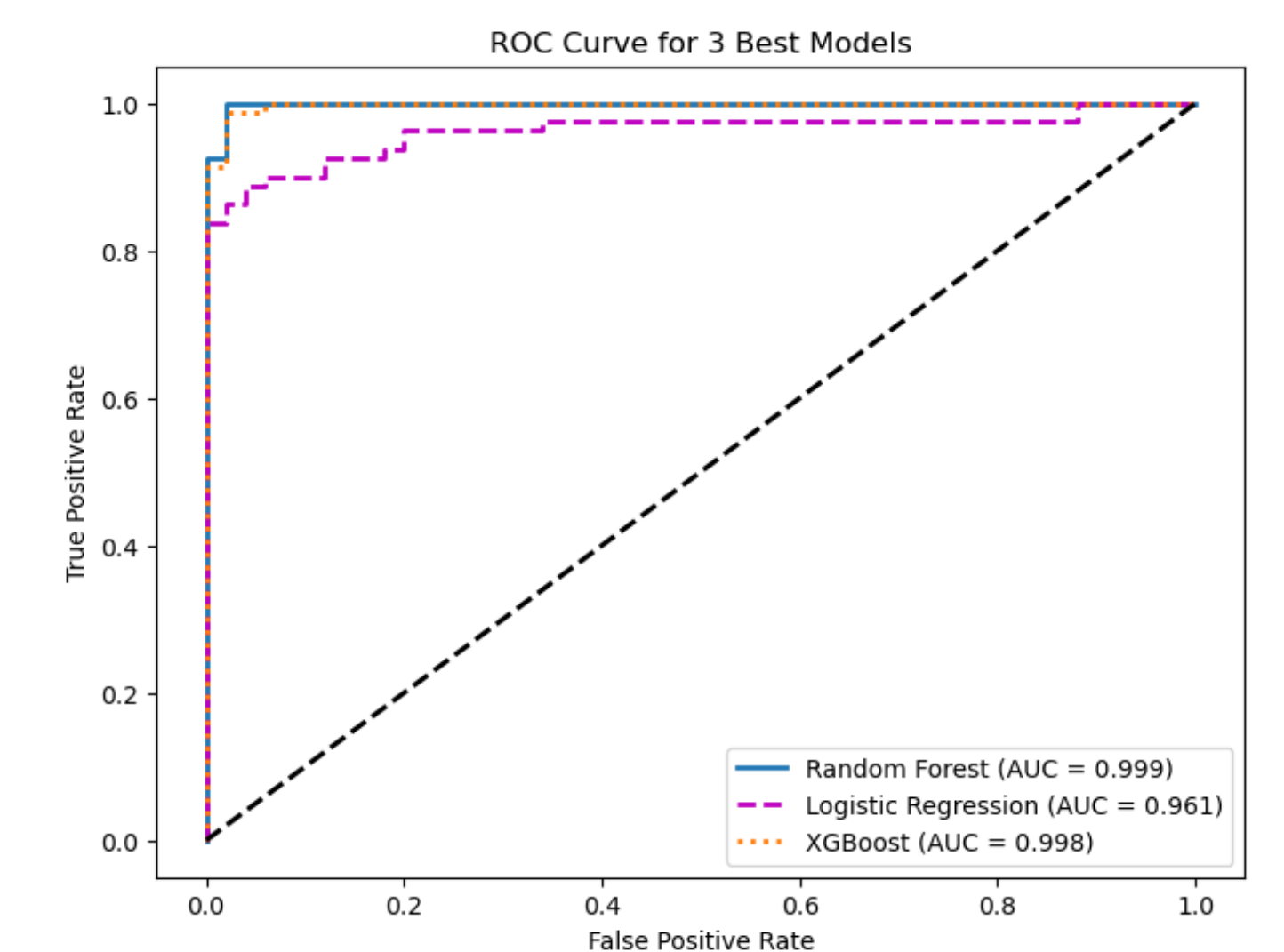


Figure 4. ROC for 3 models

Conclusions

Highlights

- Processed given dataset by encoding all categorical values into numerical terms.
- Scaled data to ensure model reliability without outliers majorly impacting results.
- Applied three different machine learning models using a 75/25 train-test split to compare accuracy, F1, and AUC values.
- Identified the best performing machine learning model using graphical elements such as ROC and confusion matrix.

Future Work

- Identify** outliers in dataset using z-score or +/- three standard deviations to improve accuracy.
- Utilize** larger samples to ensure robustness of model.
- Validate** best parameters to clean up dataset further.

References

- CDC. (2024, May 15). *National diabetes statistics report*. Centers for Disease Control and Prevention. <https://www.cdc.gov/diabetes/php/data-research/index.html>
- Rawat, S., & Chakoli, A. *Early Stage Diabetes Risk Prediction Dataset*. UCI Machine Learning Repository.