Classifying Diabetes Conditions Using Various Machine Learning & Artificial Intelligence Models

Eddie McGowan: Lehigh University Department of Data Science

https://github.com/EddieMcGowan/Diabetes_Classification_Model

Abstract

This project investigates the use of supervised machine learning to classify various types of diabetes and prediabetic conditions. Using a 70,000-record dataset from Kaggle, I trained and evaluated ten classification models, prioritizing both predictive performance and interpretability for non-technical audiences. While the XGBoost model delivered the highest accuracy and fastest training time, its complexity limits practical interpretability. In contrast, the stacking model achieved nearly identical performance and offers greater explainability, though at the cost of a longer training duration.

Problem Description/Motivation

Societal Motivation: Over 800 million people globally are affected by diabetes, a condition that, if left unmanaged, can lead to serious, life-threatening complications. For this project, I developed ten classification models that predicts the type of diabetes or prediabetic condition a patient may have based on both genetic markers and environmental factors. As a professional currently working in healthcare, my goal is for this model to eventually assist doctors in diagnosing a patient's condition, enabling more effective preventive measures or targeted treatment strategies.

Practical Motivation: I currently work in data analytics but am transitioning into more advanced data science projects. Since no single model fits all use cases, this project was designed to explore a range of classification models, giving me practical exposure to their strengths and limitations. This experience will inform more strategic model selection for future corporate data science initiatives.

Objective and Parameters

Goal: Using ten classification model, find the model which best predicts the patient's condition.

Evaluation Metrics:
- Accuracy
- F1 Score
- Training Time
- Testing Time
- Interpretability for non-technical audiences.

Train Test: 5-fold cross validation on all models
Hyperparameter Tuning: Grid search for all learning models
Graph: Created feature importance plot for relevant models

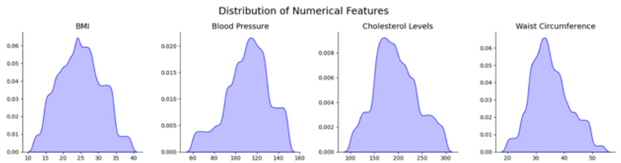
Dataset Description

- **Source**: Kaggle Diabetes Dataset
- **Size**: 70,000 records and 34 Features

Relevant Features:
- **Target**: The type of diabetes or prediabetic condition
- **Genetic Markers**: Indicators of genetic predisposition to diabetes
- **Autoantibodies**: Presence of autoantibodies commonly associated with autoimmune diabetes
- **Family History**: Information on whether there is a known family history of diabetes
- **Environmental Factors**: Details about environmental influences that may contribute to diabetes
- **Insulin Levels**: Measured insulin levels in patients
- **Age and BMI**: Demographic information including age and Body Mass Index (BMI)
- **Physical Activity and Dietary Habits**: Lifestyle factors that could influence diabetes risk

EDA/Preprocessing

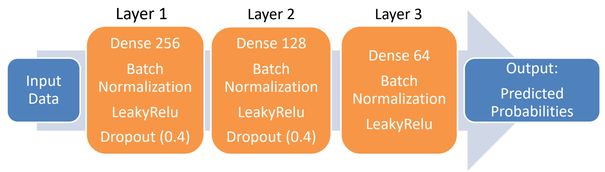
- Analyzed feature distributions to check for outliers
- Created a correlation matrix to check for collinearity
- Chi-square test confirmed obese patients are significantly more likely to have type 2 diabetes (p = 0.0): validating data

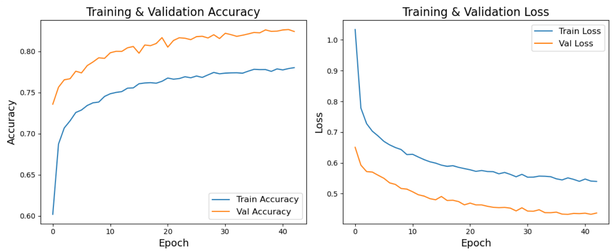


Models Tested & Parameters Optimized

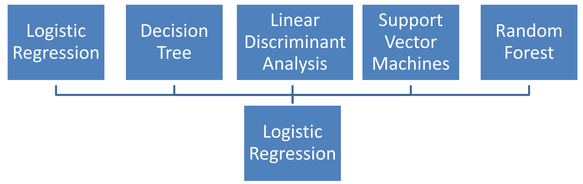
Baseline Models
- **K-Nearest-Neighbors**: n_neighbors
- **Logistic Regression**: c, solver
- **Decision Tree**: max_depth, criterion, min_samples_split

Advanced Methods
- **Random Forest:** n_estimators, max_depth, min_samples_split
- **Linear Discriminant Analysis**
- **Support Vector Machines (LinearSVC):** c
- **XGBoost**: n_estimators, max_depth, learning_rate
- **MultiLayerPerceptron**: LabelEncoding





Custom
- **Ensemble Voting**: a majority vote classifier utilizing the k-nearest neighbors, logistic regression, decision tree, random forest, linear discriminant analysis, and support vector machine models
- **Stacking**: logistic regression, decision tree, linear discriminant analysis, support vector machine, and random forest models are input to a logistic regression mode

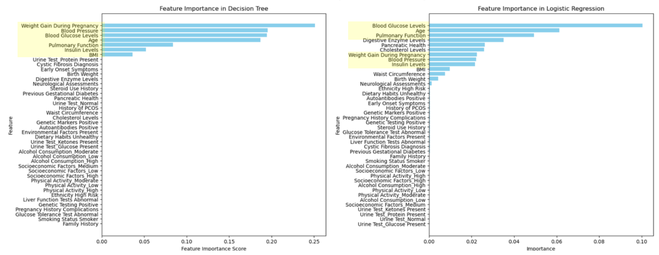


Evaluation: Quantitative Summary

Model	Epochs or Trees	Accuracy (%)	F1 Score (%)	Training Time (s)	Testing Time (s)	Interpretability
K-Nearest Neighbors	NA	67	67	234.5	98.83	Great
Logistic Regression	100	77	77	79.0	0.03	Great
Decision Tree	NA	87	87	2.2	0.01	Great
Random Forest	100	90	90	32.8	0.75	Average
Linear Discriminant Analysis	NA	74	74	0.4	0.07	Average
Support Vector Machines (LinearSVC)	55	40	35	229.6	0.02	Average
XGBoost	100	91	91	318.0	1.16	Poor
Multilayer Perceptron	55	84	84	306.7	1.14	Poor
Ensemble Voting	NA	83	83	578.6	93.16	Average
Stacking	2000	89	89	377.7	0.99	Average

Evaluation (Key Findings)

- **XGBoost**: Highest accuracy and F1 score (~91%); fast training (93s); difficult to explain to non technical audiences
- **Stacking Model**: Comparable accuracy to XGBoost; training time longer (2 minutes); easier to explain compared to XGBoost
- **Random Forest**: Strong performance with balanced tradeoff between speed and interpretability
- **MLP (Keras)**: Good performance, but below XGBoost; required one-hot encoding for labels
- All other models had comparatively poor accuracy and F1 Scores
- Across all models, the same key features consistently ranked at the top of the feature importance plots



Future work

- Apply to different diseases/conditions (e.g., neurological disorders or cancers)
- Integrate the model with the suggested next steps for the patient