AI Boot Camp Project 2

Group 3: Diabetes Predictor ML

Team Members:

Michael Cirino
Alexander Iruthaya
Jaylen Guevara-Kirksey Bey
Jean Clark

Project Overview

Project Purpose

The problem we chose to solve is applying machine learning to look into what features are predictive for diabetes diagnosis. Our dataset was from patients who self-reported 18 different data points with over 70,000 participants.

Project Overview

Goals to be answered

- **Goal 1** - Create a predictive model that can accurately predict if someone has potential for a diabetes diagnosis
- **Goal 2** - See what factors are most important and least important when predicting diabetes in future patients.
- **Goal 3** - Who is susceptible, or are most likely, to receive a diabetes diagnosis

Project Overview

Overview of data collection and cleanup

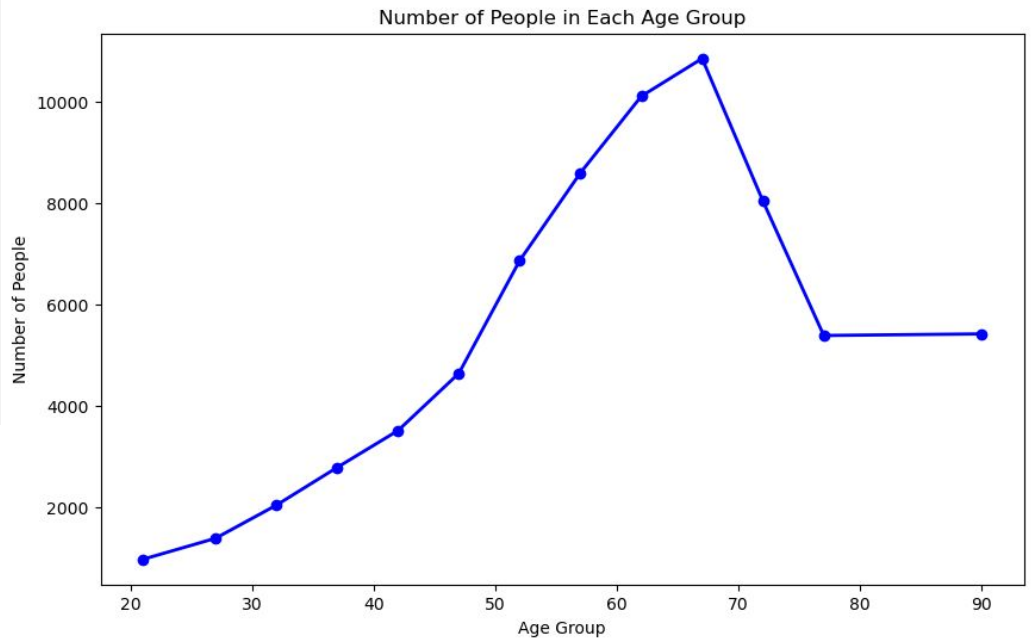
First we surveyed publicly available datasets on kaggle and open.gov

After parsing a few different sets, we ended on a survey conducted in 2015 by the Behavioral Risk Factor Surveillance System (BRFSS), a branch of the CDC.

- https://www.kaggle.com/datasets/prosperchuks/health-dataset?select=diabetes_data.csv

Cleanup

- Age Group - normalized ages into meaningful brackets with a map function that
- There were no null values in the data to remove
- Renamed Columns to be more accurately descriptive



Project Overview

Overview of data exploration process

We then took the cleaned data and dropped the “DiabetesDiagnosis” column, and split the data 80/20.

Created a pipeline for each model

- SVM, Log Regression, Decision Tree, KNN, Random Forest, Extra Trees, Gradient Booster, AdaBoost

Results of SkLearn Pipeline

Logistic Regression Training Accuracy: 0.747
Logistic Regression Test Accuracy: 0.746

SVM Training Accuracy: 0.747
SVM Test Accuracy: 0.745

Decision Tree Training Accuracy: 0.972
Decision Tree Test Accuracy: 0.658

K-Nearest Neighbors Training Accuracy: 0.798
K-Nearest Neighbors Test Accuracy: 0.716

Random Forest Training Accuracy: 0.972
Random Forest Test Accuracy: 0.728

Extremely Random Trees Training Accuracy: 0.972
Extremely Random Trees Test Accuracy: 0.712

Gradient Boosting Training Accuracy: 0.754
Gradient Boosting Test Accuracy: 0.753

Project Overview

Approach taken to achieve goals

Our approach to achieve our goals was to create a machine learning notebook.

To create this notebook we used tools including the Python programming language, Supervised Learning with the SciKit-Learn library, Pandas for data manipulation, and MatPlotLib for data visualization,

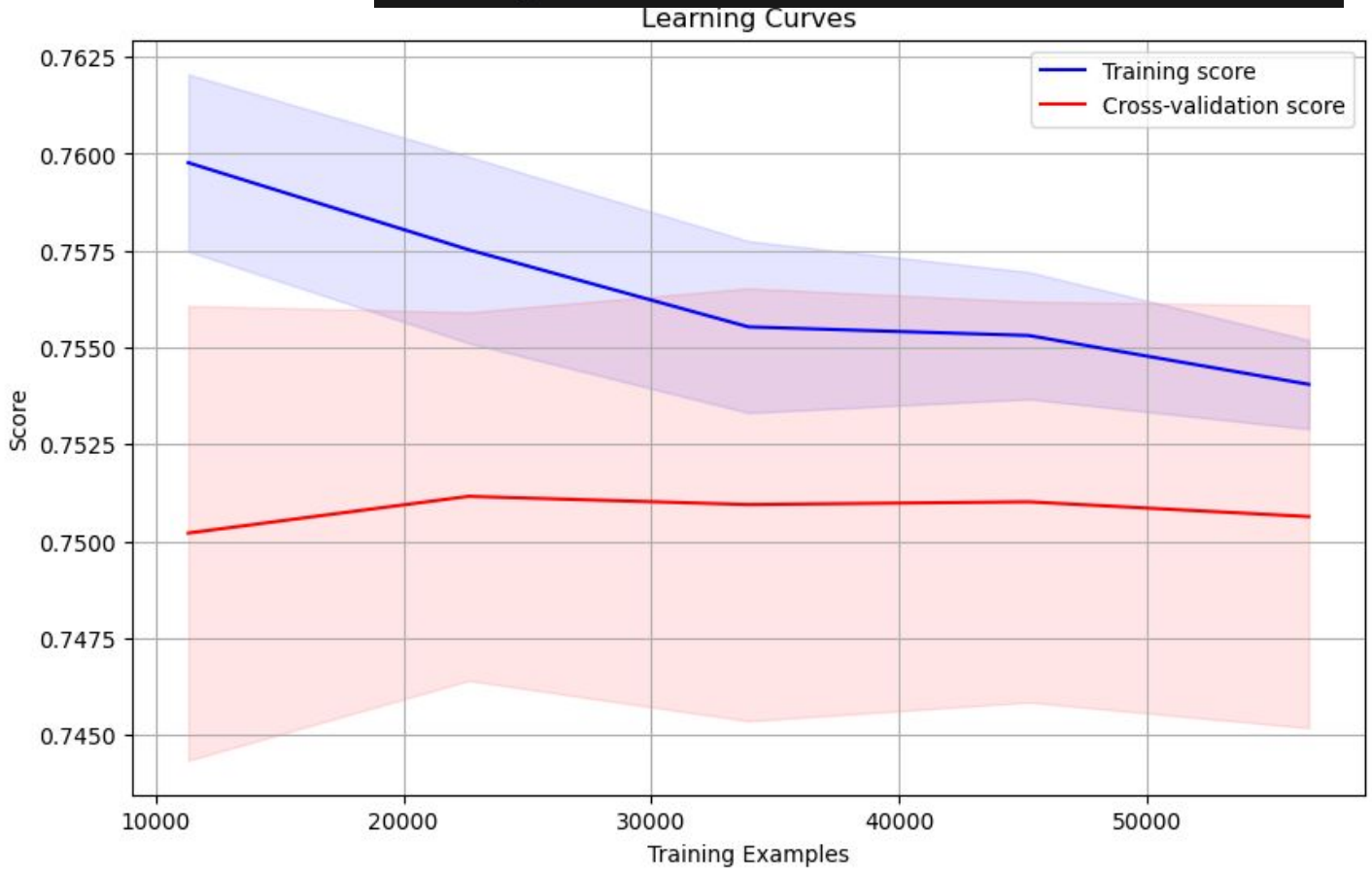
Goal 1: Create a Predictive Model

	precision	recall	f1-score	support
0.0	0.78	0.71	0.75	7090
1.0	0.73	0.80	0.77	7049
accuracy			0.76	14139
macro avg	0.76	0.76	0.76	14139
weighted avg	0.76	0.76	0.76	14139

Accuracy: 0.756

We were able to reach an f1-score of 0.76 for diabetes diagnosis with a mean accuracy of 75.6%

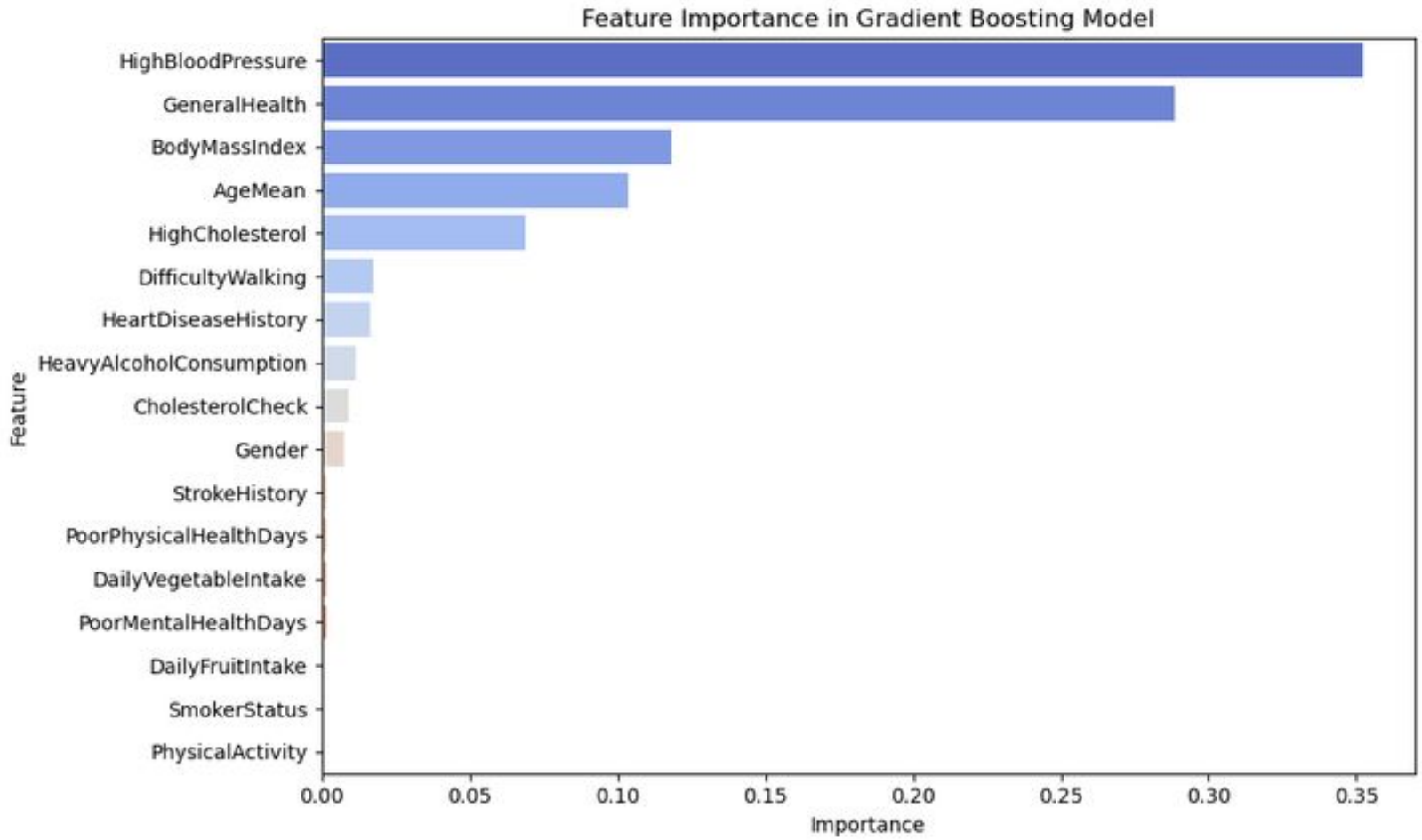
- Looking at the Learning Curve graph, we can see how our model learns as we give it more training data.
- The blue line is the training score, training score starts high and slightly decreases as we feed it more data.
- Cross validation score starts lower and rises slightly with more data
- **The small gap at the end shows that the model is not overfitting, and adding more data will not significantly improve model accuracy.**



Goal 2: See What Factors are most important

We were surprised by the results of Factor importance

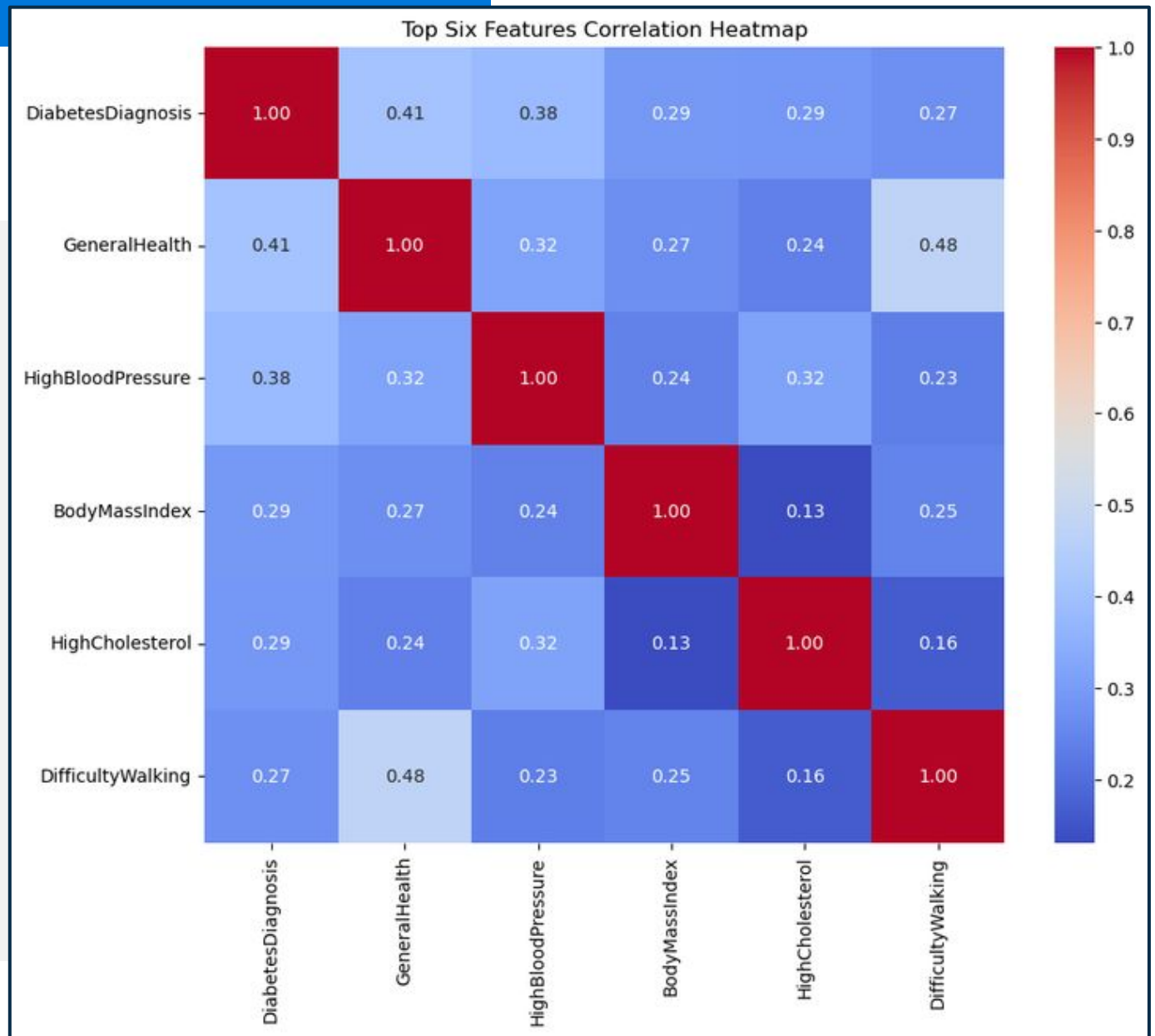
- High Blood Pressure, General Health, and Body Mass Index were most important
    - General Health was a self reported metric!
- Physical Activity, Fruit/Veggie Intake, and Smoker Status were all under 1% importance!



Goal: Who is most susceptible

A few connected data points:

- **Diabetes and General Health (+0.41)**
  - People who report a lower general health are more likely to have diabetes
- **Diabetes and High Blood Pressure (+0.38)**
  - Moderate positive correlation with High Blood Pressure, this aligns with medical findings that high blood pressure is a comorbidity in diabetic patients
- **Diabetes, BMI and High Cholesterol (+0.29)**
  - Weaker than the two above, but both high BMI and Cholesterol have a positive correlation with Diabetes.



Summary

➔ First:

- The biggest factors for predicting diabetes are in our control!
- In our analysis **4 of the top 5 features are directly related to the health choices that we make everyday**. Relative to other chronic conditions, Diabetes is possible to avoid.

➔ Second

- Medical conditions are hard to predict with Machine Learning alone!
- Despite our best efforts and trying a few different routes, getting past the 75% accuracy, and 75% precision rate was quite difficult. A wide variety of factors impact medical conditions.

Problems Encountered

The biggest problem we encountered was model selection.

Initially we went with a random_forest, because it had the highest training data score. However when working with it to tune the hyperparameters and improve it, we found that models accuracy was hard to increase.

We switched to Gradient Boosting because of its:

- **High Performance:** it often achieves high accuracy
- **Flexibility:** Can be used in both regression and classification
- **Handle Complex Data:** Capable of capturing complex patterns and relationships

Drawbacks to Gradient Boosting:

- **Computationally Intensive:** can be slower to train compared to simpler models
- **Overfitting:** If not properly tuned, can overfit the training data.

Future Considerations

1) Create a frontend where users could enter their own data in and see what result they would get.

2) Find data with more medical rigor, this dataset was self-reported

3) Create an ensemble machine-learning model with SciKit Learn for higher accuracy and potentially lower false positives