Predicting Diabetes with Machine Learning

APRIL 8, 2023

Brad Anderson
Bryson Webb
Alexandria Orvis

PROJECT 4

Can diabetes be predicted with machine learning using survey questions?

THE DATA SOURCE

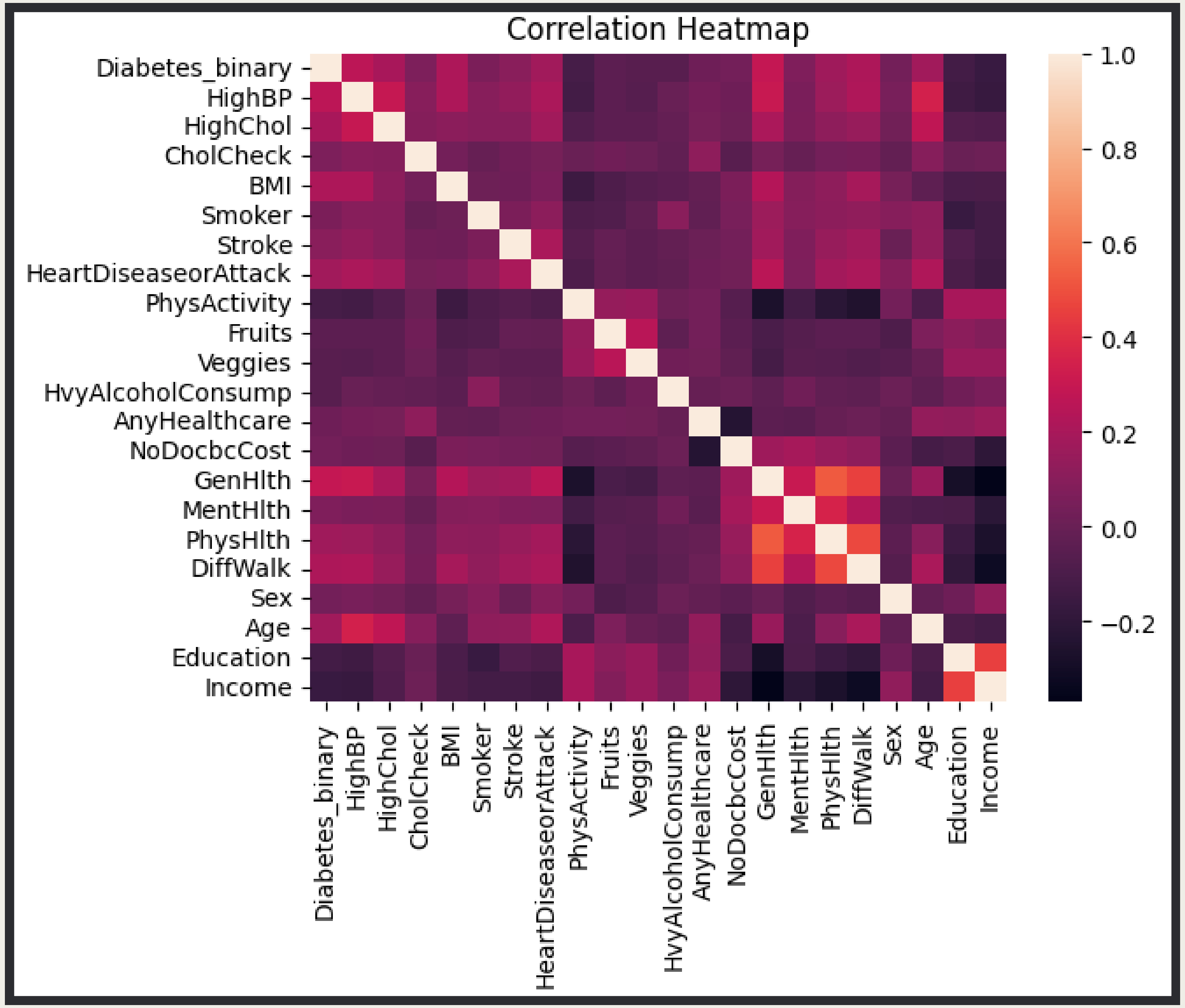
The data originated from the 2015 Behavior Risk Factor Surveillance System (BRFSS) survey.

The dataset was cleaned and binned for machine learning prior to its use for this project.

PROJECT ROADMAP

- Identified dataset
- Loaded dataset into PostgreSQL
- Performed exploratory data analysis
- Executed classification machine learning models
  - Logistic Regression
  - K Nearest Neighbors (KNN)
  - Random Forest
- Model optimization
  - Applied scaling on only categorical features
  - Feature engineering
  - Converted categorical features to dummy variables
  - Dropped less important features
- Final model evaluation

EXPLORATORY DATA ANALYSIS



Dataset Details

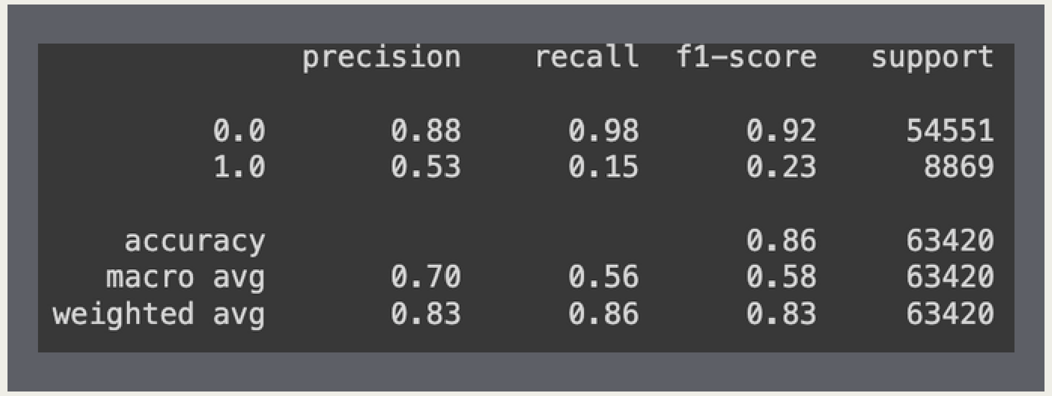
- 253,680 rows of data
  - 86% did not have Diabetes
  - 14% did have Diabetes
- Clean dataset with no missing values
- Categories have already been applied to the columns

Key Takeaways

- Very clean dataset with little transformation needed
- On average people with diabetes are/have:
  - Higher BMI
  - High Cholesterol
  - High Blood Pressure
  - Older
  - Lower Education
  - Lower Income
- Very low correlation between Diabetes and features
- Features with the highest correlation were Physical Health and General Health (r-value .524)
- Generated a box plot and found the BMI column had some outliers that could be removed

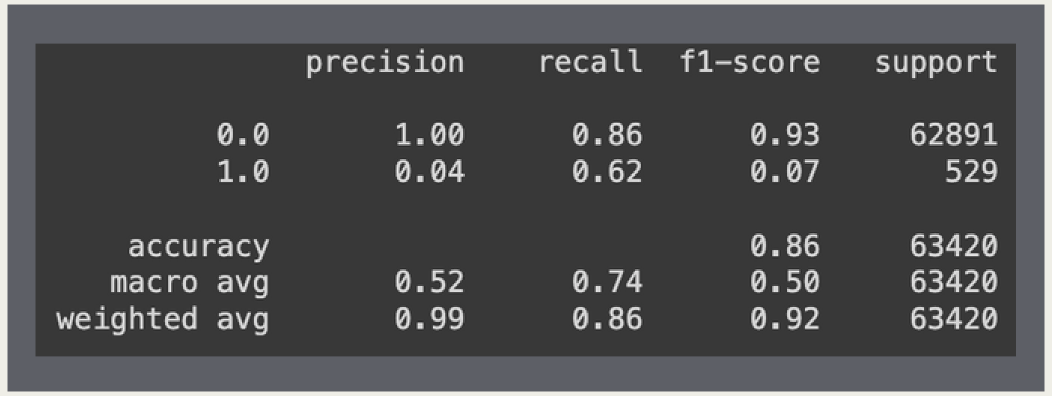
MACHINE LEARNING MODELS EMPLOYED

Logistic Regression



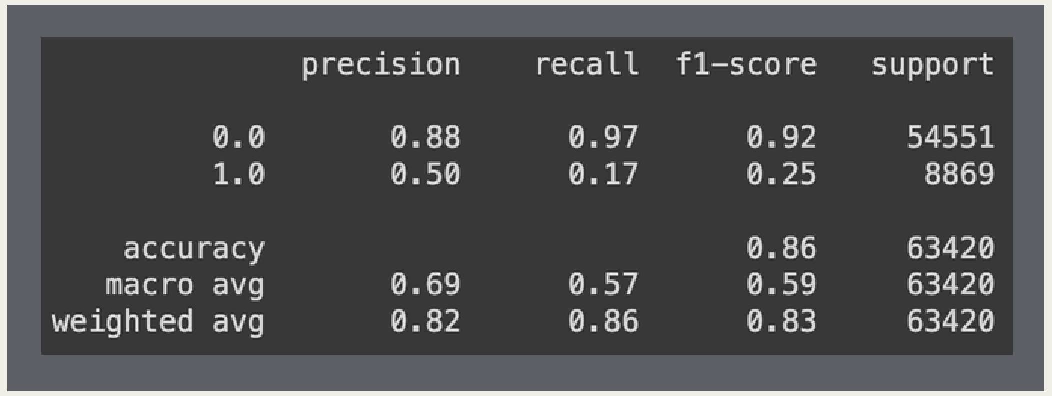
```
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.92     54551
         1.0       0.53      0.15      0.23      8869

    accuracy                           0.86     63420
   macro avg       0.70      0.56      0.58     63420
weighted avg       0.83      0.86      0.83     63420
```

KNN



```
              precision    recall  f1-score   support

         0.0       1.00      0.86      0.93     62891
         1.0       0.04      0.62      0.07       529

    accuracy                           0.86     63420
   macro avg       0.52      0.74      0.50     63420
weighted avg       0.99      0.86      0.92     63420
```

Random Forest

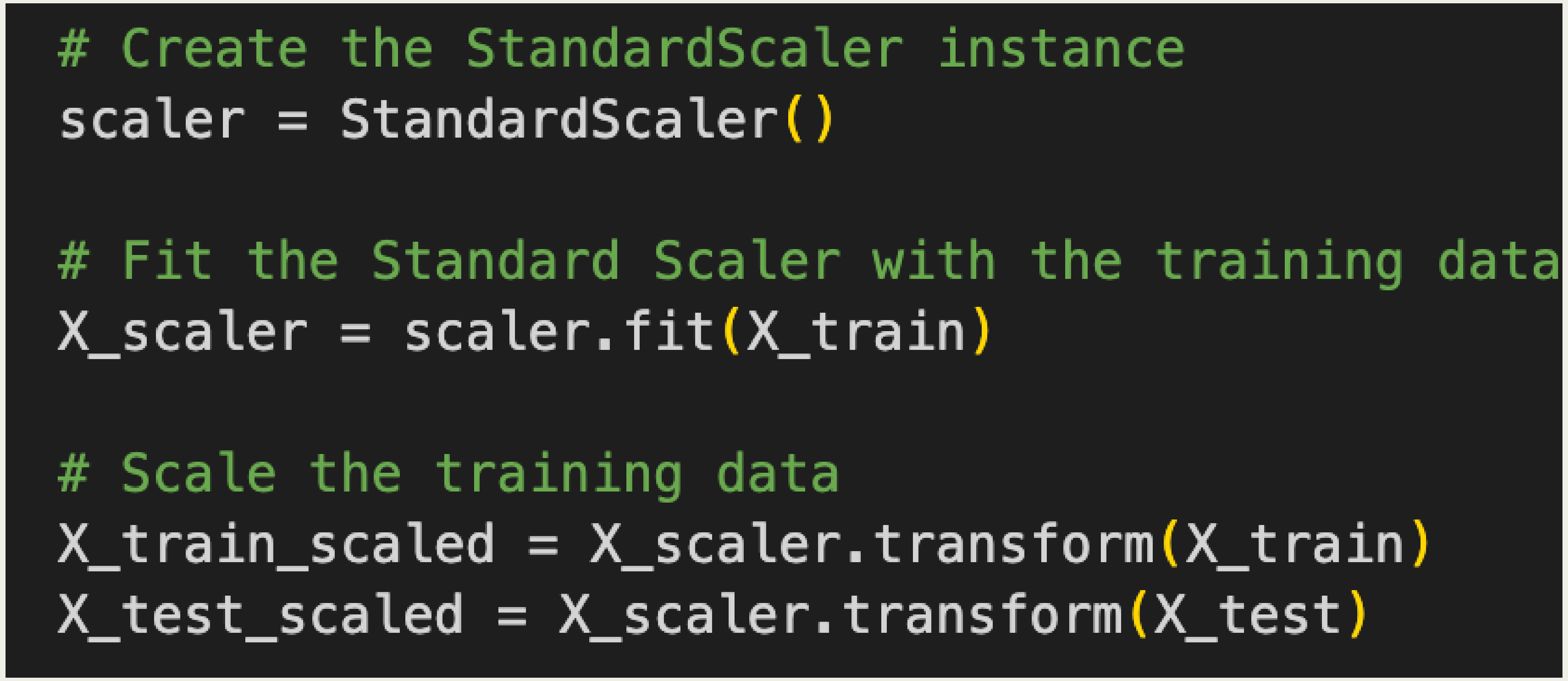


```
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92     54551
         1.0       0.50      0.17      0.25      8869

    accuracy                           0.86     63420
   macro avg       0.69      0.57      0.59     63420
weighted avg       0.82      0.86      0.83     63420
```

FEATURE SCALING

The group attempted to optimize the model's accuracy by scaling in two different ways.

- All of the features were processed with the **sklearn** Standard Scaler function.
- ONLY continuous variables (BMI, PhysHlth, MentHlth) were processed with the sklearn Standard Scaler function



```
# Create the StandardScaler instance
scaler = StandardScaler()

# Fit the Standard Scaler with the training data
X_scaler = scaler.fit(X_train)

# Scale the training data
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```

FEATURE ENGINEERING

We created a number of features using the five most important features (based on the Random Forest Model importance function) to attempt to increase each models effectiveness.

Age / BMI

MentHlth * BMI

Age * BMI

Income * BMI

PhysHlth * MentHlth

CONVERTING CATAGORICALS

A majority of the variables were encoded as 0's and 1's relating to Yes or No survey questions.

Other answers were encoded on different scales (0-5 for example) but were still categorical. Any categorical variable that was not binary was converted to a dummy variable to attempt to improve the accuracy of the various models.

OPTIMIZED LOGISTIC REGRESSION

Accuracy

0.86

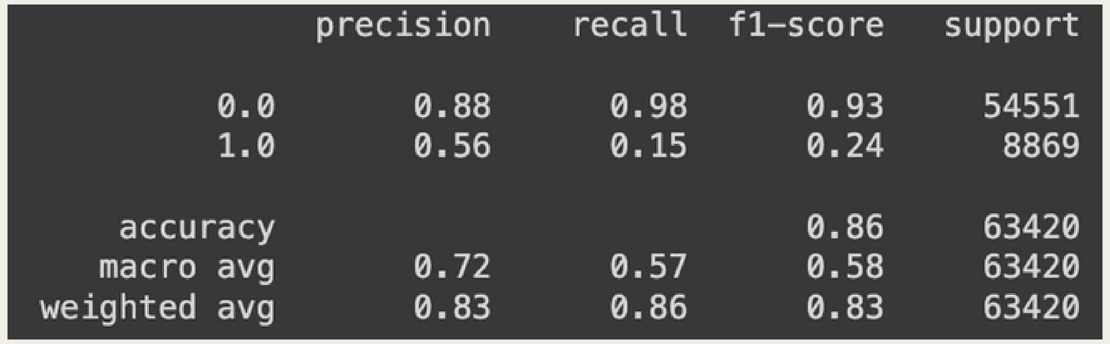
Precision

0 : 0.88 *1 : 0.53*

Recall

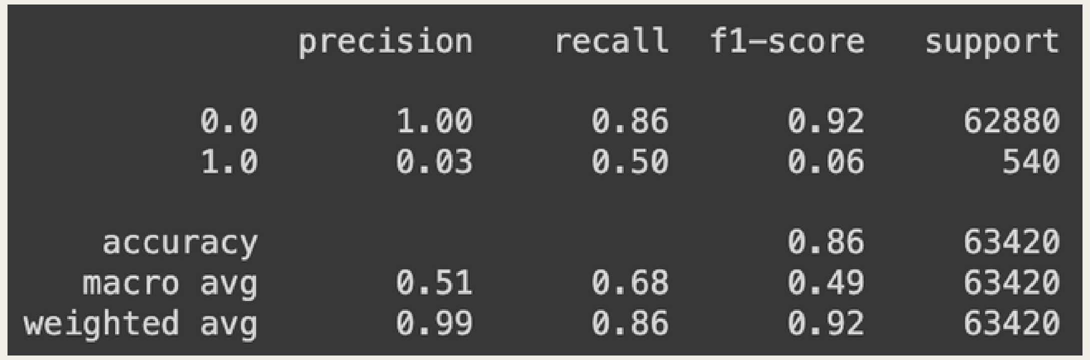
0 : 0.56 *1 : 0.15*

LOGISTIC REGRESSION



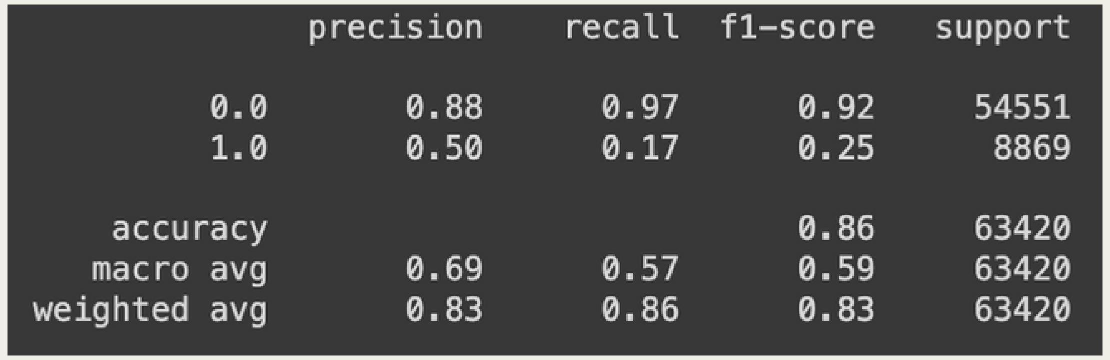
```
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     54551
         1.0       0.56      0.15      0.24      8869

    accuracy                           0.86     63420
   macro avg       0.72      0.57      0.58     63420
weighted avg       0.83      0.86      0.83     63420
```

KNN



```
              precision    recall  f1-score   support

         0.0       1.00      0.86      0.92     62880
         1.0       0.03      0.50      0.06       540

    accuracy                           0.86     63420
   macro avg       0.51      0.68      0.49     63420
weighted avg       0.99      0.86      0.92     63420
```

RANDOM FOREST

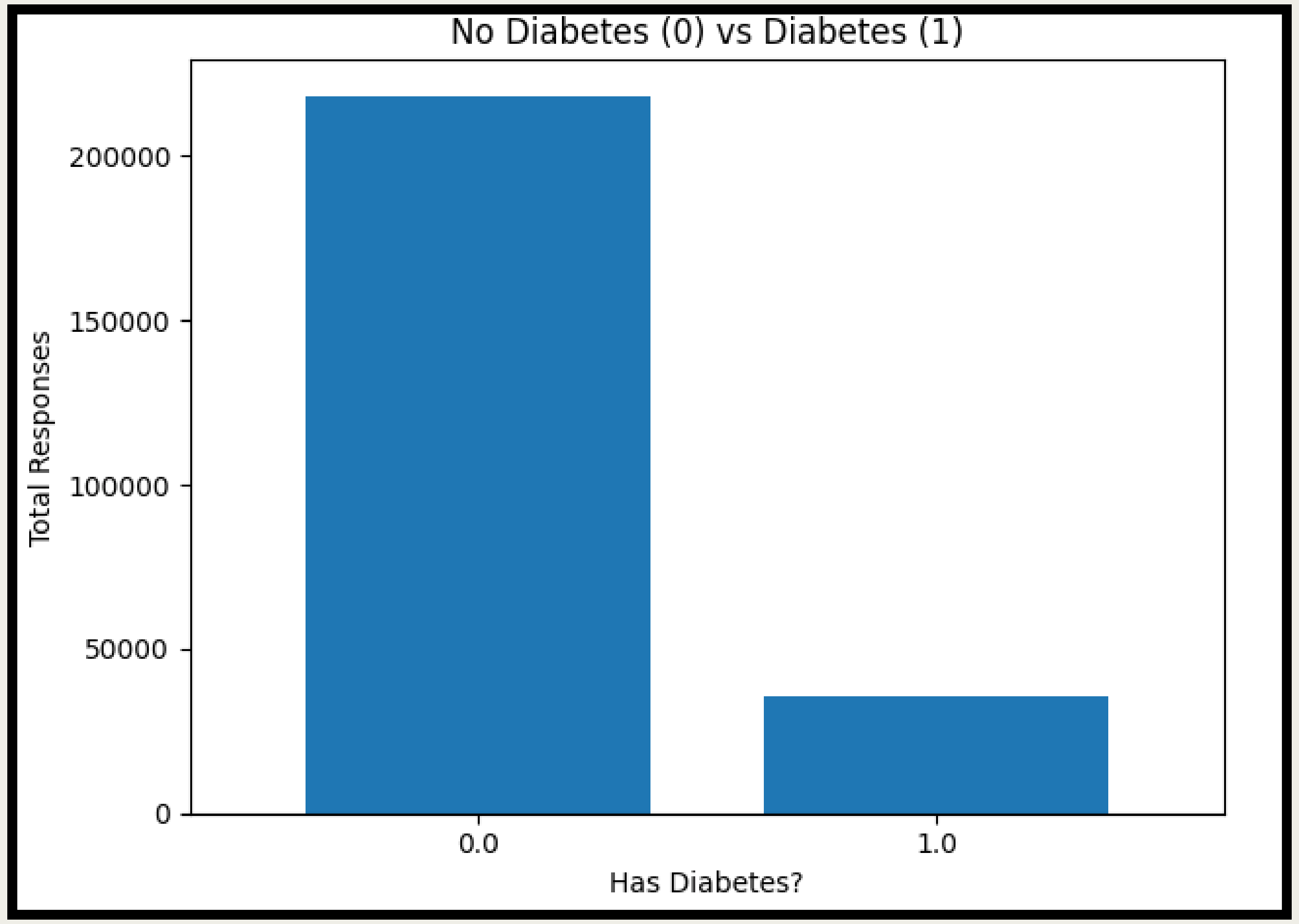


```
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92     54551
         1.0       0.50      0.17      0.25      8869

    accuracy                           0.86     63420
   macro avg       0.69      0.57      0.59     63420
weighted avg       0.83      0.86      0.83     63420
```

DISCUSSION :

IMBALANCED DATA

The original data set was significantly imbalanced. The majority of the survey respondents (86%) did NOT have diabetes. This makes it challenging to train a model to identify positive diabetes cases as it has so few to draw from.



MISLEADING ACCURACY

Though our original accuracy for all models was reasonably high (86%) the data failed to capture MOST of the positive diabetes cases. The model almost always predicted a negative result. Given 86% of the population did not have diabetes, the negative prediction was correct 86% of the time.

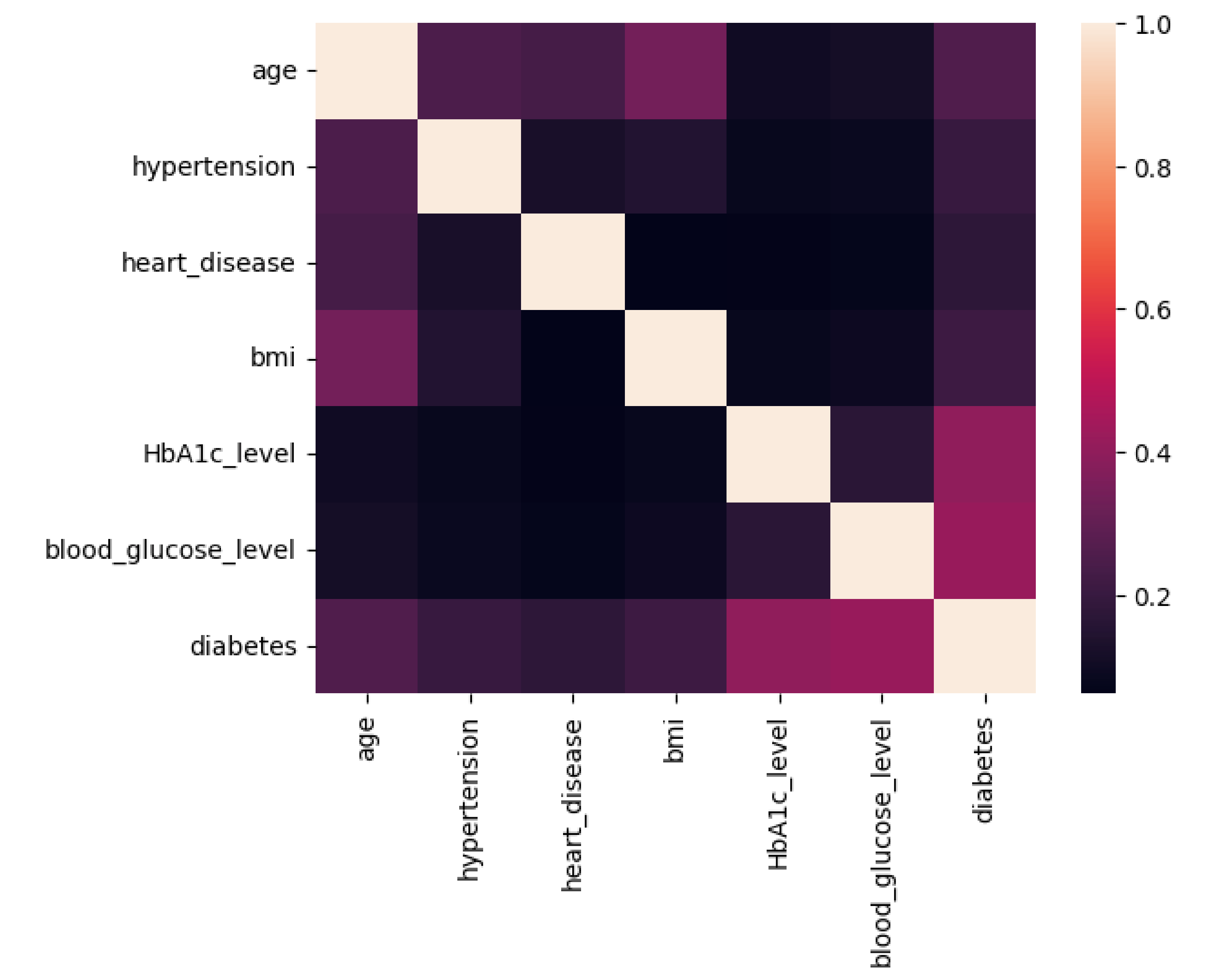
CONTINUOUS V. CATAGORICAL

The data was processed prior to being used for this project. Most of the features were encoded into categorical variables. To account for this we transformed the models in various was to attempt to make them more useful for the model to interpret

DIABETES DATASET WITH IN HOSPITAL TESTING

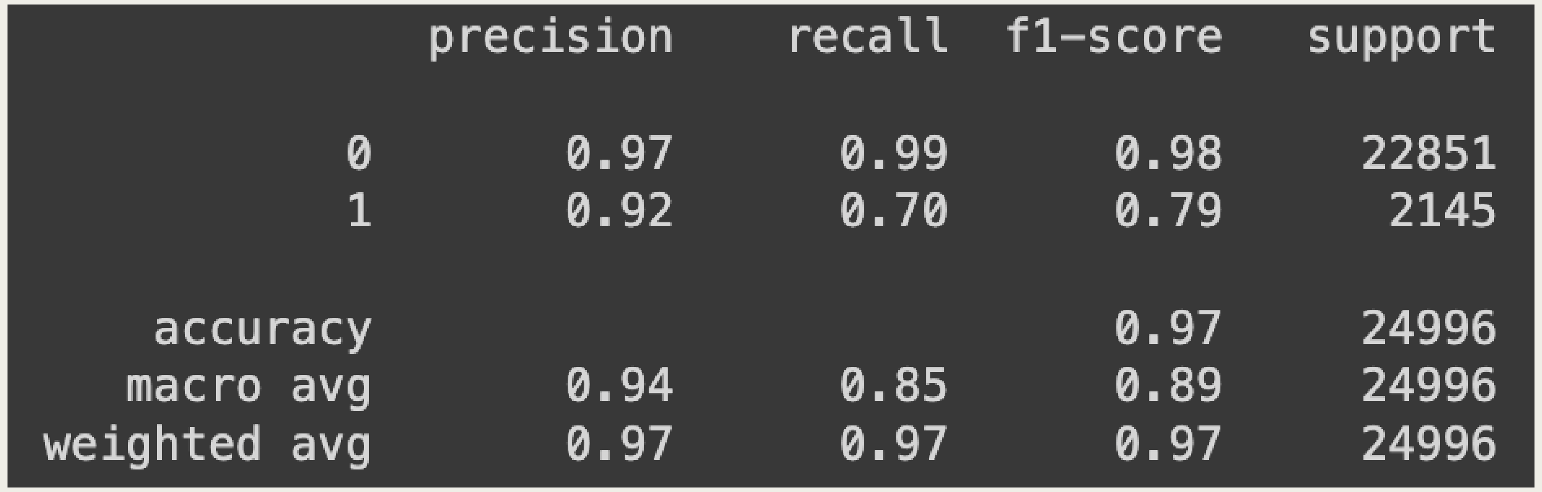
DIABETES PREDICTION DATASET

- We wanted to compare the results of our survey model against the results of a model that used actual medical data taken from patients
- Diabetes Prediction is a second dataset we found that included medical and demographic data about patients
  - Dataset was taken from Kaggle
- Dataset was also severely imbalanced
  - 92% did not have diabetes
  - 8% did have diabetes
  - Total of 100,000 rows
- Dataset included data on Hemoglobin A1C (HbA1c) and Glucose, which are the two main ways to test/track diabetes



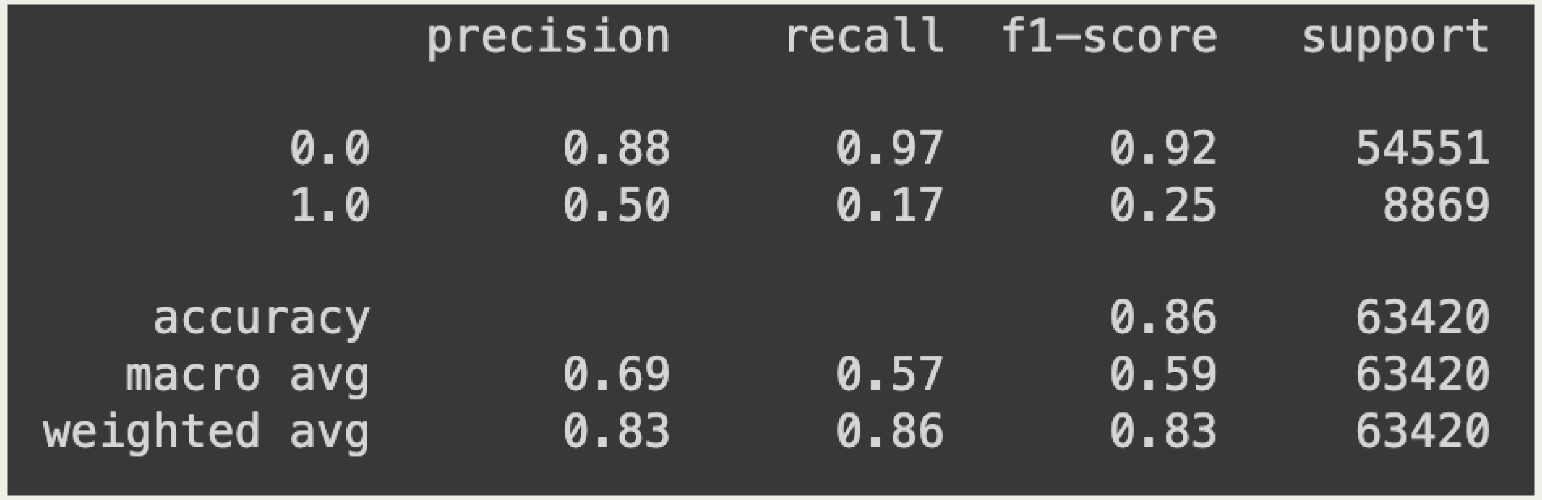
PREDICTING DIABETES WITH MEDICAL TESTING : A COMPARISON

Medical Test RF Outcomes



```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98     22851
           1       0.92      0.70      0.79      2145

    accuracy                           0.97     24996
   macro avg       0.94      0.85      0.89     24996
weighted avg       0.97      0.97      0.97     24996
```

Survey Questions RF Outcomes



```
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92     54551
         1.0       0.50      0.17      0.25      8869

    accuracy                           0.86     63420
   macro avg       0.69      0.57      0.59     63420
weighted avg       0.83      0.86      0.83     63420
```

Thank you!

RESOURCES

- **Diabetes Health Indicators Dataset (Main)**
    - https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data
- **Diabetes Prediction Dataset (secondary)**
    - https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
- **Seaborn Library Code**
    - https://www.almabetter.com/bytes/tutorials/data-science/exploratory-data-analysis