

heathlikethecandybar / phase_5

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

☆ 0 stars

🔗 0 forks

👁 1 watching

🏠 Activity

🌐 Public repository

main

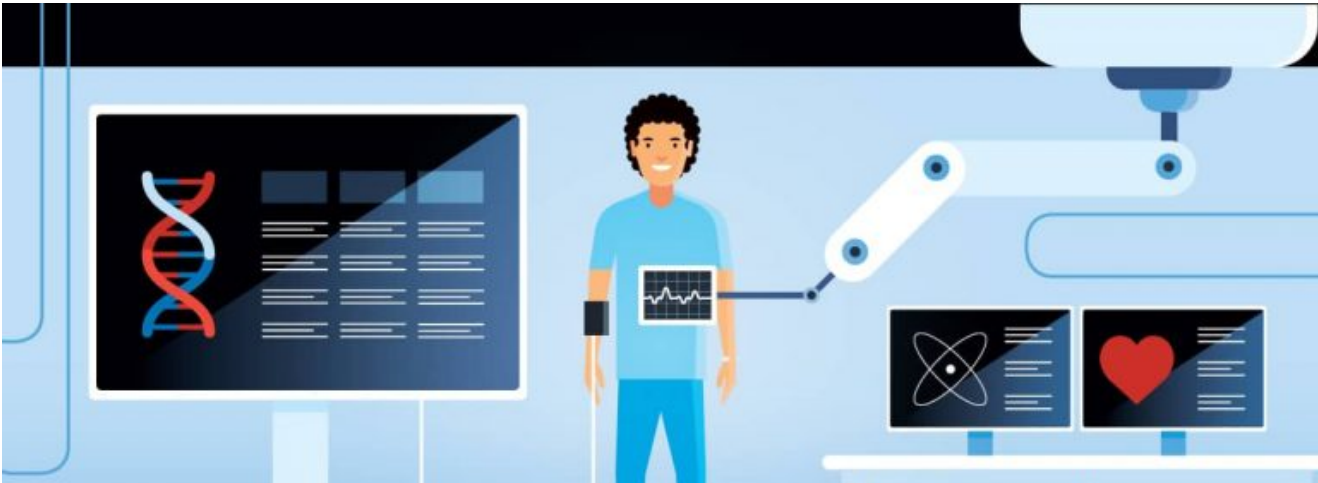
Branches

Tags


Heath Rittler and Heath Rittler Updates from presentation and feedback ...

2 hours ago 20

View code



☰ README.md



Diabetes Classification

Author: [Heath Rittler](#)

Overview

According to the International Diabetes Federation, in 2021, an estimated 537 million people worldwide had diabetes, and this number is projected to rise to 642 million by 2040. This is roughly 1 in 15 people and around 7% of the total population living with Diabetes. In the United States, the American Diabetes Association reports that the total estimated cost of diagnosed diabetes was \$327 billion in 2017, including direct medical costs and reduced productivity.

Individuals with diabetes tend to have higher healthcare expenses compared to those without the condition. Diabetes also incurs indirect costs, such as lost productivity and disability. These costs arise from missed workdays, reduced productivity at work, early retirement, and disability due to diabetes-related complications.

Business Problem

An insurance company wants to develop a predictive model to assess the risk of diabetes among their policyholders based on a limited set of available data points. By accurately identifying individuals at high risk of developing diabetes, the company aims to take proactive measures to reduce healthcare costs and improve the overall health outcomes of their customers.

The challenge for the company is to build a robust and accurate predictive model that can handle the complexity and non-linear relationships between the available data points and the risk of developing diabetes. The model should consider factors such as age, gender, BMI, hypertension status, heart disease history, smoking history, HbA1c level, and blood glucose level. We will use a classification model to predict diabetes within the population of interest.

My background and work history has been in healthcare which makes this an interesting problem for me. Being able to accurately predict risk within a population and provide resources and preventive measures are important now more than ever.

Data

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information.

This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

A link to the data can be found on Kaggle located here:

<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

The data are contained in a single file:

diabetes_prediction_dataset.csv

- `gender` - Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories in it male ,female and other.
- `age` - Age is an important factor as diabetes is more commonly diagnosed in older adults.Age ranges from 0-80 in our dataset.
- `hypertension` - Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values a 0 or 1 where 0 indicates they don't have hypertension and for 1 it means they have hypertension.
- `heart_disease` - Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values a 0 or 1 where 0 indicates they don't have heart disease and for 1 it means they have heart disease.
- `smoking_history` - Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes.In our dataset we have 5 categories i.e not current,former,No Info,current,never and ever.
- `bmi` - BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.
- `HbA1c_level` - HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes.
- `blood_glucose_level` - Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.
- `diabetes` - Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

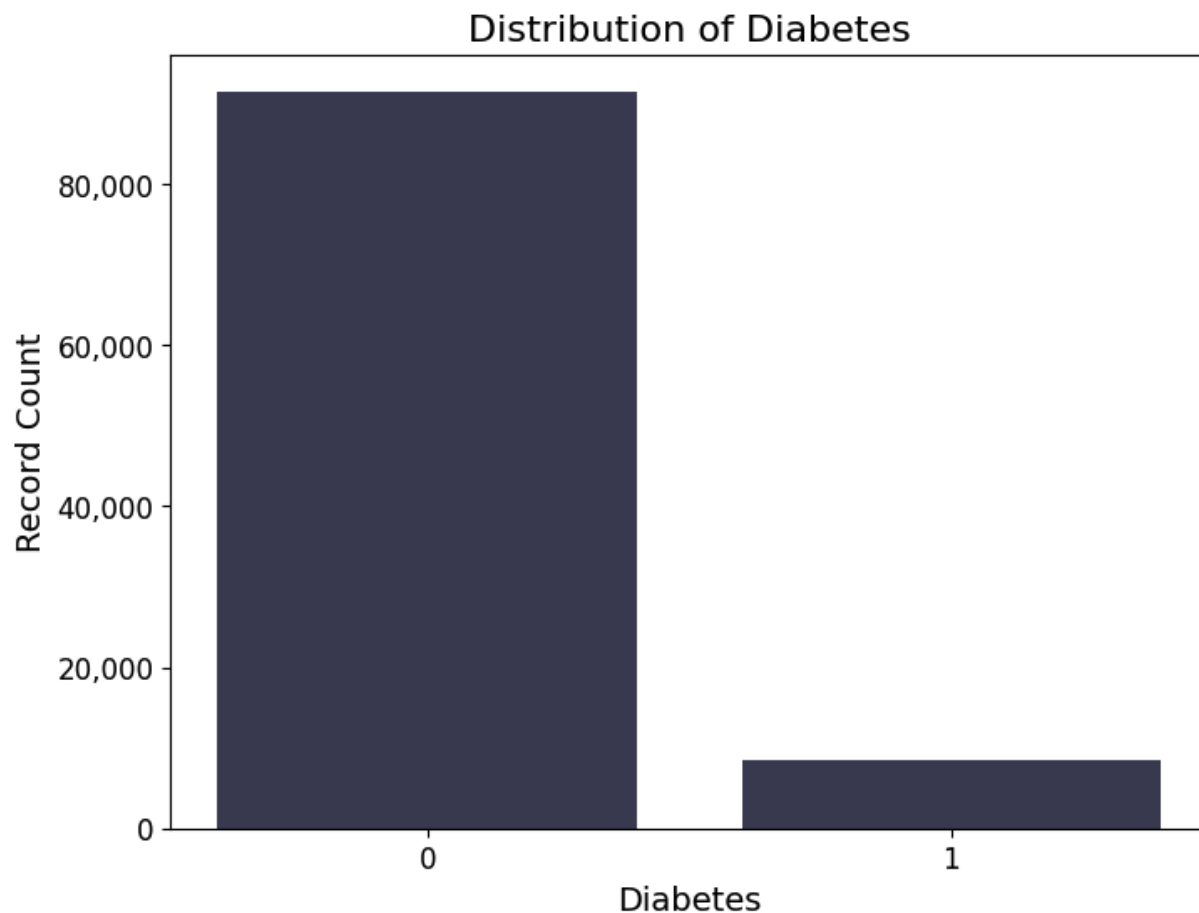
Additional information about this dataset can be found on the [Kaggle](#) website.

Methods

After our exploratory analysis, we employed classification methodologies to see if we could accurately predict a diabetes diagnosis within our data. We trained our model on 80% of the dataset, while saving the remaining 20% to test our assumptions in what our algorithms learned. We leveraged a Logistic Regression model, and tuned the regression's hyperparameters to arrive at our baseline model.

That model was then iterated on, leveraging multiple models such as Decision Tree, Random Forest, XGBoost, and MLP Neural Network to evaluate the best model. Each approach was modeled with and without tuned hyperparameters.

Our target class or variable was 'diabetes' column. This included either a 1 or a 0, and was present approximately 8.5% of the records included.



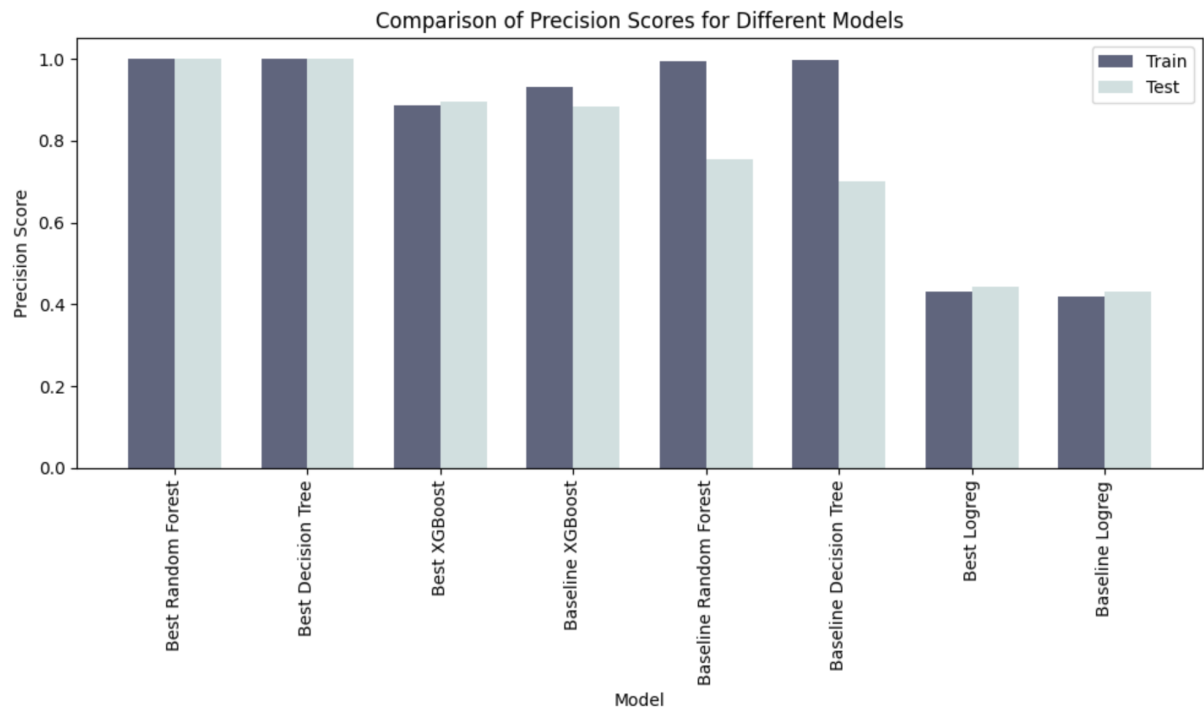
Because of this target imbalance we used SMOTE (Synthetic Minority Oversampling Technique) to generate new instance from existing minority cases that we used as the input. In addition to SMOTE, we also used Standard Scaling in our pipeline to ensure that no features were weighted because of their difference in scale value.

Overall, our best-performing model for our metrics of interest was the tuned XGBoost, achieving a precision score of 88% and 89% on the train and test sets respectively. It demonstrated excellent accuracy in identifying true positive cases of diabetes while minimizing false positives.

When considering the F1 score, which balances precision and recall, the best decision tree and random forest models showed the highest scores. These however were being influenced by the extremely high precision score, not taking into account the recall performance (as much). When looking at our XGBoost models, these models achieved F1 scores of around 80% on the test set, indicating a good balance between precision and recall. They were not the highest, but they also had performed better and more evenly across all metrics.

In summary, our models demonstrated strong performance in accurately classifying diabetes cases. The XGBoost, with its balanced performance across precision, F1 score, and ultimately recall show promising potential for accurately predicting diabetes in future applications.

For our final model, we were able to predict positive diabetes diagnoses 89% of the time.

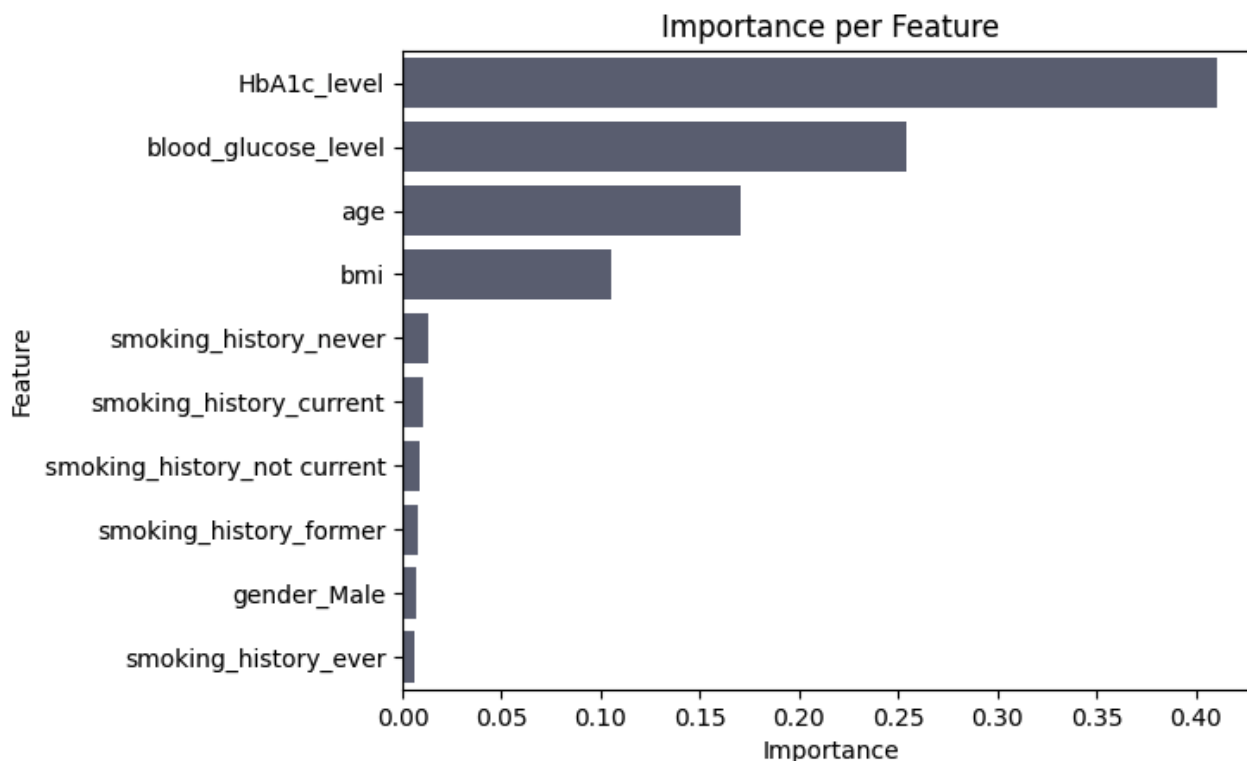


	Model	Train Precision Score	Test Precision Score	Train F1 Score	Test F1 Score
3	Best Decision Tree	1	1	0.801026	0.803713
5	Best Random Forest	1	1	0.801026	0.803713
7	Best XGBoost	0.884868	0.89547	0.784971	0.788142
6	Baseline XGBoost	0.93223	0.883058	0.836596	0.78499
4	Baseline Random Forest	0.994138	0.753302	0.994374	0.740895
2	Baseline Decision Tree	0.996498	0.699173	0.994441	0.715686
1	Best Logreg	0.430175	0.442249	0.558478	0.568822
0	Baseline Logreg	0.419432	0.432226	0.551109	0.561701

The top 3 features that influence diabetes diagnosis are:

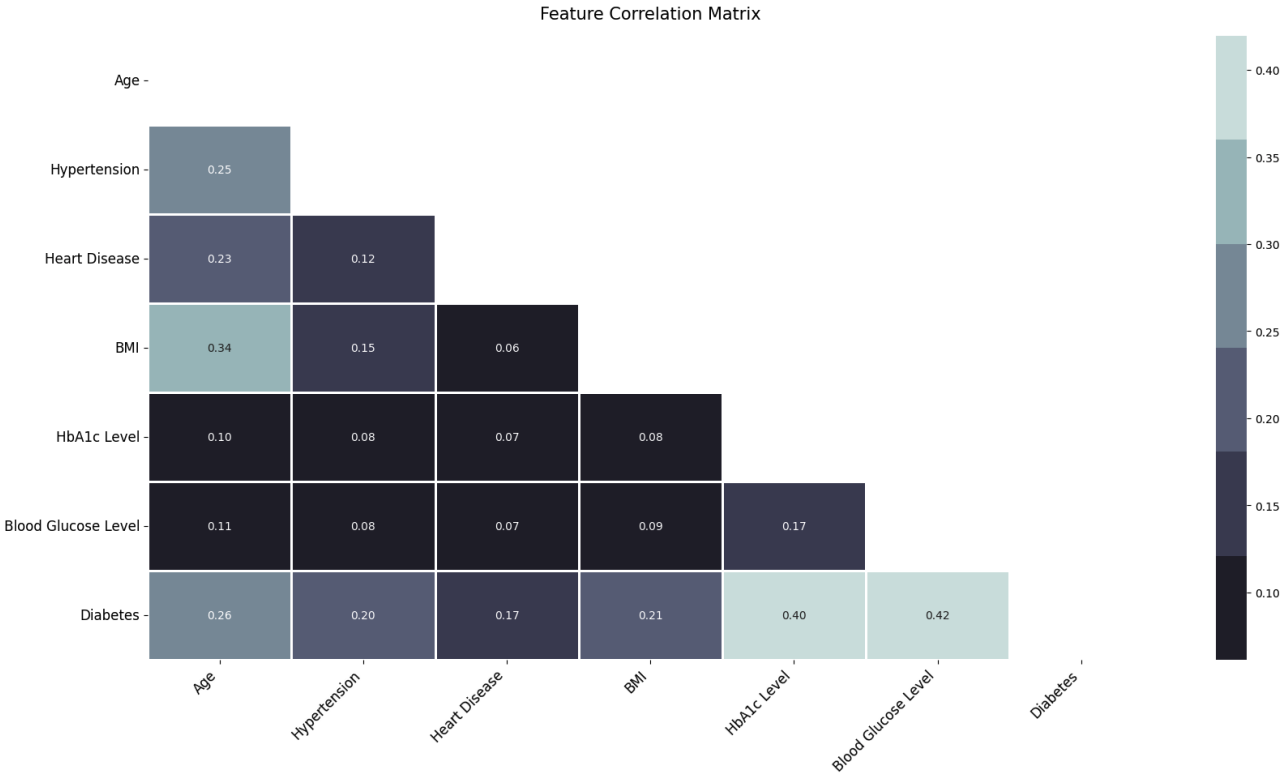
HbA1c Blood Glucose Level Age

These top features make a lot of sense as they are essential in definiing and measuring diabetes diagnosis. So patients that have high markers in these features are very highly likely to have diabetes. Age, and other comorbid factors such as smoking history, and Body Mass Index (BMI) are also high risk factors for acquiring the preventable disease. These latter factors are very tightly coupled with lifestyle choices, which is where our interventions can make a difference in reducing risk for the patient.



Results

As we were seeing in our EDA & in our feature importance from our modeling, there is a high correlation between HbA1c and blood glucose levels within diabetes cases. BMI is the 3rd highest correlated value, while age, hypertension and heart disease also show moderate correlations. We also see a high correlation between age and BMI which makes sense, and hypertension and heart disease. These are all risk factors associated with Diabetes, and with the exception of age, are lifestyle based decisions that can be impacted.



As we look to intervene with these lifestyle markers, it would be interesting to evaluate additional data such as physical activity, caloric intake, heart rate data (maybe other wearable information), to see if we could predict one of the markers within our dataset. Timing is really critical in disease progression, especially when considering interventions to prevent a diagnosis.

Conclusions & Recommendations

In our diabetes classification problem, we aimed to develop models that could accurately predict the presence of diabetes based on various features. We evaluated the performance of several models, including logistic regression, decision trees, random forest, XGBoost, and MLP neural network.

Overall, our best-performing model for our metrics of interest was the tuned XGBoost, achieving a precision score of 88% and 89% on the train and test sets respectively. It demonstrated excellent accuracy in identifying true positive cases of diabetes while minimizing false positives.

When considering the F1 score, which balances precision and recall, the best decision tree and random forest models showed the highest scores. These however were being influenced by the extremely high precision score, not taking into account the recall performance (as much). When looking at our XGBoost models, these models achieved F1 scores of around 80% on the test set, indicating a good balance between precision and recall. They were not the highest, but they also had performed better and more evenly across all metrics.

In summary, our models demonstrated strong performance in accurately classifying diabetes cases. The XGBoost, with its balanced performance across precision, F1 score, and ultimately recall show promising potential for accurately predicting diabetes in future applications.

Next steps:

- Run the algorithm on new data.
- Continually evolve the datasets that are being used for prediction to incorporate more features and potentially time based data.
- Analyze time, and the impact of additional metrics in #2 and early diagnosis.
- Evaluate interventions impact on classified population vs those that were not classified for programming.
- Load data into centralized repository for sharing into operational systems.

For More Information

The full analysis is located in the [Jupyter Notebook](#) or review this summary [presentation](#).

For additional info, contact Heath Rittler at hrittler@gmail.com

Image courtesy of <https://pharmaceuticalintelligence.com/2021/05/29/developing-machine-learning-models-for-prediction-of-onset-of-type-2-diabetes/>

Repository Structure

```
|— data
|— images
|— README.md
|— phase_5_presentation.pdf
|— phase_5_notebook.pdf
|— phase_5_notebook.ipynb
```

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 100.0%