dragunat2016 /
Diabetes_Prediction

Code Issues Pull requests Actions Projects Wiki Security Insights Set

Study on predicting diabetes from features in a dataset.

☆ **0** stars **0** forks **1** watching Activity

Public repository

main ▾

Branches Tags

dragunat2016 re-ran notebook made minor presentation changes ··· 3 minutes ago **13**

View code

README.md

Table of Contents

- Overview
- Repository Links
- Repository Structure
- Reproduction Steps
- Business Objectives
- Data Overview
- Exploratory Data Analysis
- Modeling
- Final Model Evaluation
- Recommendations
- Future Projects
- Contact Information

Overview

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

The BRFSS started collecting data in 2014. That year they had collected a significant amount of data across more than 400,000 people relating to health, vaccinations, and chronic conditions. Researchers saw an opportunity to apply machine learning algorithms to make predictions on the data, since it was a feature rich dataset with hundreds-of-thousands of records.

Repository Links

- Main Notebook
- Images
- notebooks
- Variables
- Neural Network Model
- presentation

Repository Structure

```
├── index.ipynb
├── Data_Cleaning-colab.ipynb
├── README.md
├── presentation.pdf
├── notebook.pdf
├── github.pdf
├── .gitignore
├── images
├── Variables
├── notebooks
```

```
|   ├── Data_Cleaning.ipynb
|   └── Neural_Network_Modeling.ipynb
```

Reproduction Steps

Download from Github to Local Machine

1. Download the 2015.CSV from this link: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system
2. Save CSV to file and run steps from the data cleaning notebook.
3. Run the main notebook.

Running on Google Colab

If you can run multiple notebooks on same runtime

1. Run the data cleaning colab notebook. first (Data_Cleaning-Colab).
2. Assuming, you have the kaggle API key, you should have downloaded the CSV to your colab space and generated the files.
3. Run the index notebook

If you cannot run multiple notebooks on the same runtime

1. Download github repo to google drive
2. Mount your google drive too colab.
3. Open the data_cleaning-colab notebook.
4. i. Run the data cleaning colab notebook. first (Data_Cleaning-Colab).
5. Run the index file

Business Objectives

BRFSS™

We have been tasked by the CDC to create models from previous BRFSS data that predicts diabetes. The CDC wants to help the people it surveys and alert them if they are at risk for diabetes given their survey results. Long-term the CDC would like to publish an application to Americans allowing them to fill out a form with questions on their vitals like BMI and habits such as excercise. Upon completing the form, the CDC would send back a diabetic risk to the person.

Accuracy and precision are our primary metrics of evaluation. Optimizing on these two metrics should reduce the amount of false positives we encounter. We want to avoid false positives because they could result in unnecessary outreach and wasting resources.

We will also be incorporating the "run time" of the model in our evaluation. Run time is the amount of time it takes to train and test the model.

A final model evaluation will be made by some heuristic combination of the accuracy, precision, and time it takes the model too run. Any gains in accuracy and precision need to justify the time it takes to train and use the model.

Data Overview

The 2015 data is available on this link from the CDC's website. The table with all the responses and the key donoting the data terms are also available. The link to the survey questions is here

The page on the CDC's website containing the data is here.

The data on the CDC's page is in an ASCII format and hard too decode with time constraints. We found a CSV version of that data on Kaggle. The download link for the CSV is specifically here.

Full Link: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system

Limitations

This is survey data where the user responses were segmented into several categories.

So the following limitations apply:

- Survey respondants may not be comfortable revealing sensitive information over the phone even if the response is anonymous.
- Many respondants who answer "no" for diabetes may actually have diabetes, but were not diagnosed. Note: That there was a significant imbalance of diabetes/pre-diabetes versus those who stated that they do not have the condition.
- Many variables that are continuous in nature were treated as ordinal in the study such as income and age. These variables were treated as ordinal as part of the models.

Data Preparation

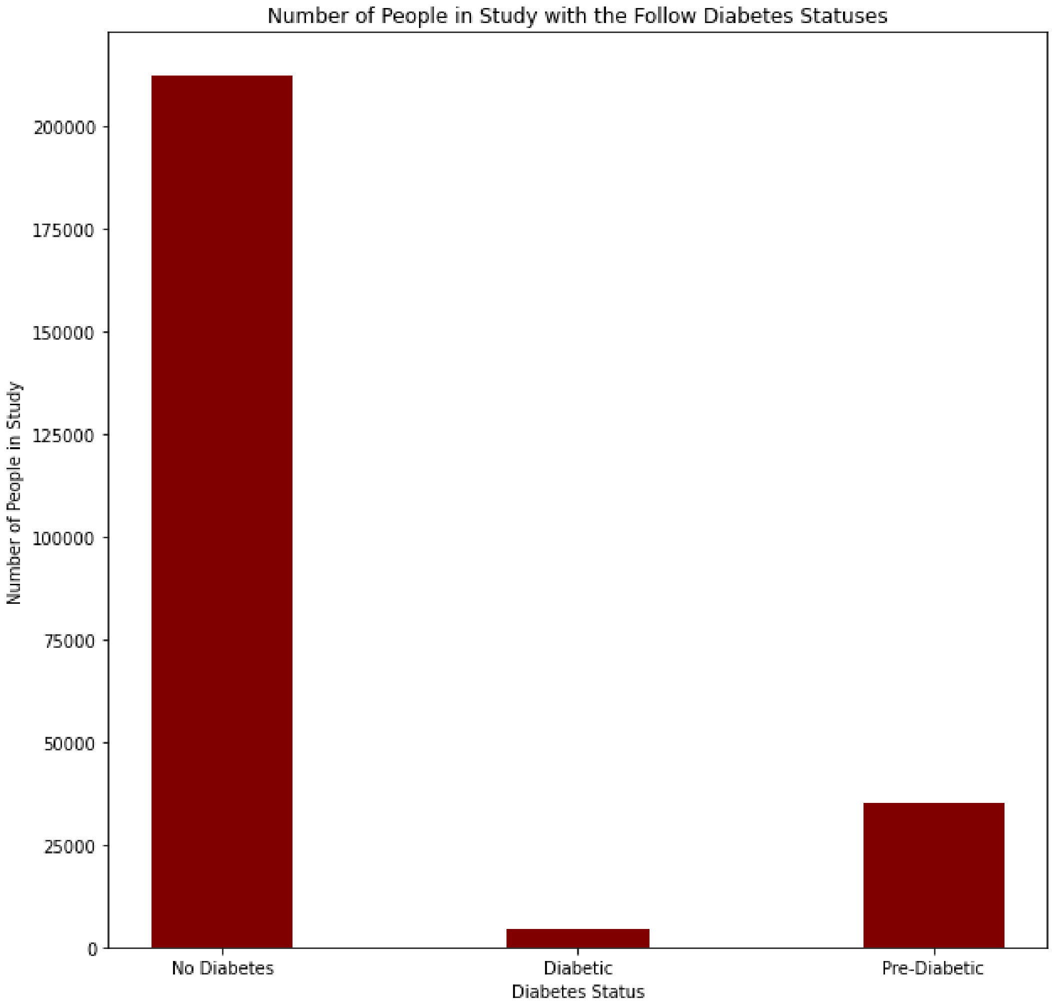
The steps for data preparation and cleaning were done in this notebook for the sake of simplifying the main notebook.

This is the short version of the data cleaning process. For more detail please click the link above.

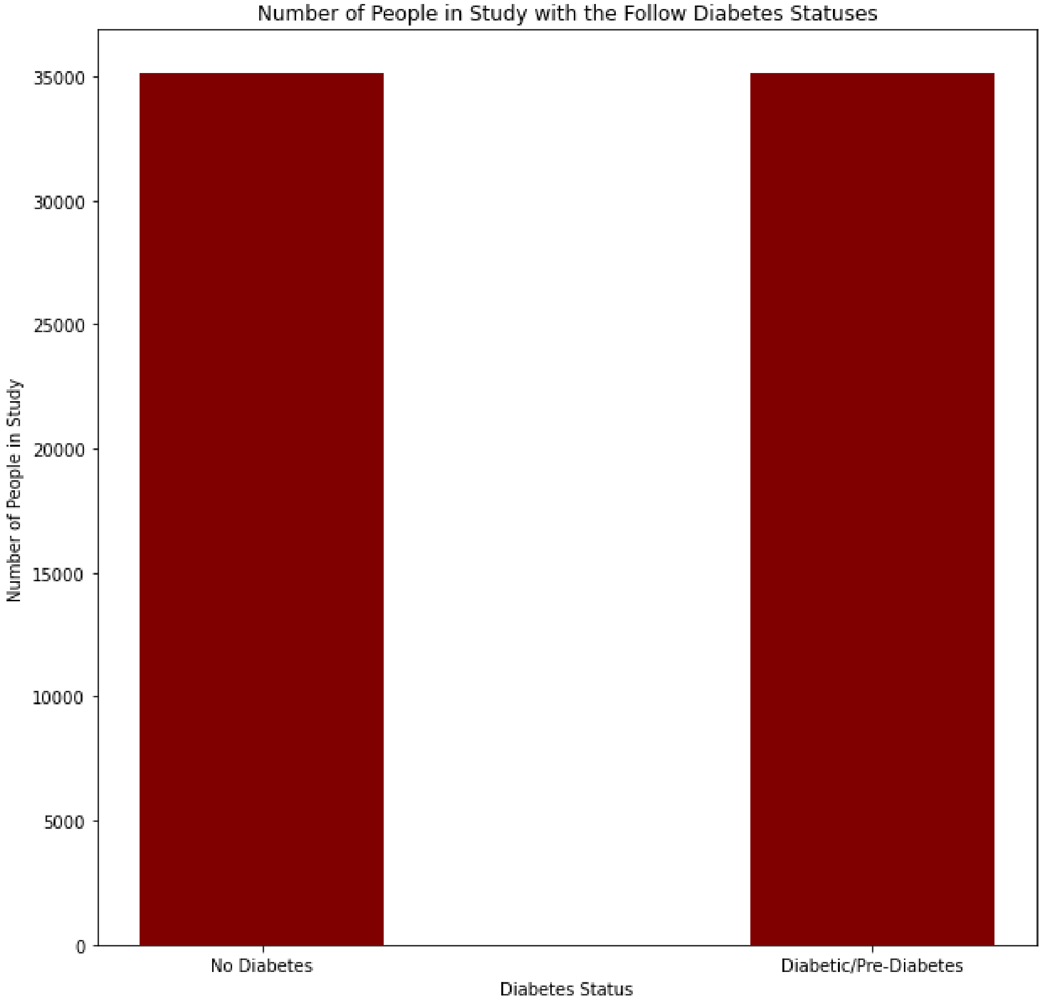
High - Level Process

- Selected for columns related to diabetes
- Dropped columns with significant data missing
- Reviewed the data in the features.
- Values within features that corresponded to information like 'N/A', 'Refused', 'Didn't Know' were dropped.
- Values were transformed to be more ordinal

- Combined Diabetes and Prediabetes data
- Addressed class imbalance by making the diabetes/non-diabetes records 50-50

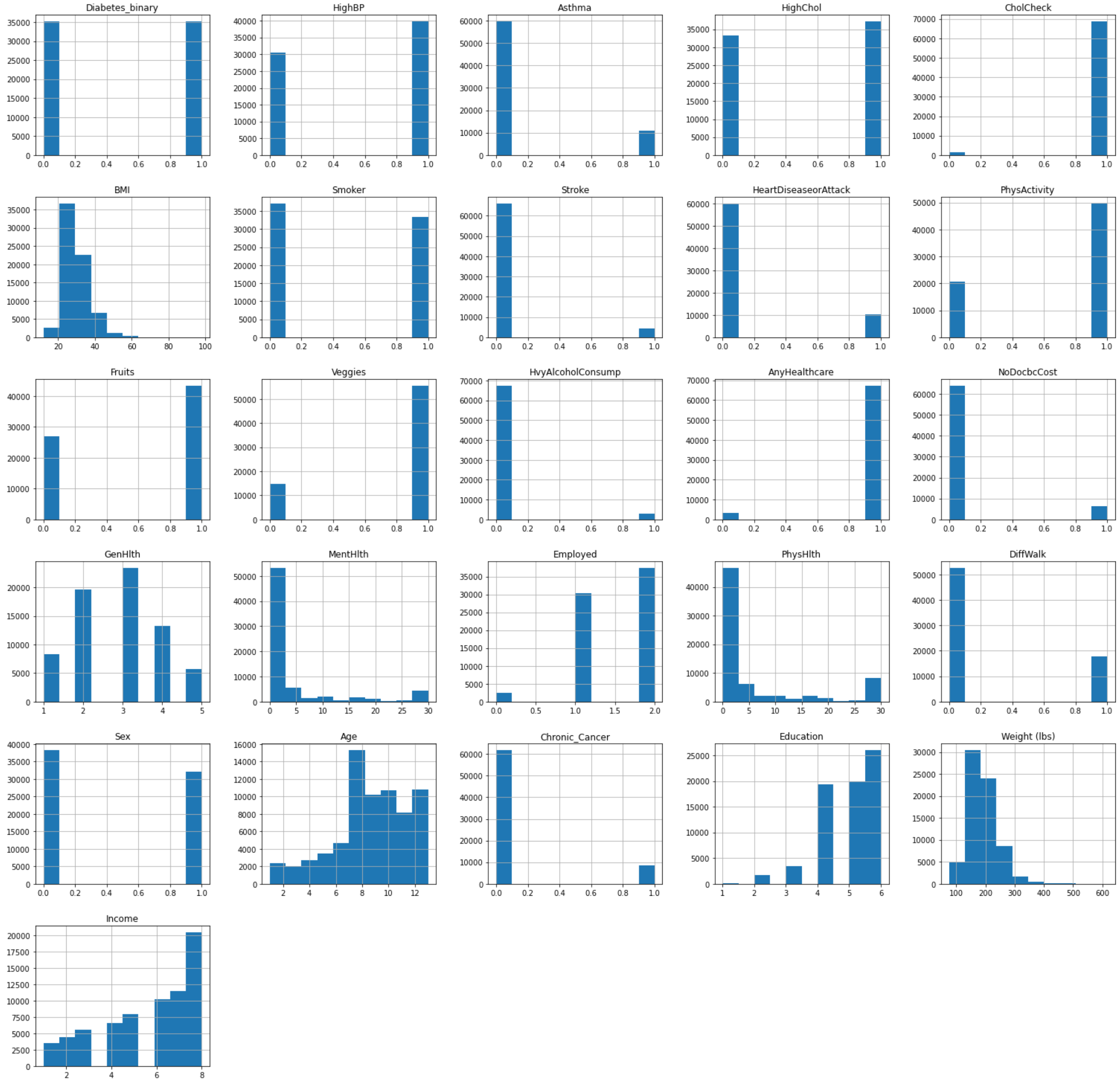


Records after cleaning. Clear class imbalance

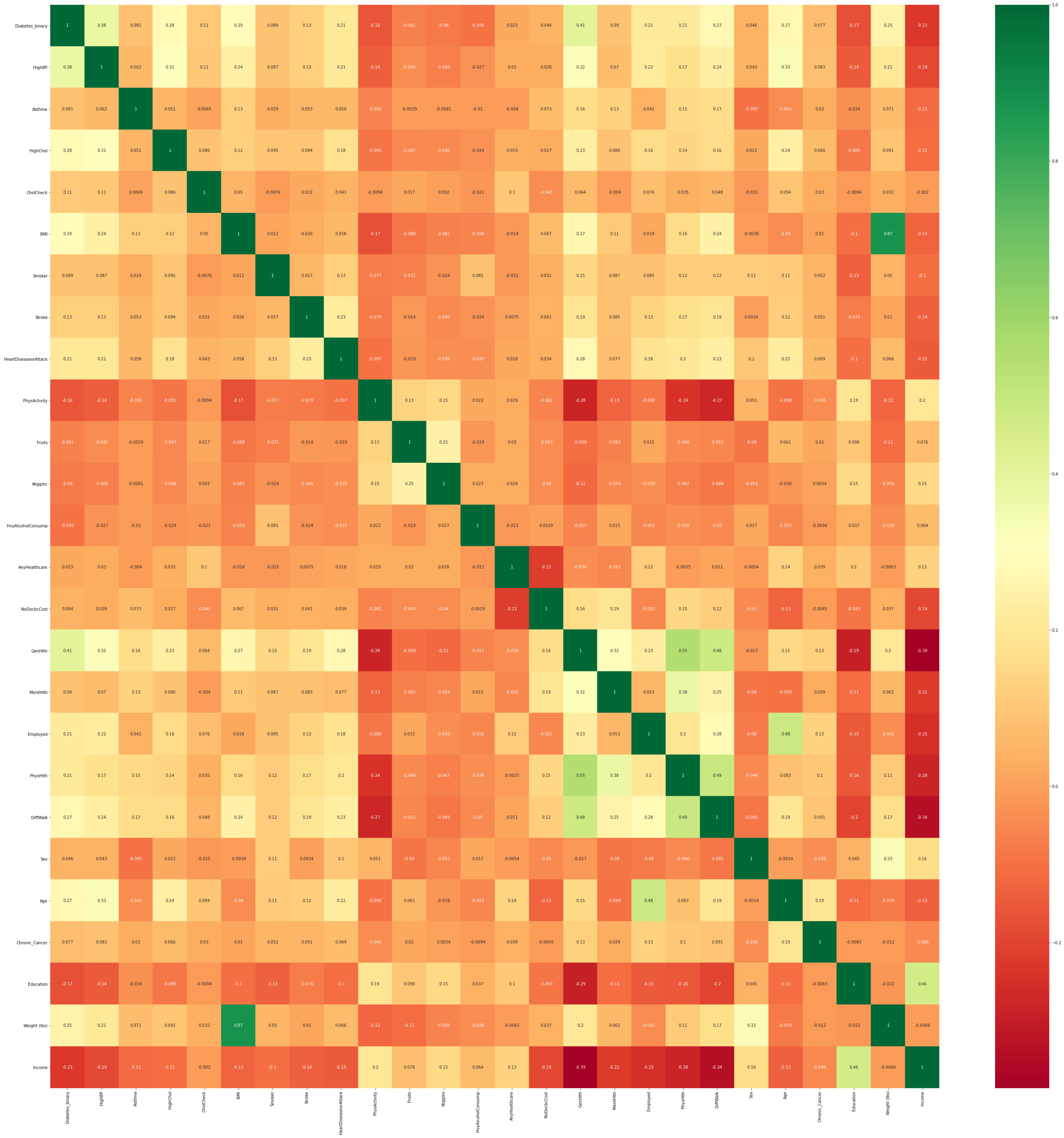


Rectified imbalance by randomly selecting non-diabetic records to match the size of diabetic records.

Exploratory Data Analysis



Here is a histogram showing the features in the data set.



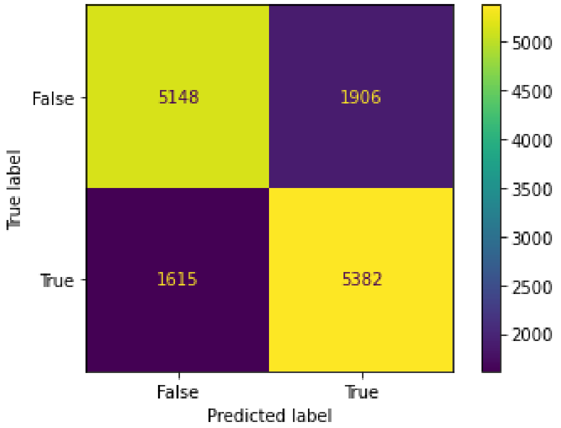
	Diabetes_binary	HighBP	Asthma	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Employed	PhysHlth	DiffWalk	Sex	Age	Chronic_Cancer	Education	Weight (lbs)	Income
Diabetes_binary	1	0.38	0.081	0.28	0.11	0.29	0.089	0.13	0.21	-0.16	-0.061	-0.08	-0.095	0.023	0.044	0.41	0.09	0.21	0.21	0.27	0.046	0.27	0.077	-0.17	0.25	-0.23
HighBP	0.38	1	0.062	0.31	0.11	0.24	0.087	0.13	0.21	-0.14	-0.045	-0.069	-0.027	0.03	0.028	0.32	0.07	0.22	0.17	0.24	0.043	0.33	0.083	-0.14	0.21	-0.19
Asthma	0.081	0.062	1	0.051	0.0069	0.13	0.029	0.053	0.058	-0.056	-0.0029	-0.0081	-0.01	-0.004	0.073	0.16	0.13	0.041	0.15	0.17	-0.095	-0.042	0.03	-0.034	0.071	-0.11
HighChol	0.28	0.31	0.051	1	0.086	0.12	0.095	0.094	0.18	-0.091	-0.047	-0.046	-0.024	0.033	0.027	0.23	0.086	0.16	0.14	0.16	0.022	0.24	0.066	-0.085	0.091	-0.11
CholCheck	0.11	0.11	0.0069	0.086	1	0.05	-0.0076	0.022	0.043	-0.0094	0.017	0.002	-0.021	0.1	-0.043	0.064	-0.004	0.076	0.035	0.048	-0.015	0.094	0.03	-0.0094	0.032	-0.002
BMI	0.29	0.24	0.13	0.12	0.05	1	0.012	0.026	0.058	-0.17	-0.088	-0.062	-0.059	-0.014	0.067	0.27	0.11	0.018	0.16	0.24	-0.0038	-0.04	0.01	-0.1	0.87	-0.13
Smoker	0.089	0.087	0.029	0.095	-0.0076	0.012	1	0.057	0.13	-0.077	-0.071	-0.024	0.081	-0.015	0.031	0.15	0.087	0.085	0.12	0.12	0.11	0.11	0.052	-0.13	0.05	-0.1
Stroke	0.13	0.13	0.053	0.094	0.022	0.026	0.057	1	0.23	-0.079	-0.014	-0.046	-0.024	0.0075	0.041	0.19	0.085	0.13	0.17	0.19	0.0034	0.12	0.051	-0.076	0.01	-0.14
HeartDiseaseorAttack	0.21	0.21	0.058	0.18	0.043	0.058	0.13	0.23	1	-0.097	-0.019	-0.039	-0.037	0.018	0.034	0.28	0.077	0.18	0.2	0.23	0.1	0.22	0.069	-0.1	0.066	-0.15
PhysActivity	-0.16	-0.14	-0.056	-0.091	-0.0094	-0.17	-0.077	-0.079	-0.097	1	0.13	0.15	0.022	0.029	-0.061	-0.28	-0.13	-0.088	-0.24	-0.27	0.051	-0.098	-0.045	0.19	-0.12	0.2
Fruits	-0.061	-0.045	-0.0029	-0.047	0.017	-0.088	-0.071	-0.014	-0.019	0.13	1	0.25	-0.019	0.03	-0.043	-0.099	-0.062	0.015	-0.048	-0.051	-0.09	0.061	0.02	0.096	-0.11	0.076
Veggies	-0.08	-0.069	-0.0081	-0.046	0.002	-0.062	-0.024	-0.046	-0.039	0.15	0.25	1	0.027	0.028	-0.04	-0.12	-0.054	-0.039	-0.067	-0.084	-0.053	-0.018	0.0054	0.15	-0.056	0.15
HvyAlcoholConsump	-0.095	-0.027	-0.01	-0.024	-0.021	-0.059	0.081	-0.024	-0.037	0.022	-0.019	0.027	1	-0.013	0.0029	-0.057	0.015	-0.052	-0.039	-0.05	0.017	-0.053	-0.0094	0.037	-0.038	0.064
AnyHealthcare	0.023	0.03	-0.004	0.033	0.1	-0.014	-0.015	0.0075	0.018	0.029	0.03	0.028	-0.013	1	-0.21	-0.034	-0.051	0.12	-0.0025	0.011	-0.0054	0.14	0.039	0.1	-0.0063	0.13
NoDocbcCost	0.044	0.028	0.073	0.027	-0.043	0.067	0.031	0.041	0.034	-0.061	-0.043	-0.04	0.0029	-0.21	1	0.16	0.19	-0.055	0.15	0.12	-0.05	-0.13	-0.0045	-0.093	0.037	-0.19
GenHlth	0.41	0.32	0.16	0.23	0.064	0.27	0.15	0.19	0.28	-0.28	-0.099	-0.12	-0.057	-0.034	0.16	1	0.32	0.23	0.55	0.48	-0.017	0.15	0.13	-0.29	0.2	-0.39
MentHlth	0.09	0.07	0.13	0.086	-0.004	0.11	0.087	0.085	0.077	-0.13	-0.062	-0.054	0.015	-0.051	0.19	0.32	1	0.053	0.38	0.25	-0.09	-0.099	0.029	-0.11	0.062	-0.22
Employed	0.21	0.22	0.041	0.16	0.076	0.018	0.085	0.13	0.18	-0.088	0.015	-0.039	-0.052	0.12	-0.055	0.23	0.053	1	0.2	0.28	-0.08	0.48	0.13	-0.15	-0.042	-0.25
PhysHlth	0.21	0.17	0.15	0.14	0.035	0.16	0.12	0.17	0.2	-0.24	-0.048	-0.067	-0.039	-0.0025	0.15	0.55	0.38	0.2	1	0.49	-0.046	0.083	0.1	-0.16	0.11	-0.28
DiffWalk	0.27	0.24	0.17	0.16	0.048	0.24	0.12	0.19	0.23	-0.27	-0.051	-0.084	-0.05	0.011	0.12	0.48	0.25	0.28	0.49	1	-0.085	0.19	0.091	-0.2	0.17	-0.34
Sex	0.046	0.043	-0.095	0.022	-0.015	-0.0038	0.11	0.0034	0.1	0.051	-0.09	-0.053	0.017	-0.0054	-0.05	-0.017	-0.09	-0.08	-0.046	-0.085	1	-0.0014	-0.038	0.045	0.33	0.16
Age	0.27	0.33	-0.042	0.24	0.094	-0.04	0.11	0.12	0.22	-0.098	0.061	-0.018	-0.053	0.14	-0.13	0.15	-0.099	0.48	0.083	0.19	-0.0014	1	0.19	-0.11	-0.079	-0.13
Chronic_Cancer	0.077	0.083	0.03	0.066	0.03	0.01	0.052	0.051	0.069	-0.045	0.02	0.0054	-0.0094	0.039	-0.0045	0.13	0.029	0.13	0.1	0.091	-0.038	0.19	1	-0.0083	-0.012	-0.046
Education	-0.17	-0.14	-0.034	-0.085	-0.0094	-0.1	-0.13	-0.076	-0.1	0.19	0.096	0.15	0.037	0.1	-0.093	-0.29	-0.11	-0.15	-0.16	-0.2	0.045	-0.11	-0.0083	1	-0.022	0.46
Weight (lbs)	0.25	0.21	0.071	0.091	0.032	0.87	0.05	0.01	0.066	-0.12	-0.11	-0.056	-0.038	-0.0063	0.037	0.2	0.062	-0.042	0.11	0.17	0.33	-0.079	-0.012	-0.022	1	-0.0066
Income	-0.23	-0.19	-0.11	-0.11	-0.002	-0.13	-0.1	-0.14	-0.15	0.2	0.076	0.15	0.064	0.13	-0.19	-0.39	-0.22	-0.25	-0.28	-0.34	0.16	-0.13	-0.046	0.46	-0.0066	1

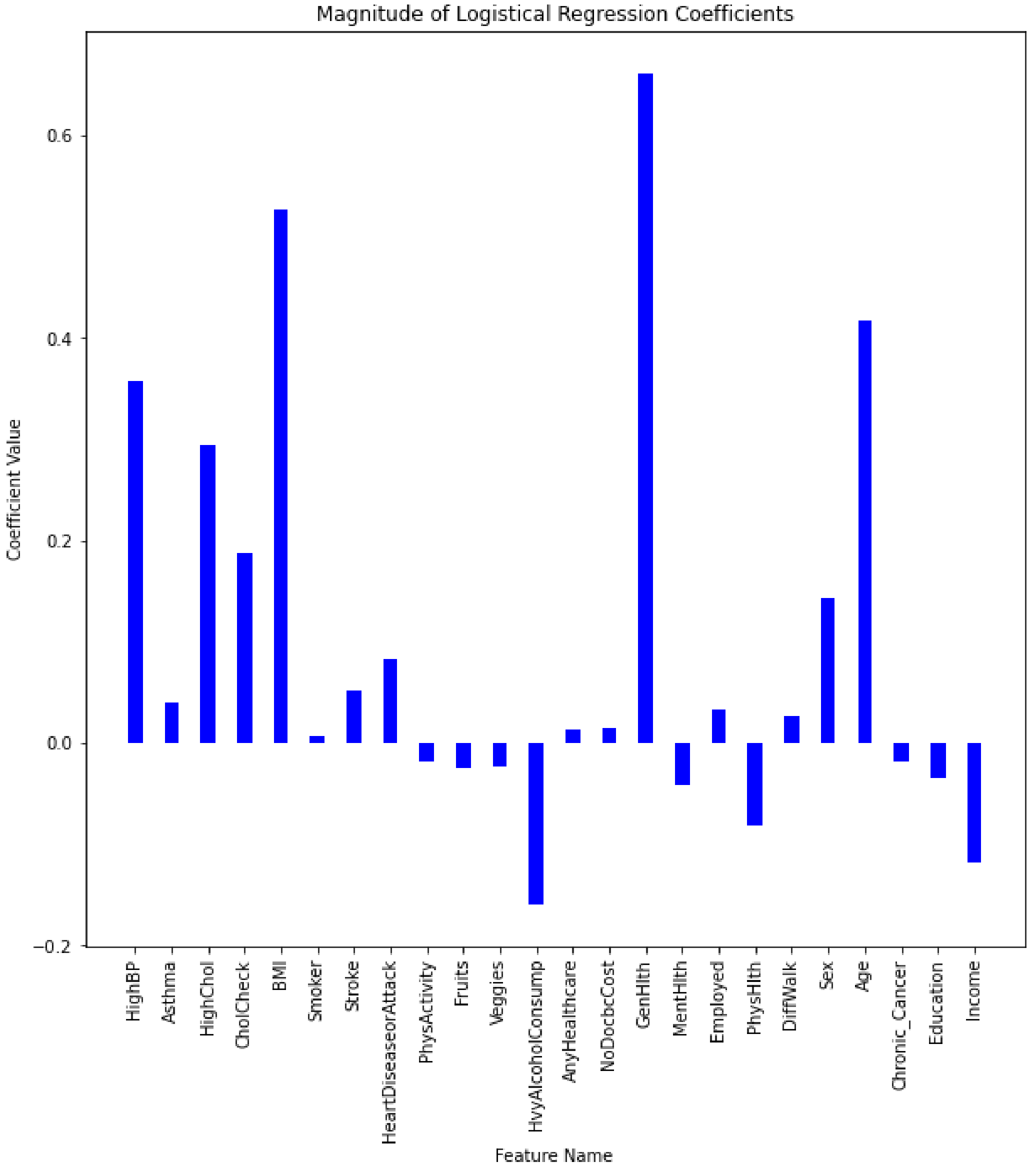
Multicolinearity was investigate by creating a heatmap

Modeling

Baseline Model

Started with a baseline logistical regression model. There are the results via a confusion matrix and it's feature importance.

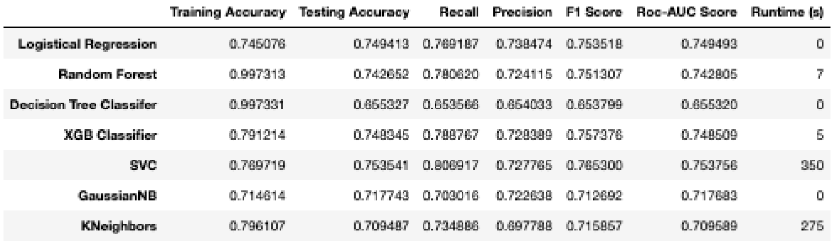




Additional Models

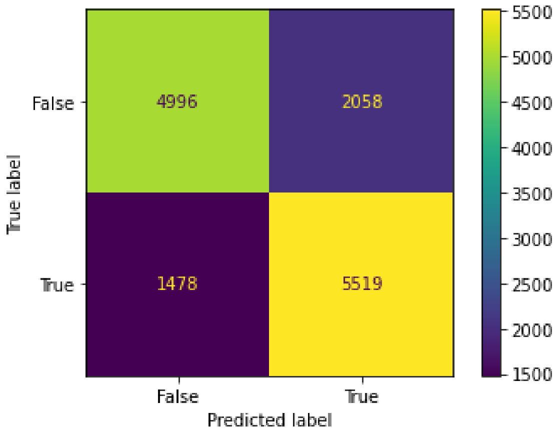
Next we ran a bunch off models. We chose XGBoost off the models to parameter tune since it had the second highest accuracy, and ran 80 times faster than the most accurate model Support Vector Machines.

Here is a table of the metrics off those models.

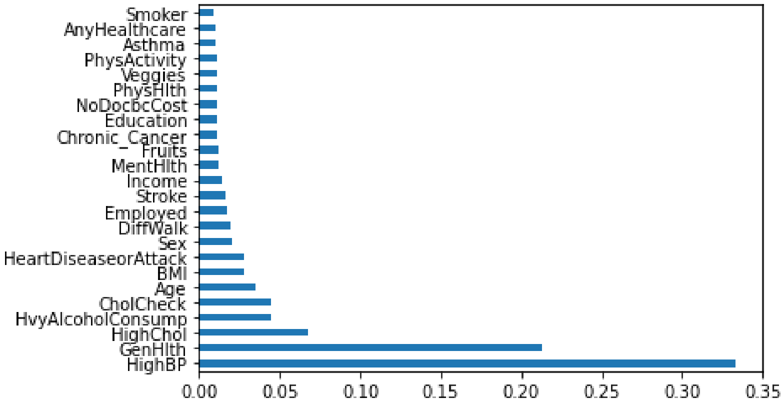


	Training Accuracy	Testing Accuracy	Recall	Precision	F1 Score	Roc-AUC Score	Runtime (s)
Logistical Regression	0.745076	0.749413	0.769187	0.738474	0.753518	0.749493	0
Random Forest	0.997313	0.742652	0.780620	0.724115	0.751307	0.742805	7
Decision Tree Classifer	0.997331	0.655327	0.653566	0.654033	0.653799	0.655320	0
XGB Classifier	0.791214	0.748345	0.788767	0.728389	0.757376	0.748509	5
SVC	0.769719	0.753541	0.806917	0.727765	0.765300	0.753756	350
GaussianNB	0.714614	0.717743	0.703016	0.722638	0.712692	0.717683	0
KNeighbors	0.796107	0.709487	0.734886	0.697788	0.715857	0.709589	275

Here is a confusion matrix of the XGB model.



Here is how it prioritized the features.



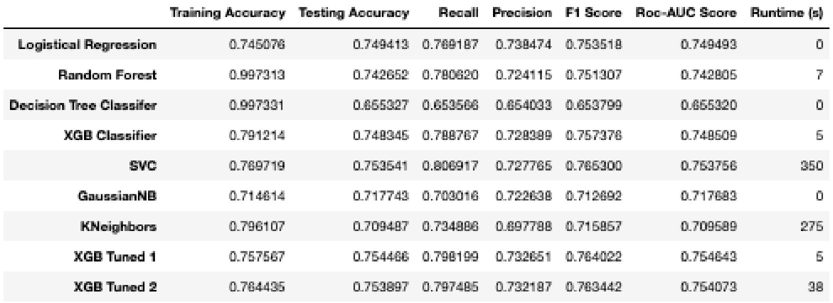
Tuning XGB

Next we tuned XGB using hyper parameter tuning. We used the Bayesian Optimization model to determine the best set of parameters for it.

The first iteration decreased the accuracy below the baseline accuracy.

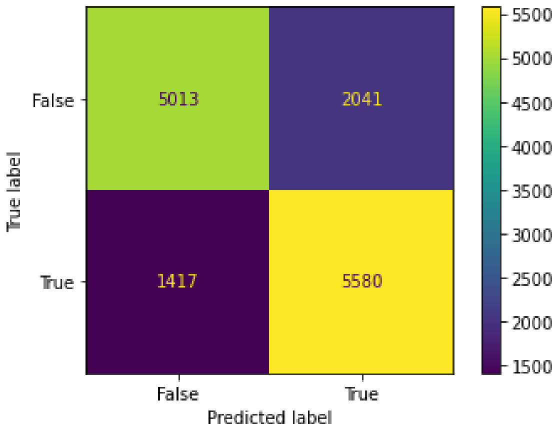
The second iteration marginal increased the accuracy of the XGB model.

Here is an updated table with the results of the tuned XGB models.

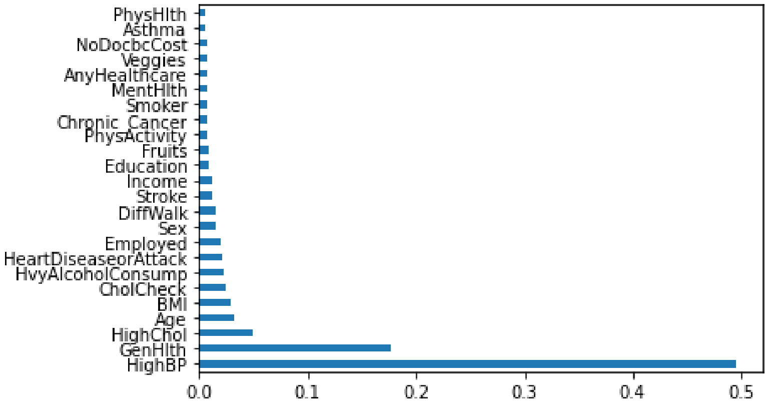


	Training Accuracy	Testing Accuracy	Recall	Precision	F1 Score	Roc-AUC Score	Runtime (s)
Logistical Regression	0.745076	0.749413	0.769187	0.738474	0.753518	0.749493	0
Random Forest	0.997313	0.742652	0.780620	0.724115	0.751307	0.742805	7
Decision Tree Classifer	0.997331	0.655327	0.653566	0.654033	0.653799	0.655320	0
XGB Classifier	0.791214	0.748345	0.788767	0.728389	0.757376	0.748509	5
SVC	0.769719	0.753541	0.806917	0.727765	0.765300	0.753756	350
GaussianNB	0.714614	0.717743	0.703016	0.722638	0.712692	0.717683	0
KNeighbors	0.796107	0.709487	0.734886	0.697788	0.715857	0.709589	275
XGB Tuned 1	0.757567	0.754466	0.798199	0.732651	0.764022	0.754643	5
XGB Tuned 2	0.764435	0.753897	0.797485	0.732187	0.763442	0.754073	38

Here is the confusion matrix of the second iteration of tuning.



Here are the weights of the features from the tuned XGB model. There does not seem to be a huge difference in the features compared to the initial XGB model. Though the coefficient for HighBP increased and the rest decreased.



Neural Networks

From reviewing literature, many researchers found a neural network model too be the most accurate.

Examples are:

- *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques*, Xie et. al. link
- This article used the 2014 data from the survey to create these models.
- *Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review*, Kee et. al. link

We created our own neural network based on the data. Due to the size and amount of text generated by neural networks, we ran them on a different notebook. We saved the best model and loaded it here to create the confusion matrix, graphs, etc.

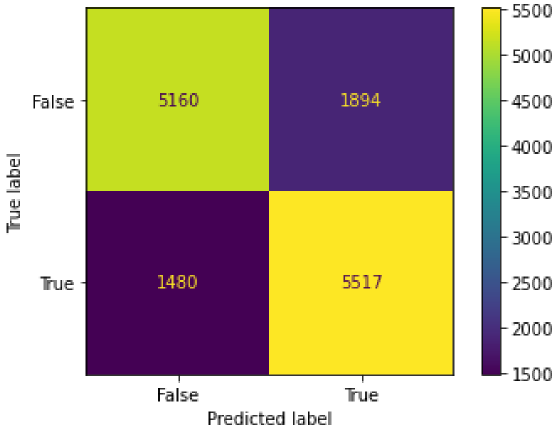
The analysis and notebook containing the optimization of the neural network is here

The neural networks architecture is:

Neural

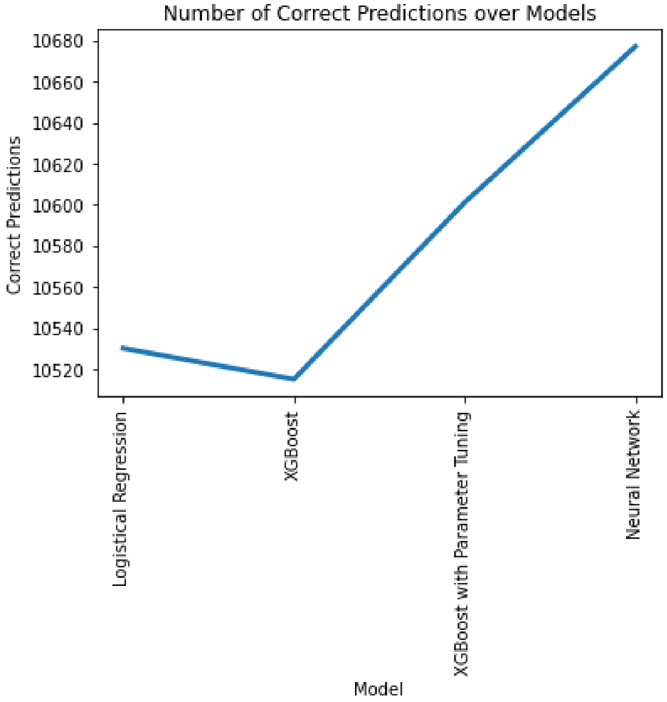
- 3 dense layers
- 40 neurons in the first layer
- 20 neurons in the second
- 10 neurons in the third
- relu activation
- Use sigmoid curve
- Early Stopping

Here are the results for the neural network depicted by a confusion matrix.

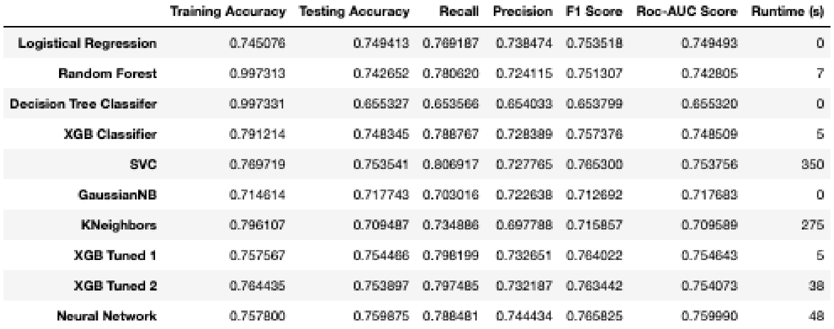


Final Model Evaluation

We have been iteratively improving our models. This line graphs shows that each models is more accurate.



Here is the table summarizing the metrics off all models.



	Training Accuracy	Testing Accuracy	Recall	Precision	F1 Score	Roc-AUC Score	Runtime (s)
Logistical Regression	0.745076	0.749413	0.769187	0.738474	0.753518	0.749493	0
Random Forest	0.997313	0.742652	0.780620	0.724115	0.751307	0.742805	7
Decision Tree Classifer	0.997331	0.655327	0.653566	0.654033	0.653799	0.655320	0
XGB Classifier	0.791214	0.748345	0.788767	0.728389	0.757376	0.748509	5
SVC	0.769719	0.753541	0.806917	0.727765	0.765300	0.753756	350
GaussianNB	0.714614	0.717743	0.703016	0.722638	0.712692	0.717683	0
KNeighbors	0.796107	0.709487	0.734886	0.697788	0.715857	0.709589	275
XGB Tuned 1	0.757567	0.754486	0.798199	0.732651	0.764022	0.754643	5
XGB Tuned 2	0.764435	0.753897	0.797485	0.732187	0.763442	0.754073	38
Neural Network	0.757800	0.759875	0.788481	0.744434	0.765825	0.759990	48

However, in-spite off the increased accuracies, we chose the Logistical Regression, our baseline model, as our final model. This model ran significantly faster compared too the others (less than a second vs 6 seconds for the second highest). Also, the Neural Network, the most accurate model, was only 0.2% more accurate than the logistical regression model. This tiny accuracy gain is not worth the significantly higher run time and more complex setup.

Recomendations

- The CDC should use the logistical regression model in their application.
- Consider a strategy around educating people to take their blood pressure on a regular basis since it was one of the top features.
- Providers who see people with high cholesterol should also screen for diabetes since high cholesterol was another top feature.
- Continue advocating for policy/strategies that aim to improve the general health and fitness of Americans. Low health was the most correlated feature with diabetes.

Future Projects

- Evaluate previous BRFSS data sets. Measure the rate of diabetes and other chronic conditions to find their trends across the country.
- Use the model to create an application on the CDC's website that allows a person to enter their data and get a diabetic risk score.
- Further investigate a strategy around making it easier for people to take and track their blood pressure. It was found to be the greatest predictor around diabetes.

Contact Information

- Linkedin: https://www.linkedin.com/in/dhruv-ragunathan-908993b1/
- Github: https://github.com/dragunat2016

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%