BIS634 Final project: Building Risk Prediction Model for Type 2 Diabetes Using Random Forest Algorithm

Mahima Kaur
Yale School of Public Health

Abstract

I planned to develop a disease risk assessment website for the final project. It helps the user to assess the risk of having diabetes on four levels: low risk, low to moderate risk, moderate risk, and high risk, and gives the user lifestyle recommendations based on their final score. I chose BRFSS-2015: Diabetes Health Indicator as the raw dataset, which met the FAIR principles of data and information. After data preprocessing and feature importance analysis, a random forest machine algorithm was applied as the classification prediction model. Based on the model risk category was defined, which would enable individuals to get the diabetes risk score and category. Finally, I deployed the website and built the page using Flask, Bootstrap, and plotly.

1. Background

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey collected annually by the Centers for Disease Control (CDC). Each year, the survey contains responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate glucose levels in the blood, which can lead to reduced quality of life and life expectancy. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes, and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics and roughly 8 in 10 prediabetics are unaware of their risk. Therefore, the project aims to build a risk prediction model and classify them into low-risk, low-to-moderate, moderate, and high-risk categories based on their risk score along with assessing the risk factor of diabetes.

2. Aim of the Project

1. To explore the distribution of the data variables in the dataset
2. To assess the association between data variable and diabetes status
3. To explore important features to predict diabetes risk
4. To find and choose a machine learning model for diabetes risk prediction.
5. To classify the individuals into low, moderate and high category for developing Type2 Diabetes Mellitus.

3. *Dataset Description*

For the project, I decided to work on a cleaned and consolidated dataset created from the BRFSS 2015 dataset available on Kaggle as the "*Diabetes Health Indicator BRFSS2015 dataset"*. The dataset was downloaded in a standard (csv) format for analysis; therefore, it was not required to be put in a standard format. The dataset contains 70,692 survey responses and has an equal 50-50 split of respondents with no diabetes and either prediabetes or diabetes. It is a. The target variable has two classes: zero is for no diabetes one is for prediabetes/ diabetes. The dataset has 22 feature variables, including the diabetes variable. I found the dataset interesting to explore as it contained variables such as perceived mental, physical, and general health, difficulty walking, and stroke. Therefore, the dataset would help find additional variables that might associate with and can be a risk predictor of diabetes.

3.1. Data Restrictions and FAIRness

The original dataset is provided by CDC from Kaggle public data repositories, named Behavioral Risk Factor Surveillance System. The metadata of the dataset is available and licensed on Kaggle. Alex Teboul created the subset of the data and made it available on Kaggle for public use.

FAIRness

I. **Findability:** The dataset is public and can be searched through the Internet.
II. **Accessibility:** People can distribute and perform the work without asking permission. The dataset is accessible and downloaded by anyone via Kaggle API.
III. **Interoperability:** The dataset is stored in .csv format, and it uses a formal, accessible, shared, and broadly applicable language for information representation.
IV. **Reusability:** The dataset is published with a clear and accessible data usage license, with a clear and detailed description of the file content, columns, provenance, and license specifications.

3.2. Data Variables

There are 22 variables included in the dataset. The variables are listed below:

Data Variable	***Definition***	***Classification***
Diabetes Status	Diabetic Status	0 = no diabetes 1 = prediabetes/ diabetes
High Blood Pressure (HighBP)	Do you have high blood pressure?	0 = no high BP 1 = high BP
High Cholesterol (HighChol)	Do you have high blood cholesterol?	0 = no high cholesterol 1 = high cholesterol
Cholesterol Check (Chol Check)	Did you get your blood cholesterol check in last 5 years?	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

Body Mass Index (BMI)	What is your BMI?	Continuous value in kg/m2
Smoking Status (Smoker)	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	0 = no 1 = yes
Stroke	Did the doctor ever told you had a stroke?	0 = no 1 = yes
Heart Disease Attack	Coronary heart disease (CHD) or myocardial infarction (MI)	0 = no 1 = yes
Physical Activity (PhyActivity)	Physical activity in past 30 days - not including job	0 = no 1 = yes
Fruits Intake (Fruits)	Consume Fruit 1 or more times per day	0 = no 1 = yes
Vegetable Intake (Veggie)	Consume Vegetables 1 or more times per day	0 = no 1 = yes
Heavy Alcohol Consumption (HvyAlcoholConsump)	(adult men >=14 drinks per week and adult women>=7 drinks per week)	0 = no 1 = yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.	0 = no 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	0 = no 1 = yes
GenHlth	Would you say that in general your health is:	1 = excellent, 2 = very good, 3 = good, 4 = fair 5 = poor
MentHlth	Days of poor mental health scale 1-30)	No. of Days
PhysHlth	Physical illness or injury days in past 30 days (scale 1-30)	No. of Days
DiffWalk	Do you have serious difficulty walking or climbing stairs?	0 = no 1 = yes
Sex	Male or Female	0 = female 1 = male
Age	Age in 13 different categories	1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80+
Education	Education Level Scale	1 = Never attended school or only kindergarten, 2 = elementary, 3 = Junior High School , 4 = Senior High

		School, 5 = Undergraduate Degree , 6 = Magister
Income	Income level	1 = less than $10,000, 2 = Between $10,000 and $15,000, 3 = Between $15,000 and $20,000, 4 = Between $20,000 and $25,000, 5 = Between $25,000 and $35,000, 6 = Between $35,000 and $50,000, 7 = Between $50,000 and $75,000, 8 = $75,000 or more

4. Data Cleaning and Preprocessing

- ***Checking for Duplicates:*** Duplicates were assessed, and it was found that there were rows with same values but there was no means to check whether they were duplicates or not since it is possible that two or more participants can have same values for all the variables. Based on this assumption the duplicate rows were not deleted.
- ***Missing Values:*** I observed that there is no missing values in dataset.
- ***Outlier Detection:*** BMI had 2181 outliers. There were BMI values more that 50 kg/m2 which seemed unreasonable, but a recent study demonstrated that severe morbid obesity could lead to BMI over 50, so we will not ignore that data. Additionally, Physical and Mental Health had outlier, but they were not removed since it was collected as a scale (in number of days) and within the range of the one-month period for instance it is possible for only few to have 30days.
- ***Renaming the variable categories:*** The variable's category was replaced in a copied dataset for better visualization e.g., 0 was replaced with No and 1 was replaced by yes.

5. ***Model Development:*** The dataset was split into training, cross-validation, and testing datasets. In the training phase of the model development, the training dataset was used to generate learned models for prediction. In the validation phase, the models were tested with the features of the testing dataset to evaluate how well they predicted the corresponding class labels of the testing dataset. A grid-search approach with parallelized performance evaluation for model parameter tuning was used for each model to generate the best model parameters. The models were also passed through 5-fold cross-validation to measure model performance accurately. In addition, to increase the accuracy of the models, hyperparameter tuning was done where the number of trees or max nodes was tweaked, and the effect on accuracy was checked.

    - ***Feature Selection:*** With a goal of creating an accurate model relying on a limited set of available features, i.e., features that did not require excessive questioning or testing of patients, I evaluated the feature dependence of the models for prediction of diabetes. The analysis was done based on the random forest classifier and chi-square. Random forests are among the most popular machine learning methods thanks to their relatively good accuracy, robustness, and ease of use. In addition, Chi-Square can also be used when the feature is categorical, or the

target variable is any way can be thought as categorical. It measures the degree of association between two categorical variables.

- ***Machine Learning Models Used for the Project:***
    a. ***K-Nearest Neighbors:*** It is a classification technique which classifies the new sample based on similarity measure or distance measure.
    b. ***Random Forest:*** It is supervised learning, used for both classification and Regression. The logic behind the random forest is bagging technique to create random sample features. The highest voted class is the final prediction of the random forest.
    c. Decision Trees Classification: It is a supervised learning method, which is used for solving classification problems. Decision tree is a technique which iteratively breaks the given dataset into two or more sample data. The goal of the method is to predict the class value of the target variable.
- ***Performance Metrics:*** Lastly the scores of the models such as accuracy, F-Scores, Recall precision were compared to evaluate their performance in predicting cases and choose the most accurate model.

6. Results

6.1. Summary Statistics

The summary statistics analysis revealed that the dataset had 70692 rows and 22 columns, there were no missing values and duplicates in the dataset. The dataset has 2181 BMI outliers, 11816 outliers in Mental Health variable and 10624 outliers in Physical Health variable. It was seen that high blood pressure count and high cholesterol count was more among the diabetic patients. Interestingly cholesterol check in the past 5year had higher count than among the non-diabetic group. The mean BMI was 29.5 ± 7.11 kg/m2 which lies in the overweight category as per the BMI classification. The minimum and maximum BMI was 12 kg/m2 and 98kg/m2 respectively. Overall, the individuals perceived general health lied in the 'good' category (Fig1)

	count	unique	top	freq
Diabetes_binary	70692	2	No Diabetes	35346
HighBP	70692	2	High BP	39832
HighChol	70692	2	High Cholesterol	37163
CholCheck	70692	2	Cholesterol Check in 5 Years	68943
Smoker	70692	2	No	37094
Stroke	70692	2	No	66297
HeartDiseaseorAttack	70692	2	No	60243
PhysActivity	70692	2	Yes	49699
Fruits	70692	2	Yes	43249
Veggies	70692	2	Yes	55760
HvyAlcoholConsump	70692	2	No	67672
AnyHealthcare	70692	2	Yes	67508
NoDocbcCost	70692	2	No	64053
GenHlth	70692	5	Good	23427
DiffWalk	70692	2	No	52826
Sex	70692	2	Female	38386
Age	70692	13	65 to 69	10856
Education	70692	6	Magister	26020
Income	70692	8	$75,000 or More	20646

Fig1: Summary Statistics of Categorical Variables

6.2. Univariate Analysis

The dataset had a balanced outcome variable i.e., it had 50-50 split of diabetic and non-diabetic participants. About 53% of the participants were female. Around 56% of the participants had high blood pressure and 54% of the participants had high blood cholesterol. It was interesting to see that majority of the participants indulged themselves in physical activity other than job, (70%) and consumed fruits (61%) and vegetables (79%) one or more times per day. Only 8% of the participants had general poor health and around 33% had a considered to have good health. Around one-third of the participants earned $75,000 or More and about 35% of the participants had a master's or equivalent degree. It was interesting to see that the dataset contained BMI more than 50 kg/m2 (Figure2).

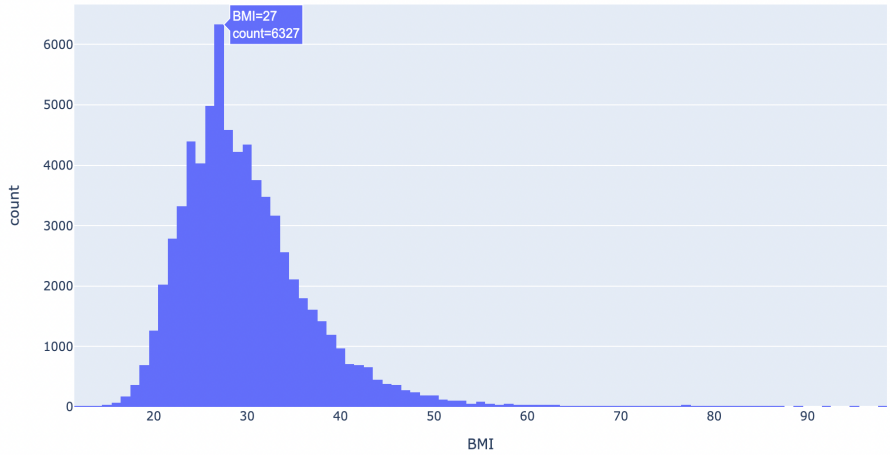

Figure2: Distribution of BMI

6.3. Bivariate Analysis

About 75% of the participants in the diabetic group had high blood pressure, while only 37% of the participants in the non-diabatic group had high blood pressure. Looking at the high cholesterol variable, it was seen that around 67% of the diabetes participants had high blood cholesterol levels. In comparison, only 38% of the participants with no diabetes had high cholesterol. Among the diabetes group, half of the female participants (52%) had diabetes; in the non-diabetes group, around 56% of the female did not have diabetes. The mean BMI of diabetes participants is 32 kg/m2, while the mean of non-diabetes individuals is 27 kg/m2. The bivariate analysis shows that high blood pressure, high cholesterol, and BMI influence diabetes outcomes (Figure 3). Variables such as fruits, veggies, smoking, and physical activity didn't have much difference in both groups, even though there is research stating the importance of these factors.

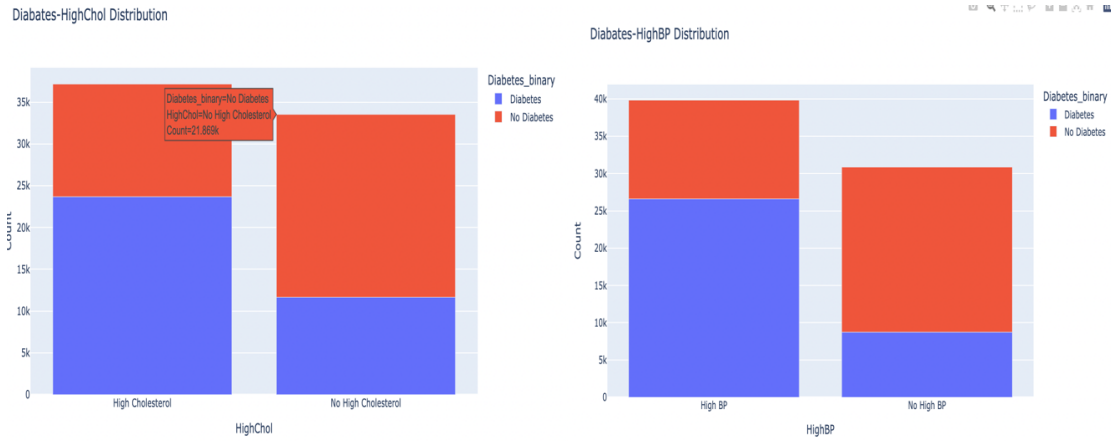

Figure 3: Distribution of High Cholesterol and High Blood Pressure among Diabatic and Non-Diabetes patients.

6.4. Correlation Matrix

Since the dataset had both numerical and categorical variables, the Dython module was used in python. The module automatically finds which features are categorical and which are numerical, computes a relevant measure of association between each feature, and plots it all as an easy-to-read heat-map. Pearson correlation was calculated to find the association between the numerical variables, while Cramer correlation was calculated to find the association between the categorical variables. Interestingly, none of the variables correlated more than 0.5 (Figure 4). High Blood Pressure and general Health has a substantial correlation with diabetes. It was interesting to see gender, smoking status, and intake of fruits and vegetables had a low correlation. BMI had a correlation of 0.29 with diabetes. Other exciting insights included:

- A substantial correlation between general Health, Age with High Blood pressure.

- A 0.50 correlation between General Health and difficulty in walking.

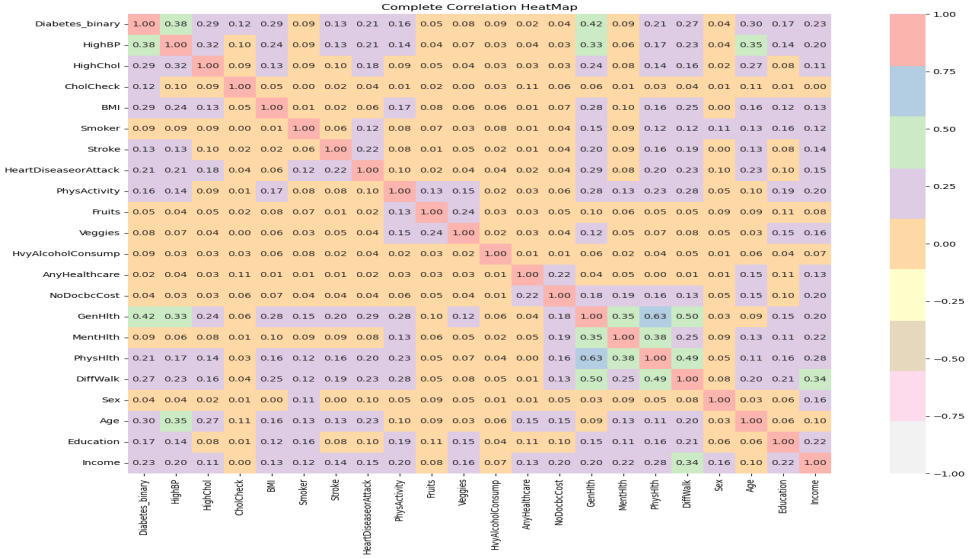

Figure4: Correlation matrix between variables

6.5.Modeling

The first involved feature selection with the goal of creating an accurate model relying on a limited set of available features so that the risk calculator can be used with minimal questioning or testing of patients. The feature selection was made based on Random Forest and chi-square. Based on the random forest, the top 5 features were BMI, Age, General Health, Income, and High Blood Pressure. It was seen that General Health seems to be the most important factor, and parameters like stroke, fruit and vegetable intake, and alcohol don't affect that much. The chi-square analysis did not give a rank to the features therefor, for the final selection, only the parameter analyzed in the Random Forest was included for future modeling (Figure 5).

Before training the model, feature scaling was done. Standardizing a dataset involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1. This can be thought of as subtracting the mean value or centering the data. Scaling the features is of utmost importance because different features are in various scales. The best practice is to use only the training set to figure out how to scale/normalize, then blindly apply the same transform to the test set. Three models were trained including K-Nearest Neighbours, Decision Tree classification, and Random Forest Classification. A grid-search approach with parallelized performance evaluation for model parameter tuning was used for each model to generate the best model parameters. In the training phase of the model development, the training dataset was used to create learned models for prediction. In the validation phase, the models were tested with the features of the testing dataset to evaluate how well they predicted the corresponding class labels of the testing dataset. Among the models used, the accuracy of the Random Forest was the highest (Figure 6). Therefore, Random Forest was used to making

final predictions on the user's data and to assess the diabetes risk among the users to give better health advice.

The Diabetes Risk Index was calculated out of 50 by using the formula risk*0.5*100. The risk categories included: less than 0.3 (low risk), 0.3 to 0.7 (low to moderate risk), 0.7 to 0.9 (moderate risk) and more than 0.9 to 1.0 high risk.

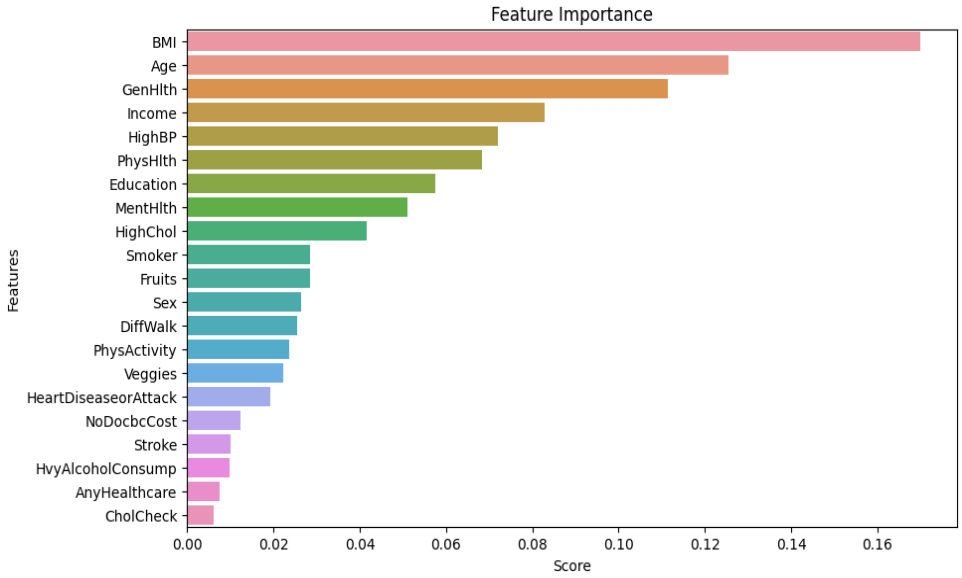

Figure 5: Graph depicting Feature selection using Random Forest

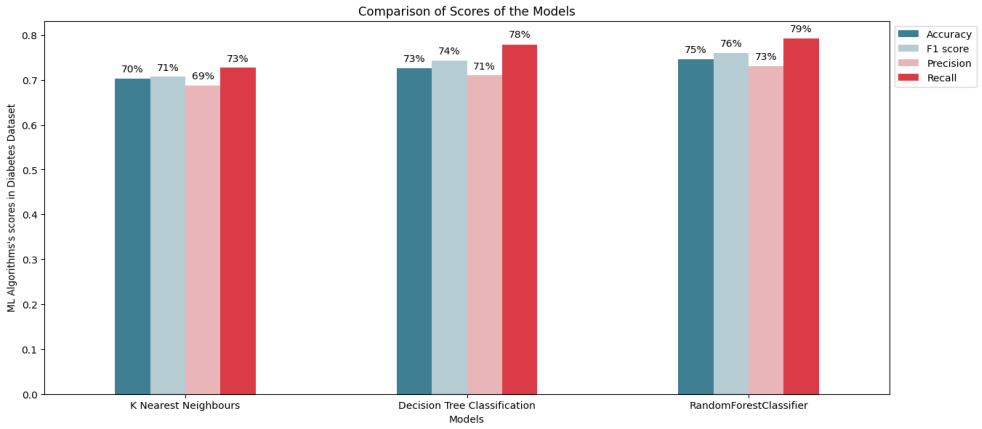

Figure 6: Graph depicting model evaluation in terms of accuracy, recall, F1 Scores and Precision

7. Deployment of the Website

The framework of API server is showed in the Fiure8. One could access the web page through port 5000. To create the website, I first created a flask application. After that, I created around ten routes which was rendered from each corresponding page from html file. In the main function, it runs the application(app.py). In the last part, we deployed our trained random forest classification model on the webpage via flask. I used Bootstrap4 for the styling the HTML pages and plotly for the interactive graphics. Bootstap4 is the fourth version of Bootstrap which is a set of open-source front-end frameworks for web site and web application development, including HTML, CSS and JavaScript frameworks and many components. Plotly is a popular and powerful data visualization module. We can create interactive plots just by feeding certain parameters. The web front-end has twelve web pages in total. User can access each page by clicking the navigation bar. The navigation bar has buttons including Home, Background, Data, Analysis and Risk Calculator. Below is in detail the webpages included on the website:

- ***The Homepage:*** The homepage shows the title of the website and contains the start button which navigates the user to the main website.
- ***The Background page:*** The background page displays the information about diabetes, its symptoms, risk factors, prevalence in USA and some prevention strategies.

- ***Under the Data Button:***
  - ***About the Dataset:*** This page gives information the dataset used to prepare the project and the risk calculator.
  - ***Data Variables:*** This page includes information the data variables, in the dataset its definition and code/categories.
  - ***Data Restrictions and FAIRness:*** This page explains the data restriction and assess the FAIRness of the dataset used.
- ***Under the Analysis Button:***
  - ***Summary Statistics:*** This page displays the summary tables including frequency, mean, standard deviation, minimum, maximum etc. of the categorical and numerical values of the whole dataset, and of the diabatic and non-diabetic individuals.
  - ***Univariate Analysis:*** This page displays the interactive figures of univariate analysis of all the 22 variables in the dataset. User can choose the variable from the dropdown button and choose the variable that they are interested to see.
  - ***Bivariate Analysis:*** This page displays the interactive figures of bivariate analysis of all the 21 variables in the dataset with the diabetes variable User can choose the variable from the dropdown button and choose the variable that they are interested to see.
  - ***Correlation Analysis:*** This page includes image and insights of the correlation matrix between different dataset variables.
  - ***Predictive Analytics:*** This page includes image and insights of the predictive models used to predict the risk of diabetes in an individual.
- ***Under the Analysis Button:***
  - ***DIARISK Calculator and Result Page:*** This page contains the calculator to assess the risk category and diabetes risk score. DIARISK is a prediction tool to identify patients at risk of developing diabetes. DIARISK uses age, BMI, physical health, general health, mental health, blood Pressure, education level, income level, and gender to determine risk of developing diabetes of an individual. The user can select the parameters themselves. After

submitting the parameters, the user can see the results which includes diabetes risk category, diabetes risk score and recommendation (Figure9).

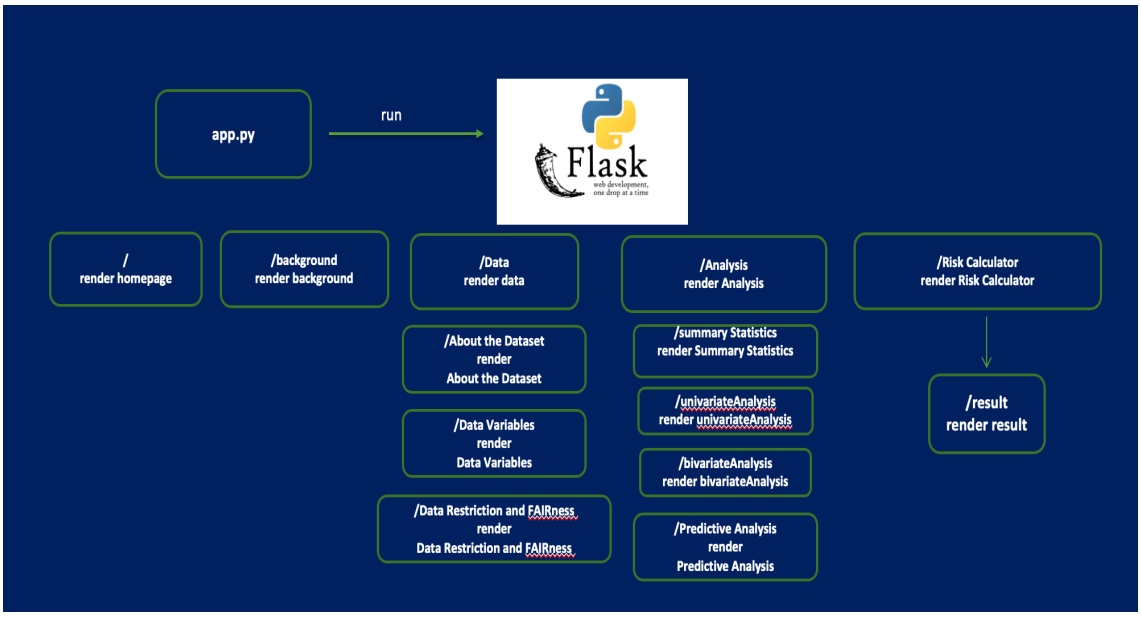

Figure8: Framework of API server

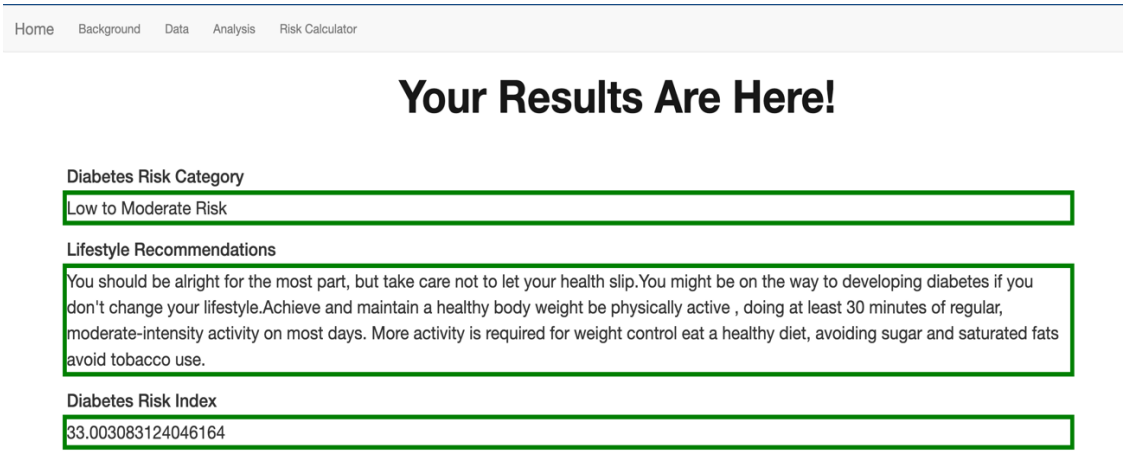

Figure9: DIARISK Results Page

8. *Discussion and Conclusion*

The project's main aim was to explore the diabetes indicator dataset available for public use on Kaggle and to classify the individuals into the low, moderate, and high categories for developing Type2 Diabetes Mellitus using a machine learning model. Some interesting insights included that the mean BMI was higher among the diabetic group, and both genders are equally vulnerable to

having diabetes. With higher education and income level, the number of diabetic individuals decreases, while with age, the number of diabetic people increases. The correlation matrix releveled that diabetes and Blood Pressure are substantially correlated. BMI, Blood Pressure, Age, Education, and General Health were among the crucial features in predicting diabetes risk compared to Heavy Alcohol Consumption, Stroke, and heart disease. The random forest had the highest accuracy score among the predictive models used. As shown in our analysis, machine-learned models show promising results in detecting diabetes in patients. Possible real-world applicability of such a model can be in the form of a web-based tool, where a survey questionnaire can be used to assess the disease risk of participants. Based on the score, the participants can opt to conduct a more thorough check-up with a doctor.

References

1. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
2. https://www.cdc.gov/brfss/index.html
3. Kermansaravi, M., Lainas, P., Shahmiri, S.S. *et al*. The first survey addressing patients with BMI over 50: a survey of 789 bariatric surgeons. *Surg Endosc* **36**, 6170–6180 (2022). https://doi.org/10.1007/s00464-021-08979-w
4. Kulkarni AR, Patel AA, Pipal KV, *et al* Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogra *BMJ Innovations* Published Online First: 09August2022. doi: 10.1136/bmjinnov-2021-000759.
5. Mujumdar, A. and Vaidehi, V., 2019. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, pp.292-299.