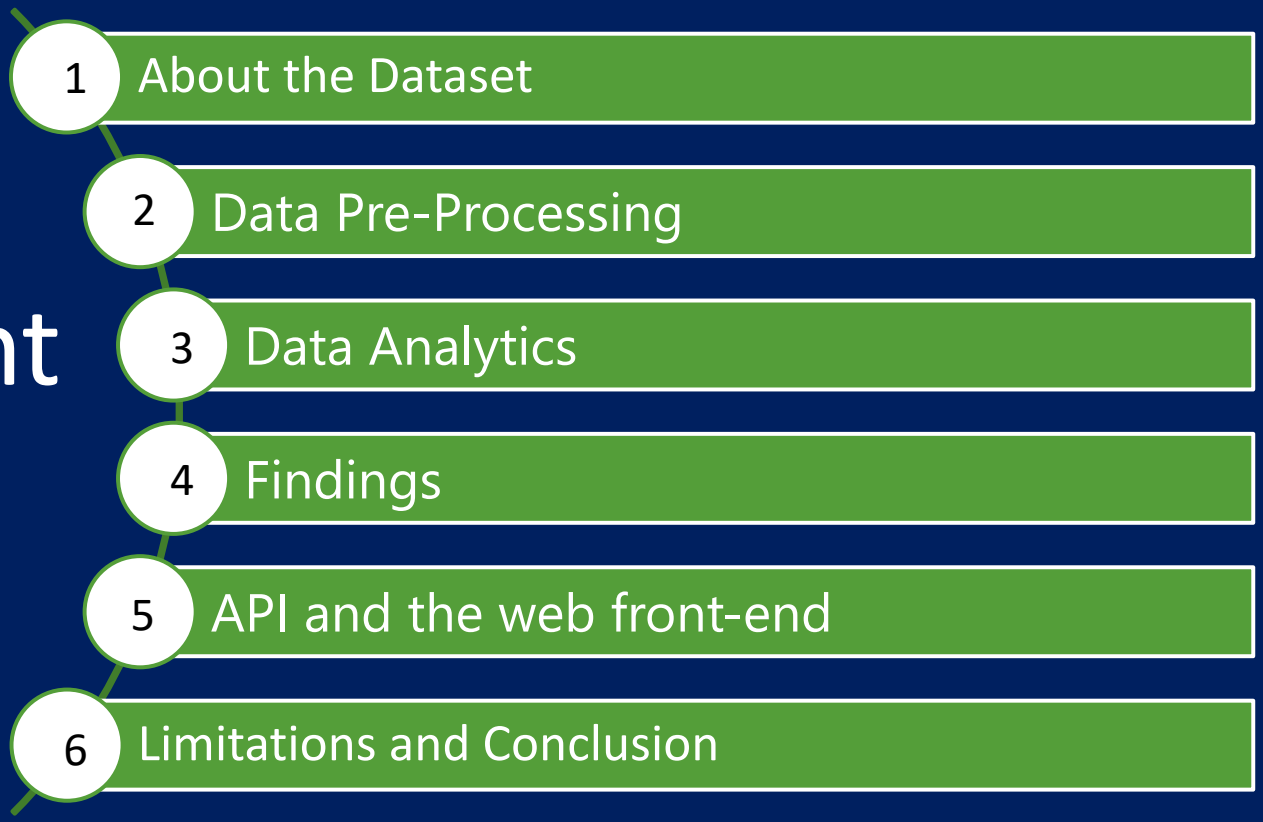
Final Project
Diabetes Risk Predictor

Mahima Kaur
MSc Health Informatics

Yale SCHOOL OF PUBLIC HEALTH

Content


1
About the Dataset
2
Data Pre-Processing
3
Data Analytics
4
Findings
5
API and the web front-end
6
Limitations and Conclusion

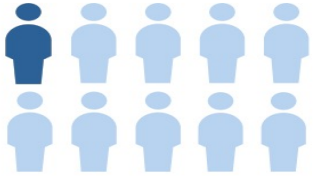
DIABETES IN THE U.S

A SNAPSHOT

DIABETES

37 Million

37 million people **have diabetes**



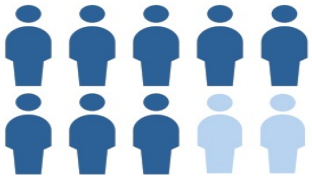
That's about **1 in every 10** people

1 in 5 people **don't know they have it**

PREDIABETES

96 Million

96 million American adults—**more than 1 in 3** —have prediabetes



More than 8 in 10
adults with prediabetes
don't know they have it

Goal?

To classify the individuals into low, moderate and high category for developing Type2 Diabetes Mellitus.

About The Behavioral Risk Factor Surveillance System (BRFSS)

- The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the Centers for Disease Control (CDC).
- Each year, the survey collects 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.
- For this project, a csv of the dataset available on Kaggle for the year 2015 was used.
- This original dataset contains responses from 441,455 individuals and has 330 features.
- These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

About The Dataset

- Diabetes Health Indictor BRFSS2015 dataset : Clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes.
- It is a subset of the original data.
- The target variable Diabetes has 2 classes : 0 is for no diabetes 1 is for prediabetes/ diabetes.
- This dataset has 22 feature variables.
- It was cleaned and consolidated dataset created from the BRFSS 2015 dataset already on Kaggle.

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

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.

Dataset Variables

Income	Education	Age	Mental Health	Gender	High Blood Pressure
Diabetes	Cholesterol Check	Smoker	Stroke	Heart Disease Attack	Fruits
Heavy Alcohol Consumption	High Cholesterol	Physical Activity	BMI	Veggies	Any Healthcare
	Physical Health	General Health	Diff Walk	NoDocbcCost	

Cleaning Process

Note : Now the data is ready for the Analysis

Analysis Questions

Exploratory data analysis (EDA) :

1. What are the variables associated with Diabetes? (Is Blood Pressure associated with Diabetes? Is BMI associated with Diabetes?)

Modeling :

1. What are the important features to predict diabetes risk?
2. Choose a machine learning model for risk prediction.

Summary Statistics of Categorical Variables

	count	unique	top	freq
Diabetes_binary	35346	1	Diabetes	35346
HighBP	35346	2	High BP	26604
HighChol	35346	2	High Cholesterol	23686
CholCheck	35346	2	Cholesterol Check in 5 Years	35105
Smoker	35346	2	Yes	18317
Stroke	35346	2	No	32078
HeartDiseaseorAttack	35346	2	No	27468
PhysActivity	35346	2	Yes	22287
Fruits	35346	2	Yes	20693
Veggies	35346	2	Yes	26736
HvyAlcoholConsump	35346	2	No	34514
AnyHealthcare	35346	2	Yes	33924
NoDocbcCost	35346	2	No	31604
GenHlth	35346	5	Good	13457
DiffWalk	35346	2	No	22225
Sex	35346	2	Female	18411
Age	35346	13	65 to 69	6558
Education	35346	6	Senior High School	11066
Income	35346	8	$75,000 or More	7195

	count	unique	top	freq
Diabetes_binary	35346	1	No Diabetes	35346
HighBP	35346	2	No High BP	22118
HighChol	35346	2	No High Cholesterol	21869
CholCheck	35346	2	Cholesterol Check in 5 Years	33838
Smoker	35346	2	No	20065
Stroke	35346	2	No	34219
HeartDiseaseorAttack	35346	2	No	32775
PhysActivity	35346	2	Yes	27412
Fruits	35346	2	Yes	22556
Veggies	35346	2	Yes	29024
HvyAlcoholConsump	35346	2	No	33158
AnyHealthcare	35346	2	Yes	33584
NoDocbcCost	35346	2	No	32449
GenHlth	35346	5	Very Good	13491
DiffWalk	35346	2	No	30601
Sex	35346	2	Female	19975
Age	35346	13	60 to 64	4379
Education	35346	6	Magister	15620
Income	35346	8	$75,000 or More	13451

Summary Statistics of Numerical Variables

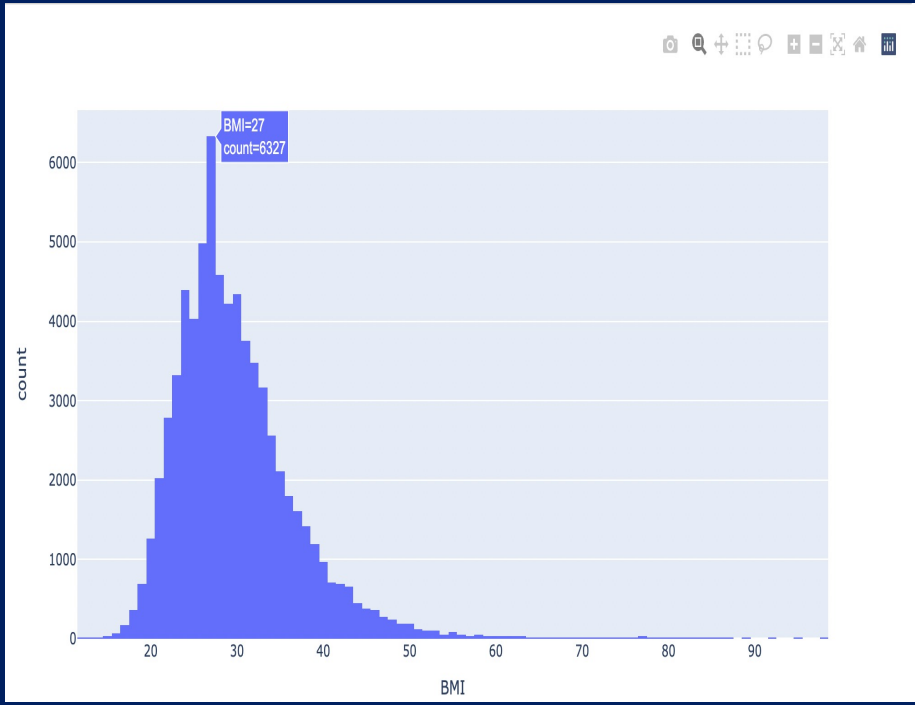
	count	mean	std	min	25%	50%	75%	max
BMI	35346.0	31.944011	7.363401	13.0	27.0	31.0	35.0	98.0
MentHlth	35346.0	4.461806	8.947717	0.0	0.0	0.0	3.0	30.0
PhysHlth	35346.0	7.954479	11.301491	0.0	0.0	1.0	15.0	30.0

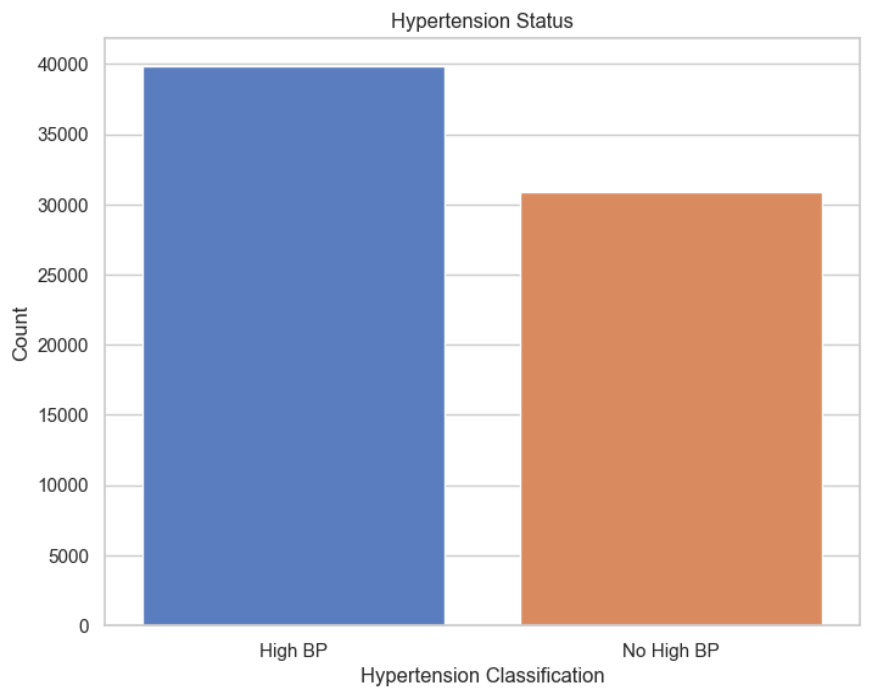
Diabetic Individuals

	count	mean	std	min	25%	50%	75%	max
BMI	35346.0	27.769960	6.187636	12.0	24.0	27.0	31.0	98.0
MentHlth	35346.0	3.042268	7.208408	0.0	0.0	0.0	2.0	30.0
PhysHlth	35346.0	3.666355	8.098339	0.0	0.0	0.0	2.0	30.0

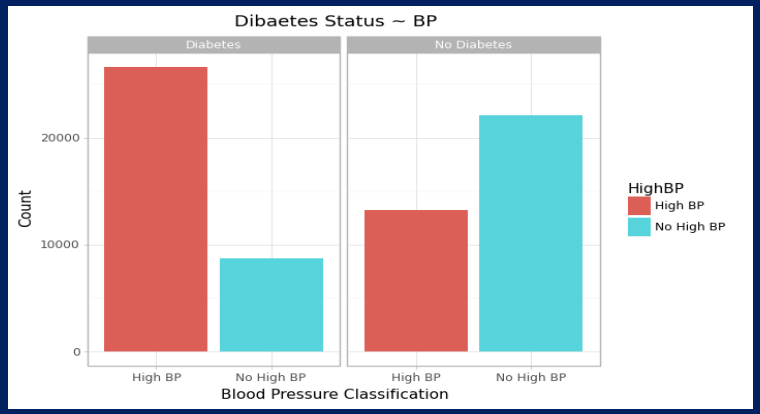
Non-Diabetic Individuals

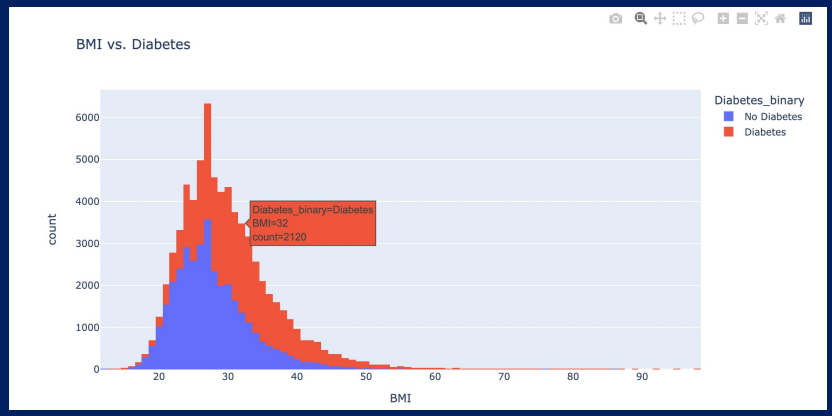
Univariate Analysis

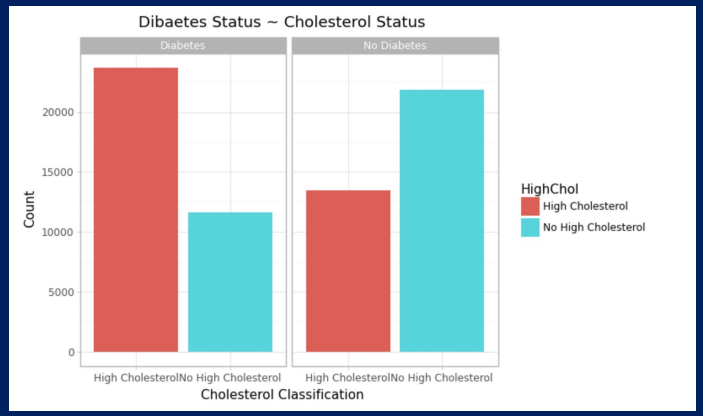




Bivariate Analysis

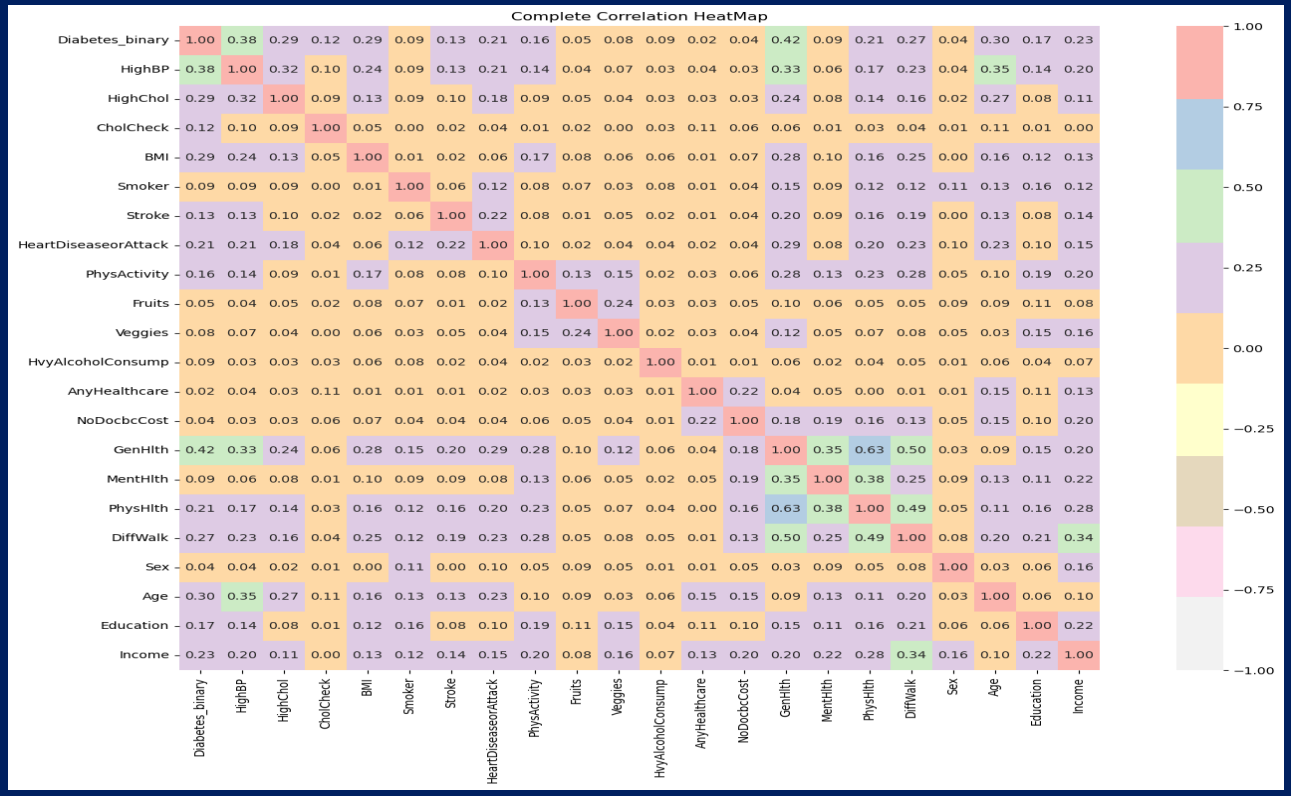






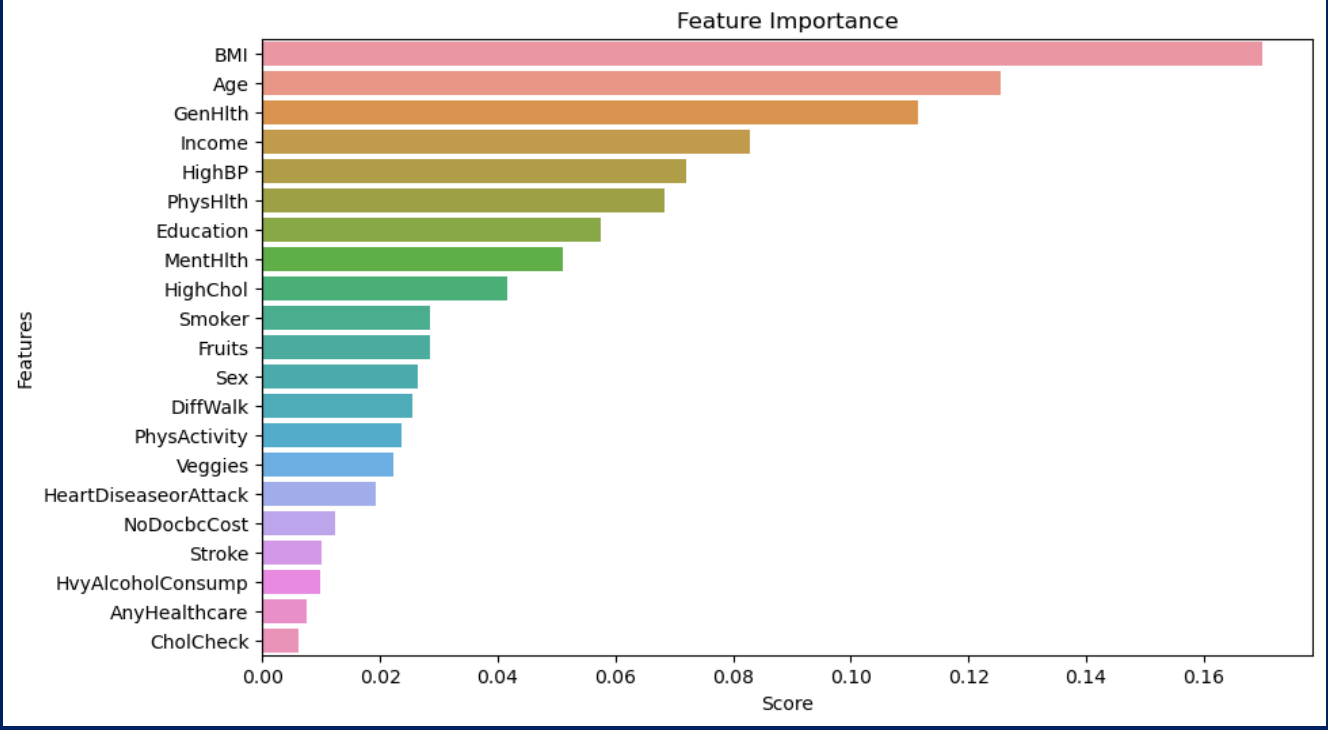
Yale SCHOOL OF PUBLIC HEALTH

Correlation Analysis

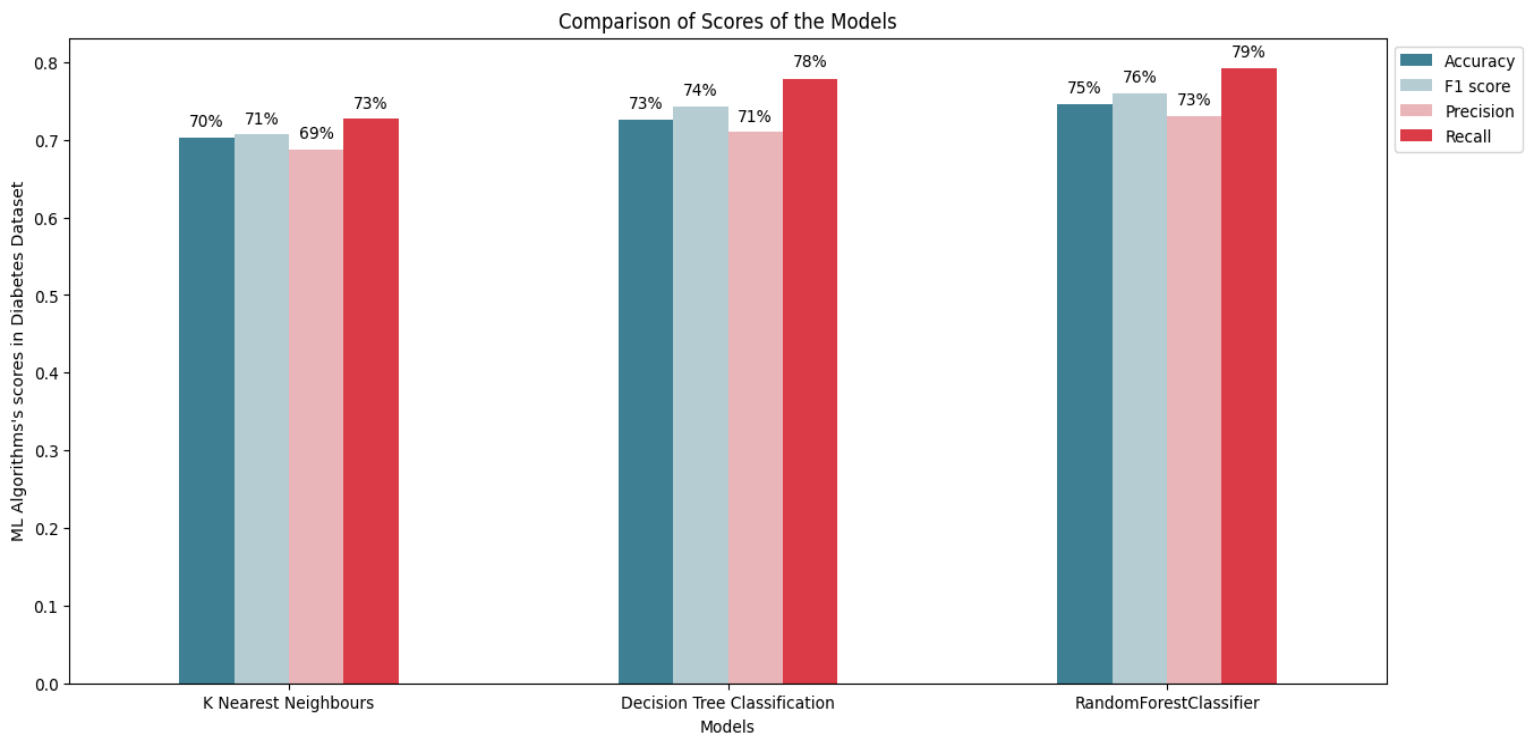


How can we Predict the Diabetes Risk ?

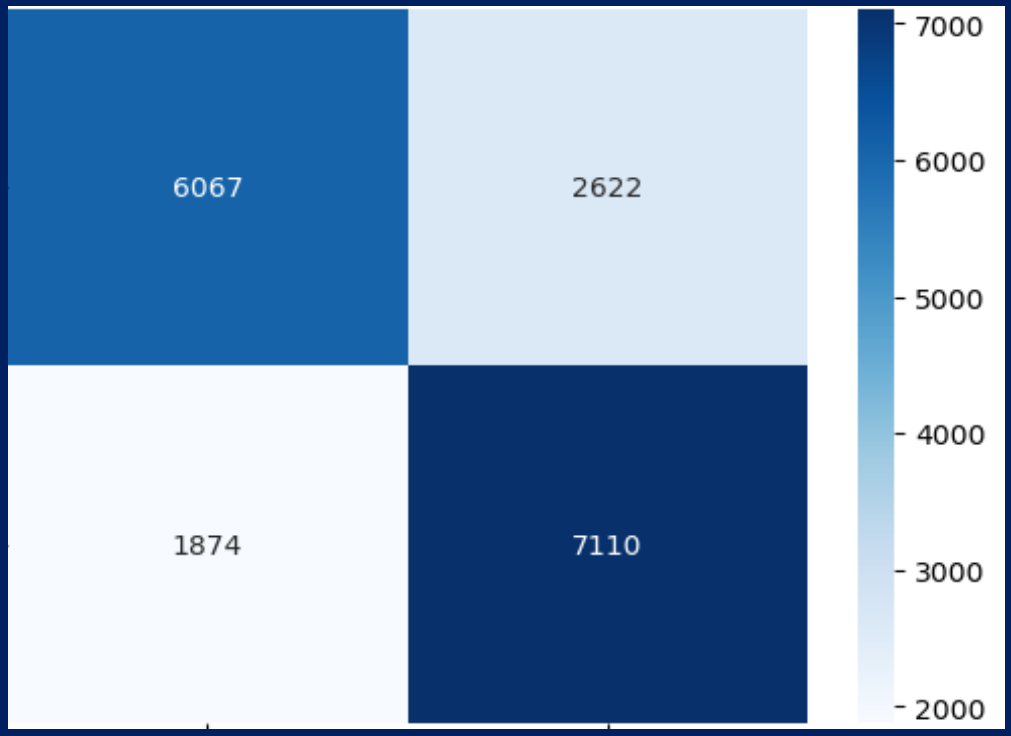
Features Identification



Modelling with Important Features



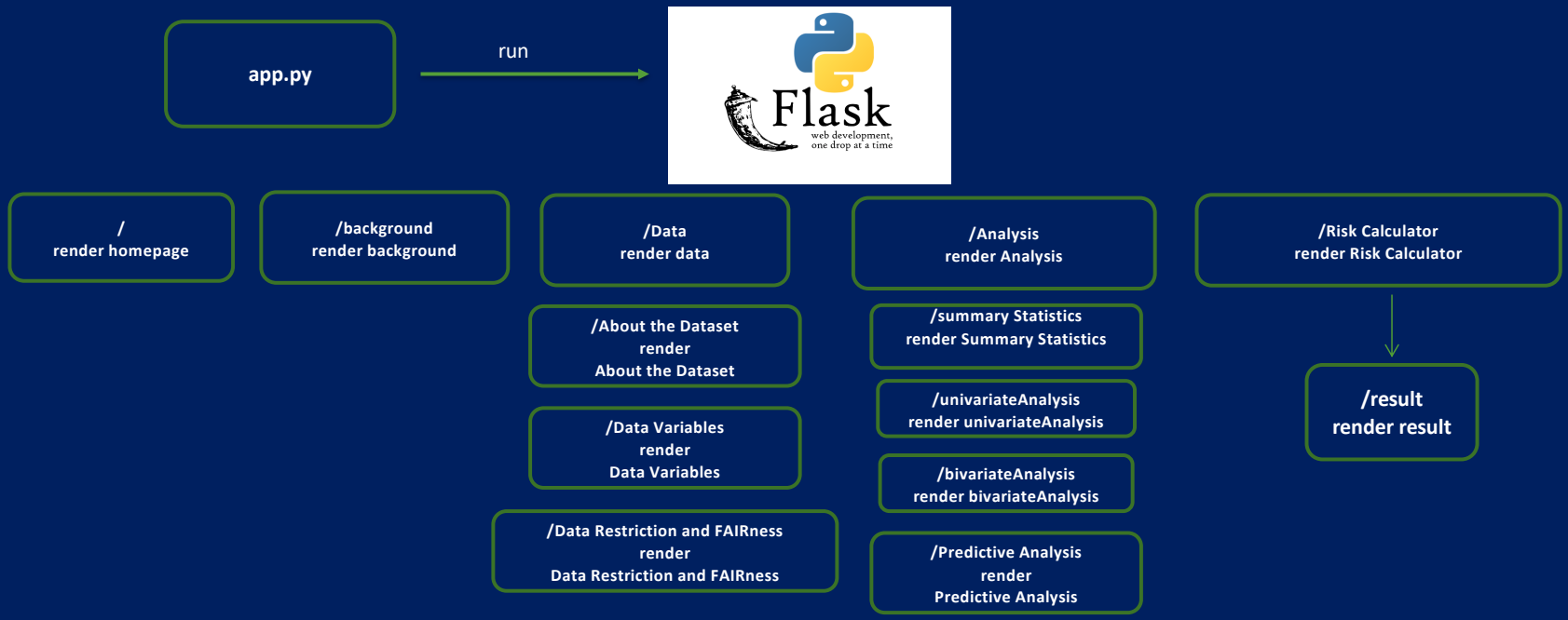
Confusion Matrix of Random Forest Classifier



Let's Find the Diabetes Risk Index

```
Do you have High Blood Pressure (0 = no high BP, 1 = high BP?):
1
Enter your Body Mass Index (kg/m2):
30
Physical illness or injury days in past 30 days (scale 1-30)?:
5
Would you say that in general your health is: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor):
4
Days of poor mental health scale 1-30 days:
6
Enter your Age category : 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+:
8
Enter Education level on the scale 1-6 : 1 = Never attended school or only kindergarten, 2 = elementary, 3 = Junior H
igh School , 4 = Senior High School, 5 = Undergraduate Degree , 6 = Magister:
6
Income scale scale 1-8 1 = less than $10,000, 2 = Between $10,000 and $15,000, 3 = Between $15,000 and $20,000, 4 = B
etween $20,000 and $25,000, 5 = Between $25,000 and $35,000 , 6 = Between $35,000 and $50,000, 7 = Between $50,000 an
d $75,000, 8 = $75,000 or more :
4
Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes :
1
What is your gender? 0 = female 1 = male :
1
-------------------------
Health Indicator Analysis
-------------------------
See a doctor as soon as you can and listen to their recommendations. You might be on the way to developing diabetes i
f you don't change your lifestyle.
Your Diabetes Risk Index is 39.60/50.
```

API and the web front-end



Limitations

- Lower Accuracy of the Models.

Conclusion

- The mean BMI was higher among the diabetic group.
- Male and Female are equally vulnerable for Diabetes.
- With higher Education and Income level the number of Diabetic individuals decreases.
- With age the number of diabetic people also increase.
- Diabetes and Blood Pressure are substantially correlated.
- BMI, Blood Pressure , Age, Education, and General Health were among the important features to predict diabetes risk in contrast to Heavy Alcohol Consumption, Stroke, Heart Disease.
- Random Forest has the highest accuracy score.

Thank you
Questions?.....