Classification and prediction of diabetes disease using machine learning algorithms

Qavamodin Soleimani

Advisor: Dr. Fereshteh Dehghani

September 13, 2023

Contents

Introduction

What is diabetes?

What do the statistics say?

Global statistics

About 422 million people worldwide have diabetes and 1.5 million deaths are directly attributed to diabetes each year.

What to do?

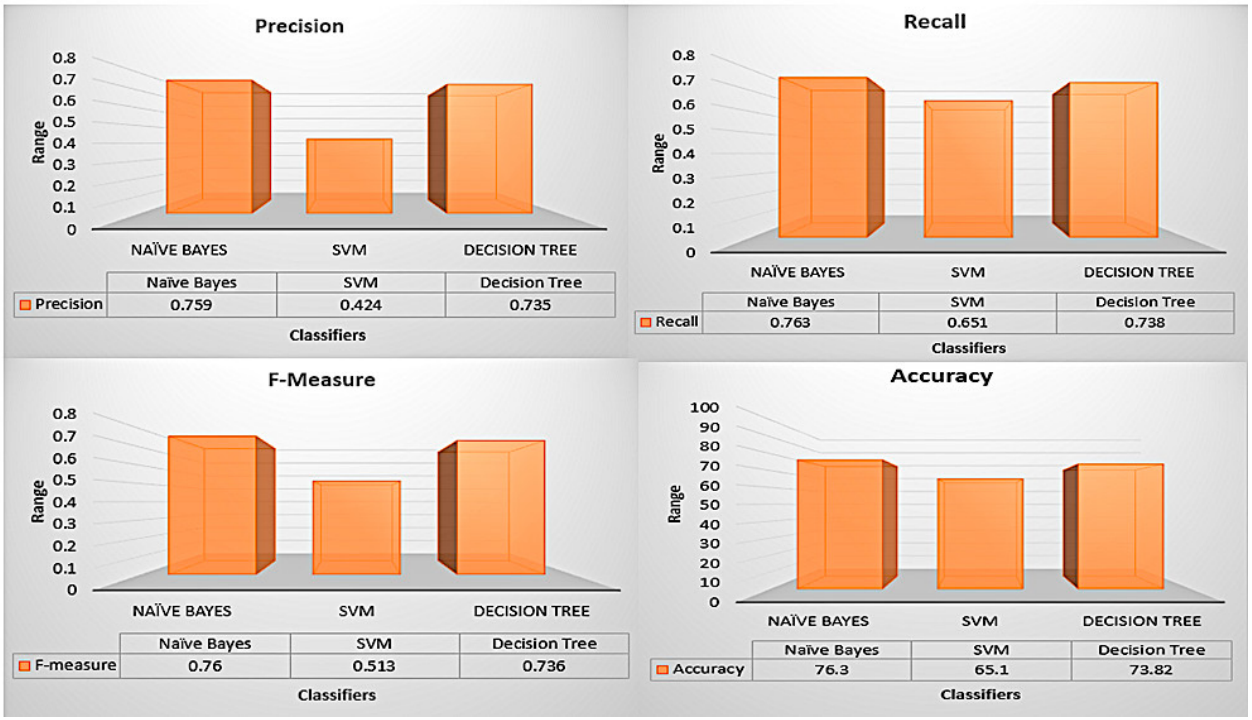
Methods of preventing diabetes.

Predicting diabetes

Scientists found that BMI, age, systolic and diastolic blood pressure, and a family history of diabetes were the most significant predictive features for prediabetes (Lama et al., 2021)

Previous researchs

- Prediction of Diabetes using Classification Algorithms 2018 (Deepti Sisodiaa, Dilip Singh Sisodia) FS:Genetic algorithm , Algo: SVM,NB(ACC=76%),DT



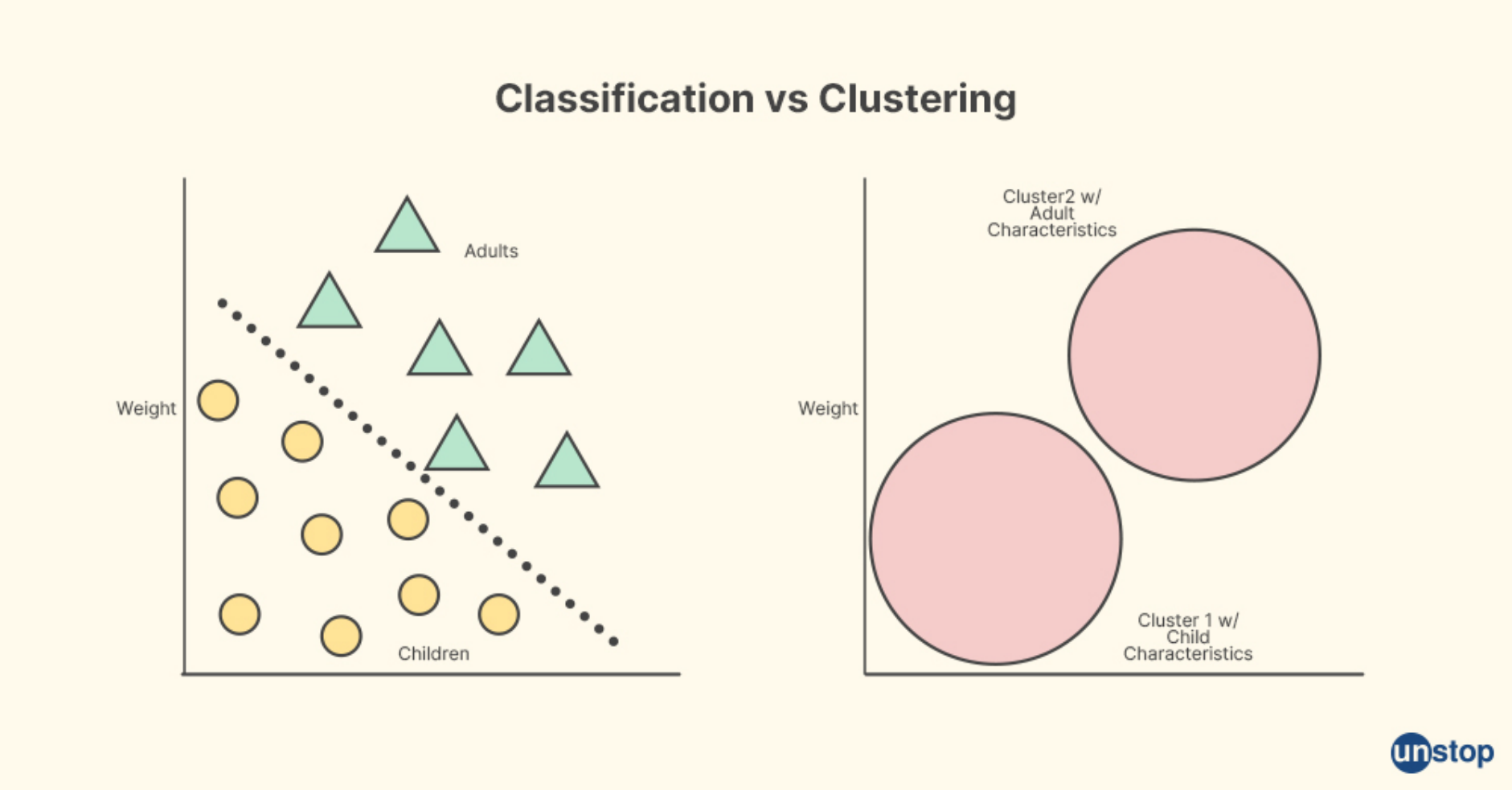
- Classification and prediction of diabetes disease using machine learning paradigm 2020 (Md. Maniruzzaman1,2*, Md. Jahanur Rahman2, Benojir Ahammed1 and Md. Menhazul Abedin) FS:LR ,

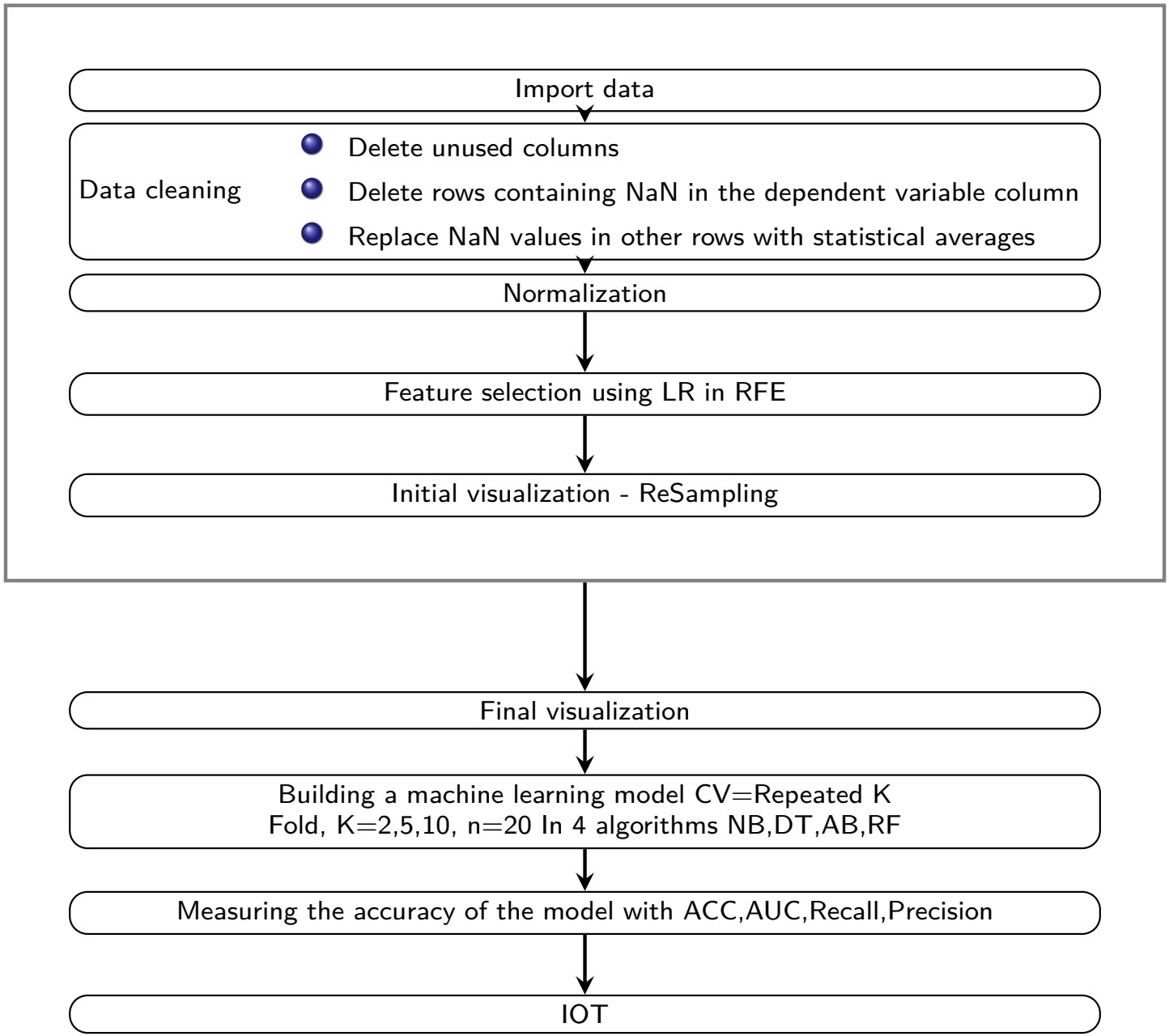
Previous researchs

- Classification and prediction of diabetes disease using machine learning paradigm 2020 (Md. Maniruzzaman1,2*, Md. Jahanur Rahman2, Benojir Ahammed1 and Md. Menhazul Abedin) FS:LR , Algo: AB,NB,DT,RF(ACC=96%) ,

Data mining methods

Supervised vs. unsupervised learning: Which is best for you?





Data preprocessing

Data and processing

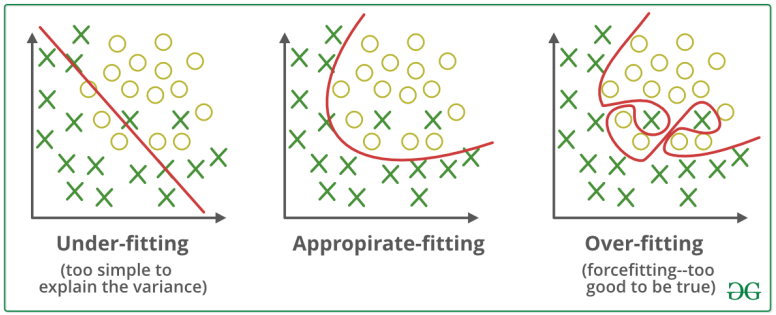
- Dataset : NHANES 2009-2012
- Preprocessing (Data cleaning) : Missing data (Statistical averages, Record deletion)
- Feature selection: LR (in RFE)

```
Column 'AgeDecade':
Number of unique values: 8
Unique values: [' 30-39' ' 0-9' ' 40-49' ' 60-69' ' 50-59' ' 10-19' ' 20-29' ' 70+']

Column 'Race1':
Number of unique values: 5
Unique values: ['White' 'Other' 'Mexican' 'Black' 'Hispanic']

Column 'Work':
Number of unique values: 3
Unique values: ['NotWorking' 'Working' 'Looking']

Column 'HealthGen':
Number of unique values: 5
Unique values: ['Good' 'Vgood' 'Fair' 'Excellent' 'Poor']

Column 'Depressed':
Number of unique values: 3
Unique values: ['Several' 'None' 'Most']

Column 'SleepTrouble':
Number of unique values: 2
Unique values: ['Yes' 'No']

Column 'SmokeNow':
Number of unique values: 2
Unique values: ['No' 'Yes']

Column 'HardDrugs':
Number of unique values: 2
Unique values: ['Yes' 'No']

Column 'SameSex':
Number of unique values: 2
Unique values: ['No' 'Yes']

Column 'SexOrientation':
Number of unique values: 3
Unique values: ['Heterosexual' 'Bisexual' 'Homosexual']

Column 'PregnantNow':
Number of unique values: 3
Unique values: ['No' 'Unknown' 'Yes']
```

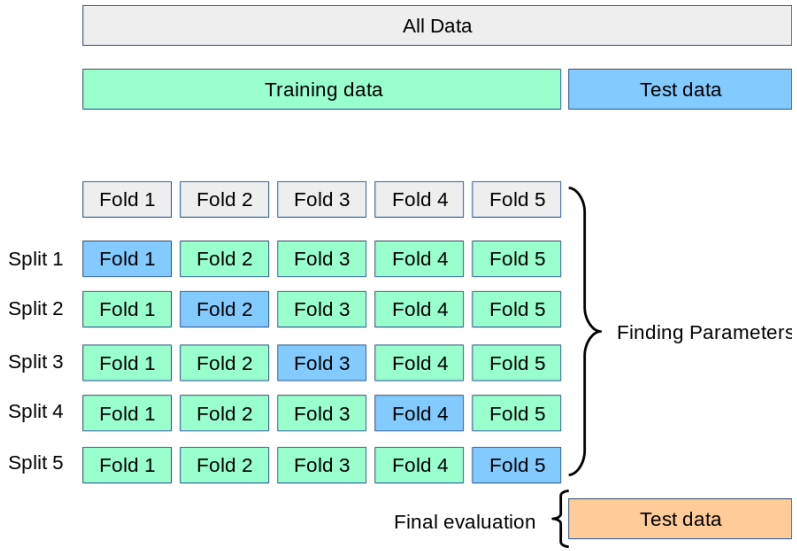
	count	mean	std	min	25%	50%	75%	max
Poverty	9858.0	2.810078	1.615772	0.00	1.32	2.810078	4.54	5.00
DirectChol	9858.0	1.365029	0.370114	0.39	1.11	1.365029	1.53	4.03
TotChol	9858.0	4.878875	0.996462	1.53	4.22	4.878875	5.40	13.65

Data and processing

- Overfitting and underfitting of the data



- Data splitting - Cross validation (Repeated K Fold)



Data and processing

- Encoded variables
- Resampling (SMOTE for Over sampling)
- Scaling

Visualization

	Poverty	DirectChol	TotChol
Poverty	1.000000	0.114370	0.078125
DirectChol	0.114370	1.000000	0.221467
TotChol	0.078125	0.221467	1.000000

Figure: Heatmap

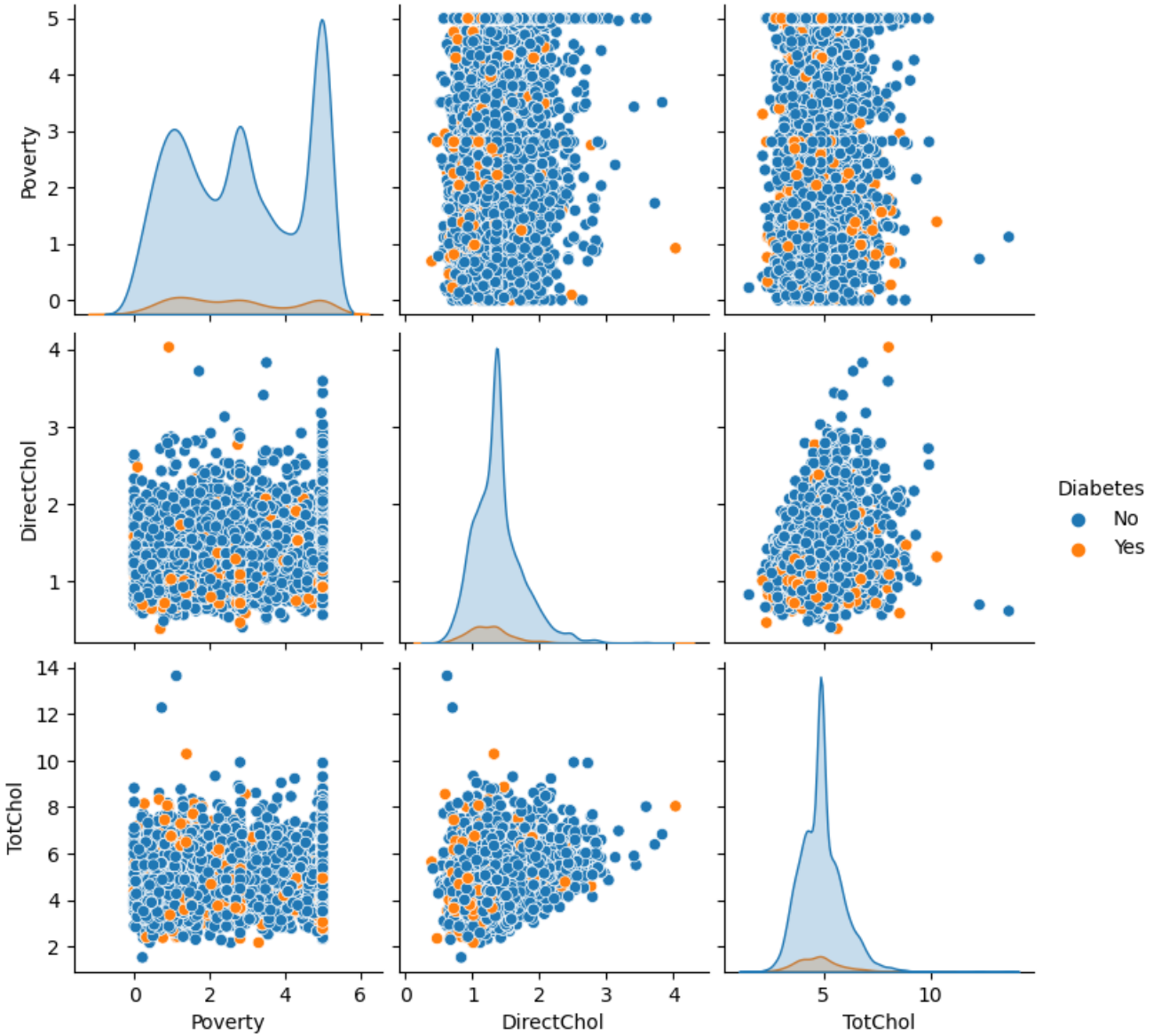


Figure: Pairplot (Not resampled)

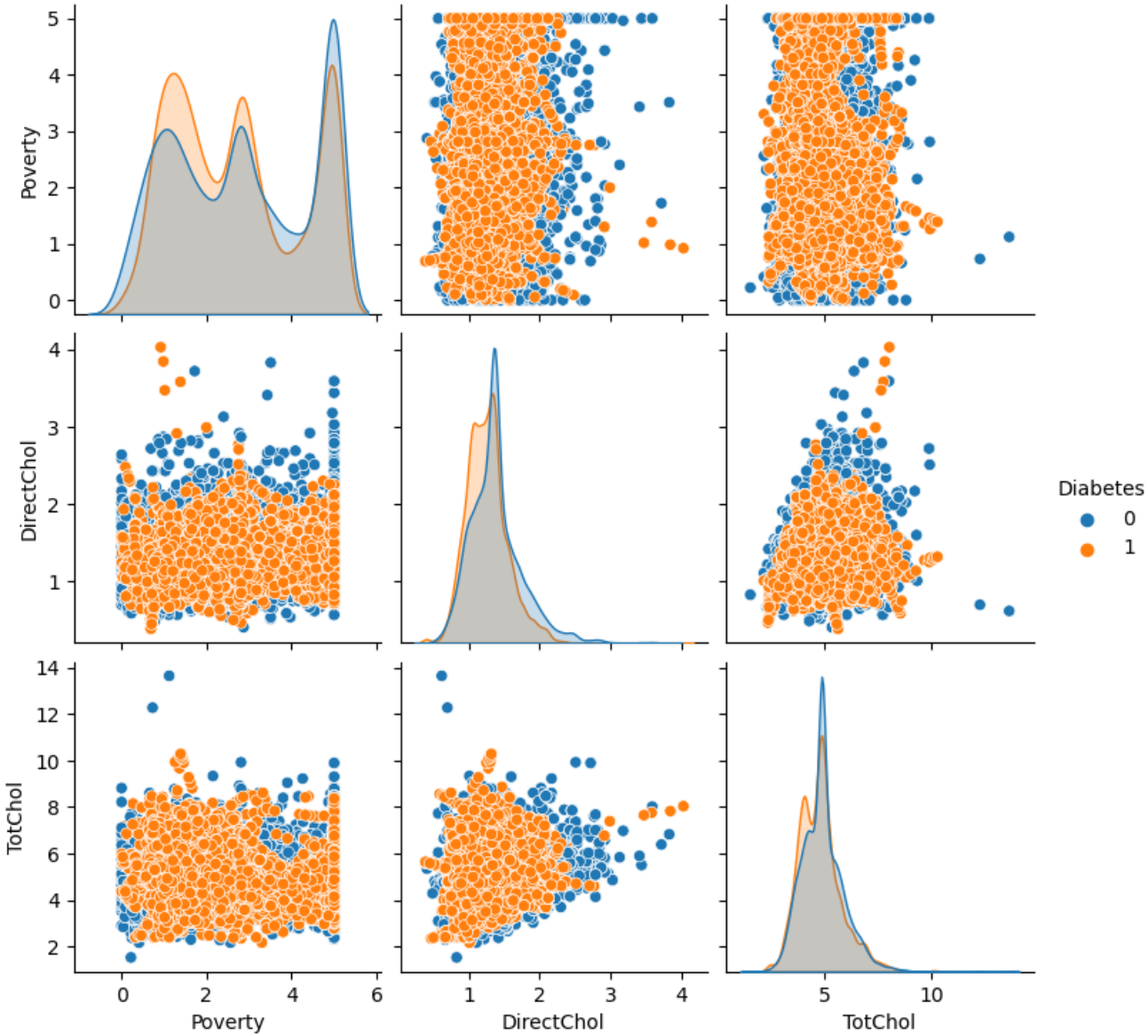


Figure: Pairplot (Resampled)

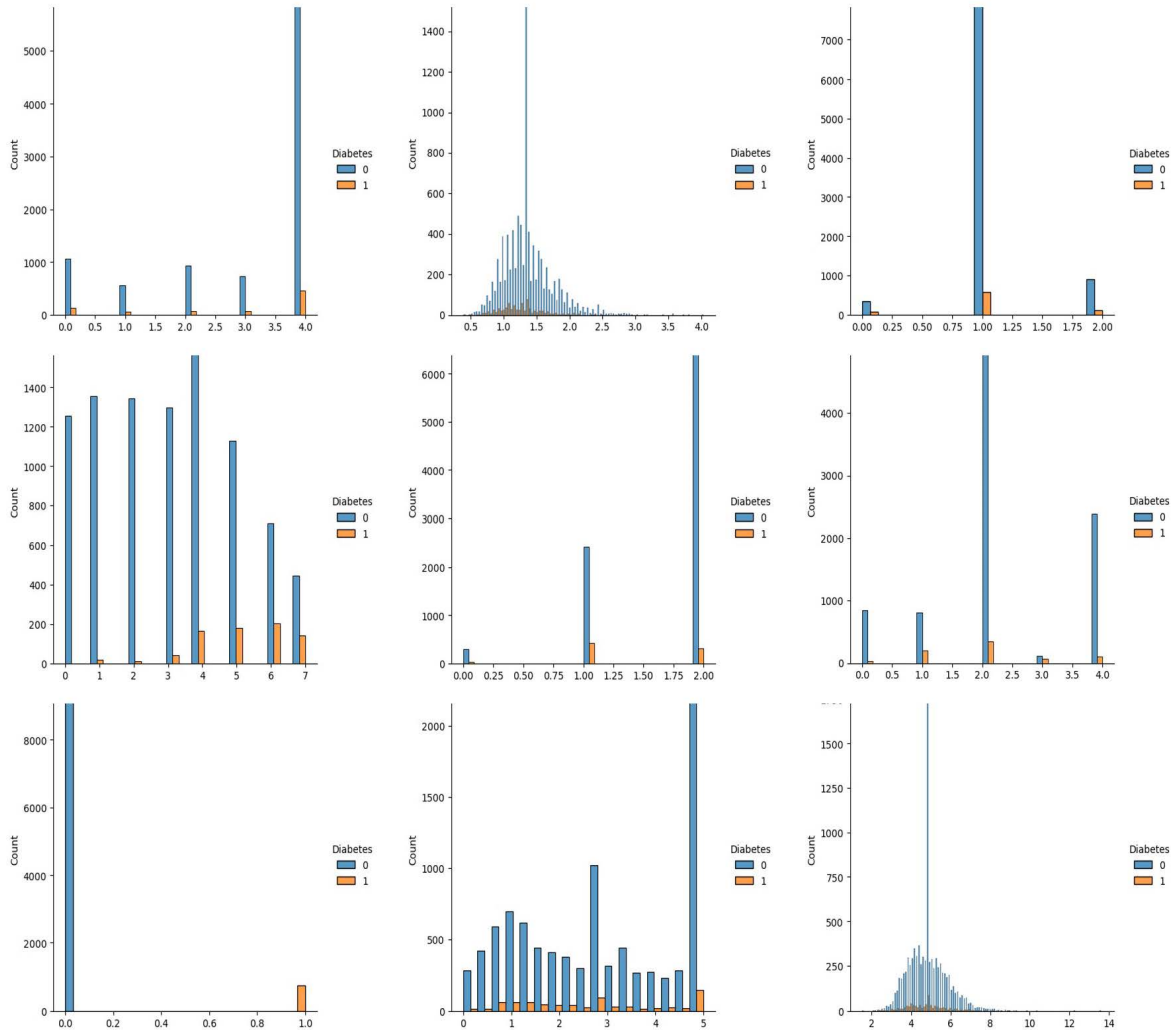


Figure: Bar chart (Not resampled)

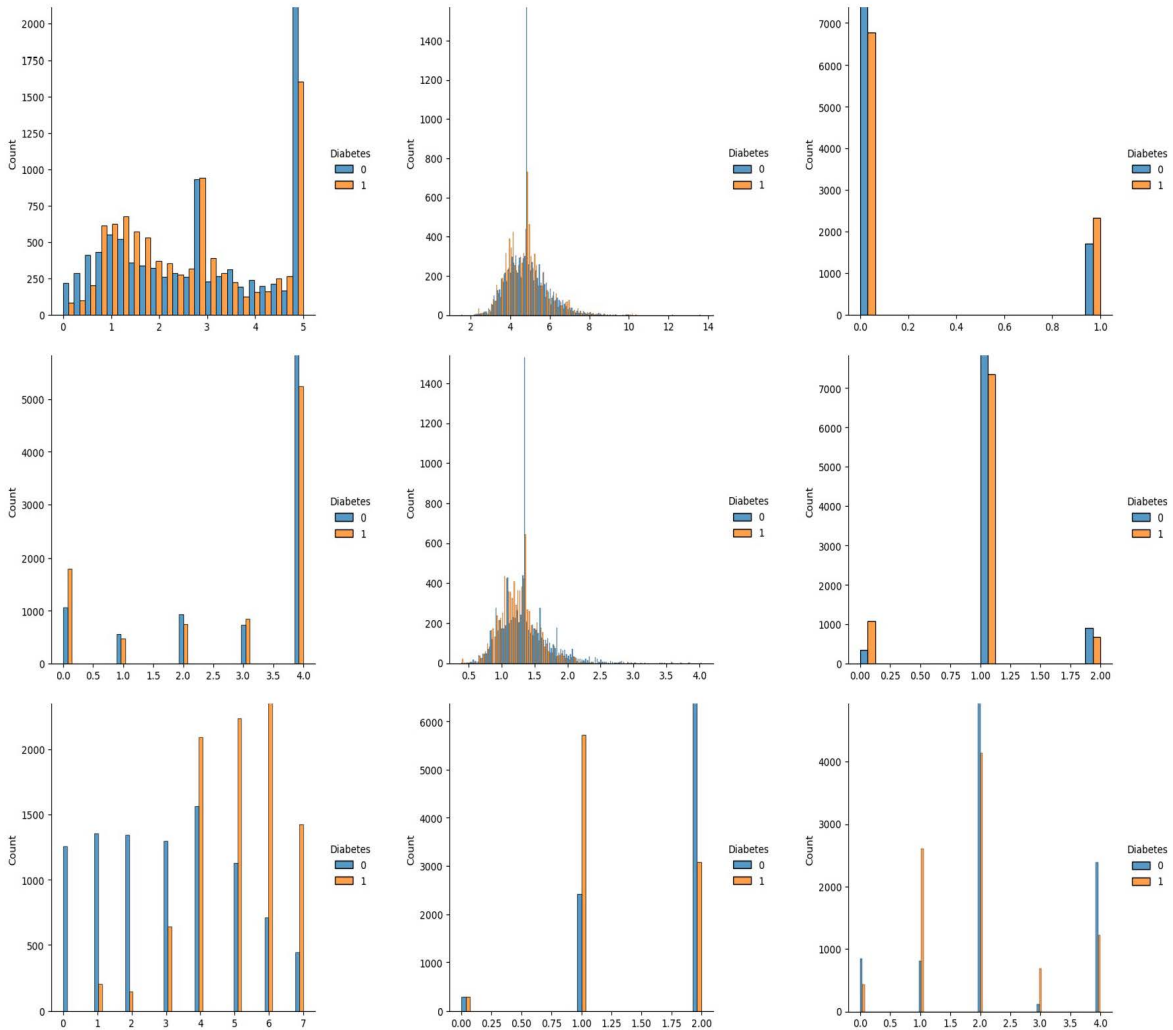


Figure: Bar chart (Resampled)

Data modeling

Algorithms

- Logistic regression for feature selection

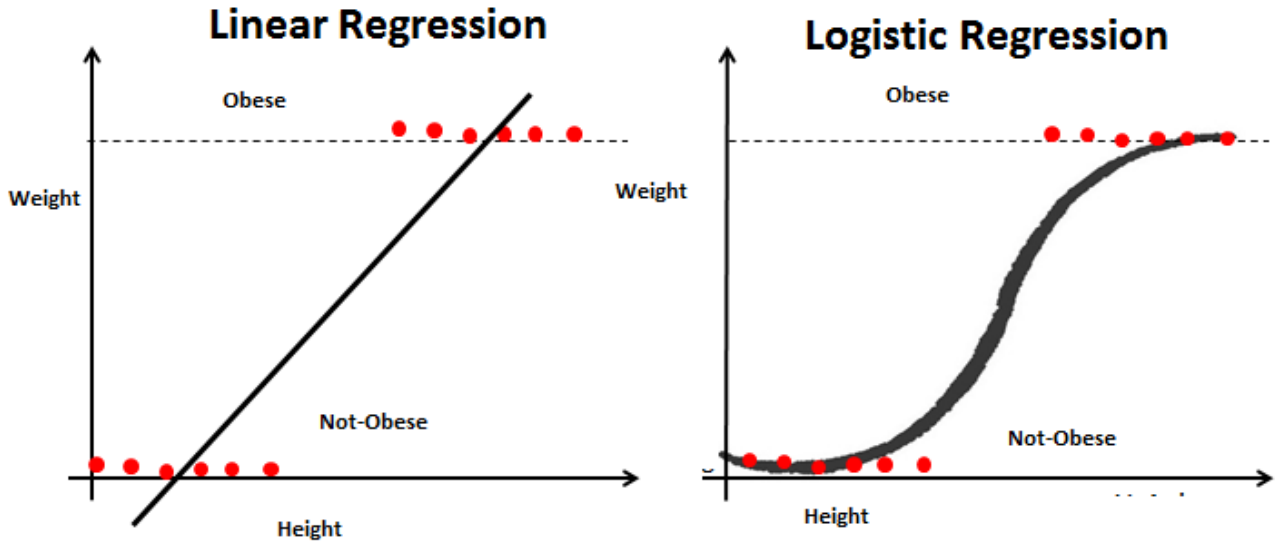


Figure: LR model

Algorithms

- Decision tree

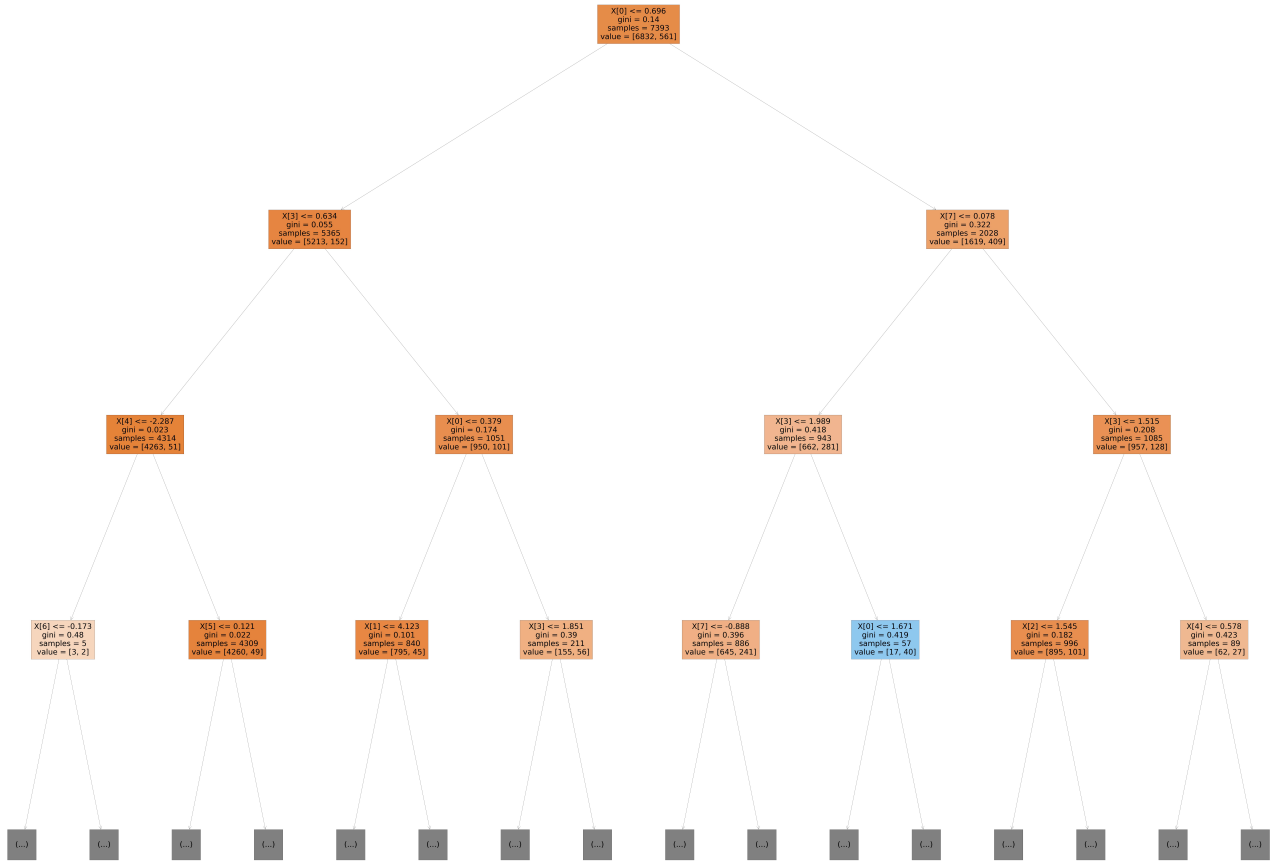


Figure: Decision tree

Algorithms

- Random forest

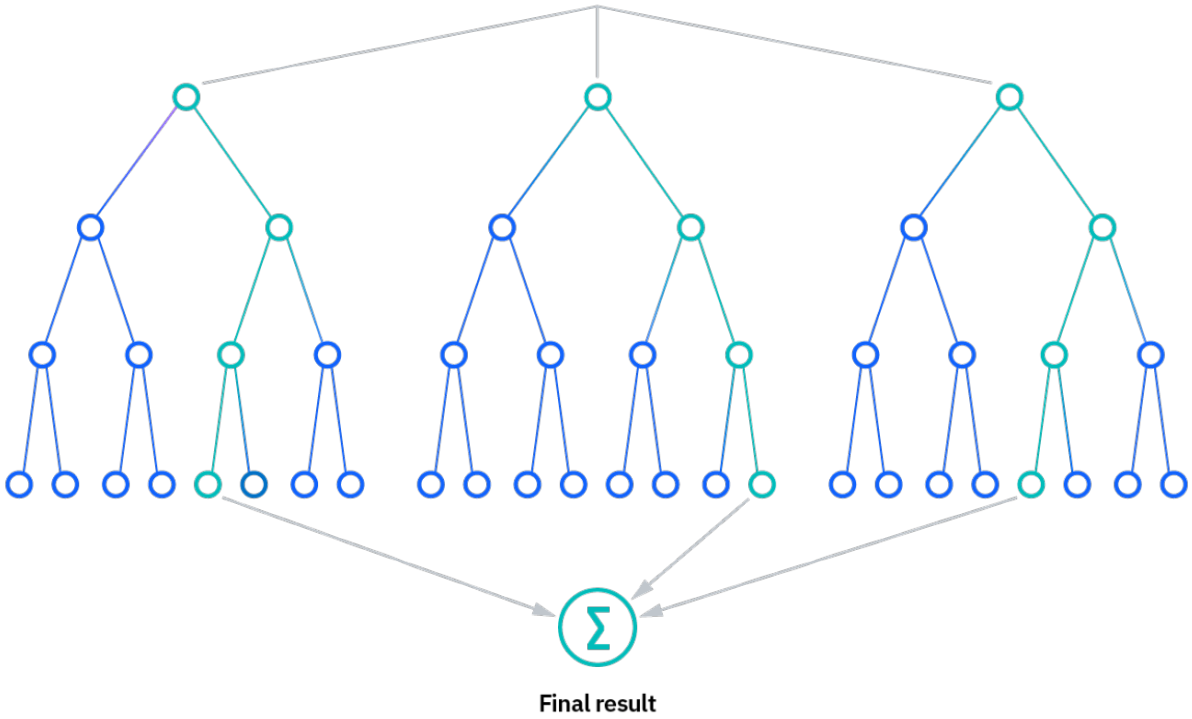


Figure: Random forest

Algorithms

- AdaBoost

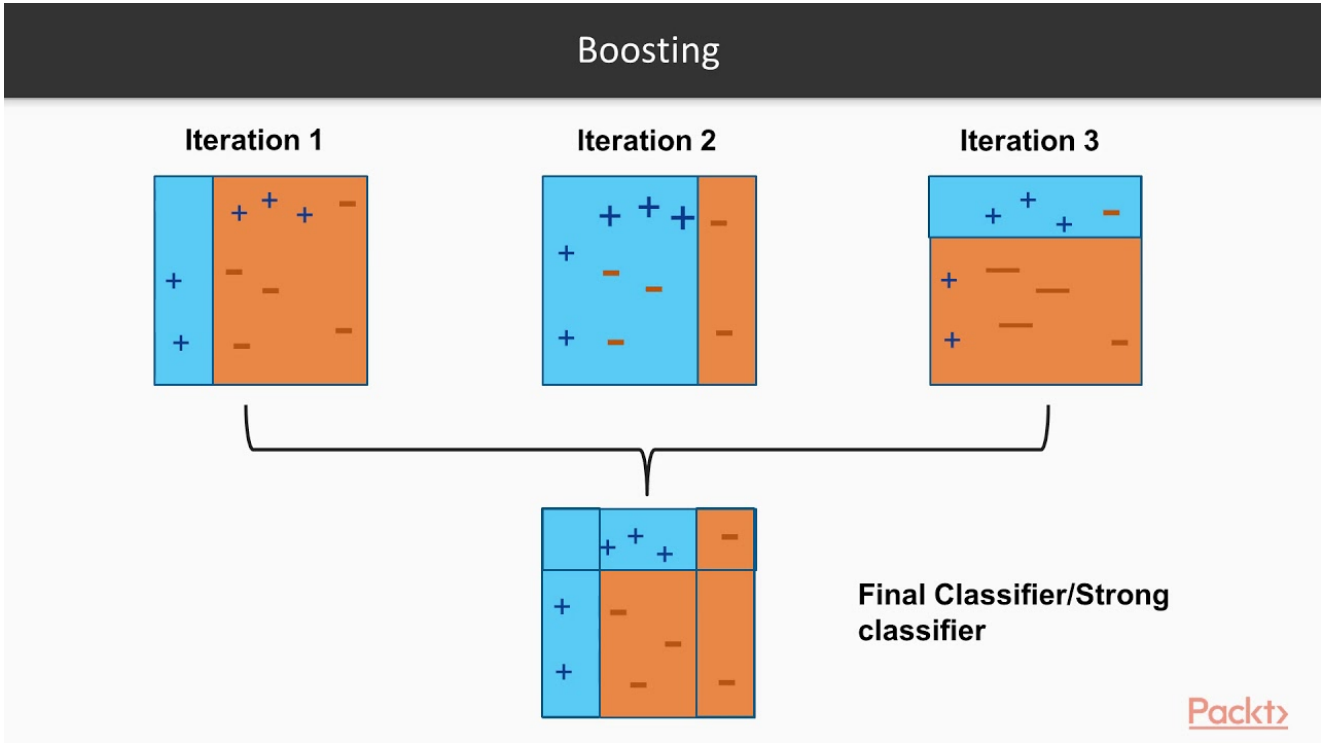


Figure: AdaBoost

Algorithms

- Naive Bayes

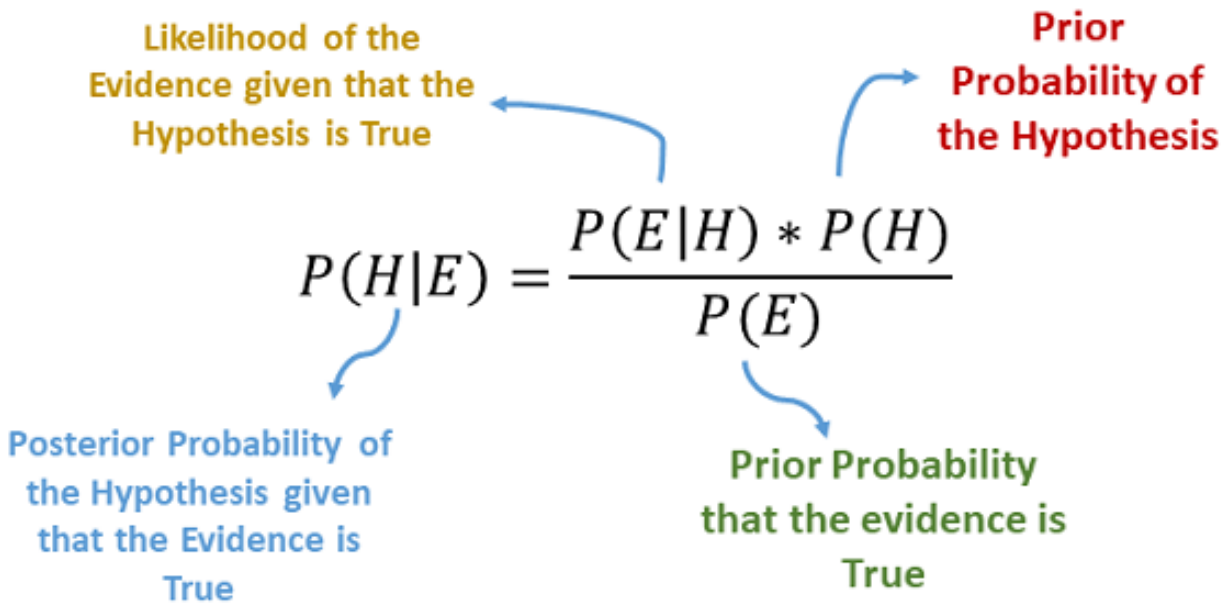
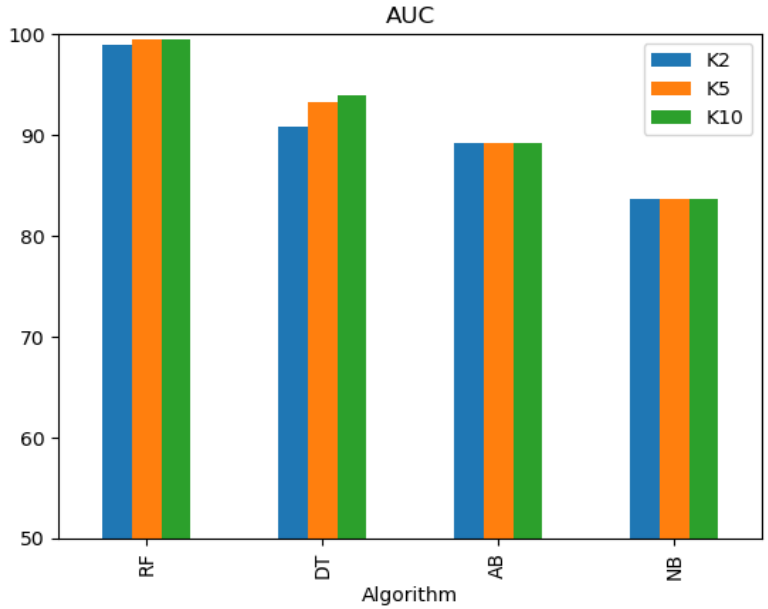


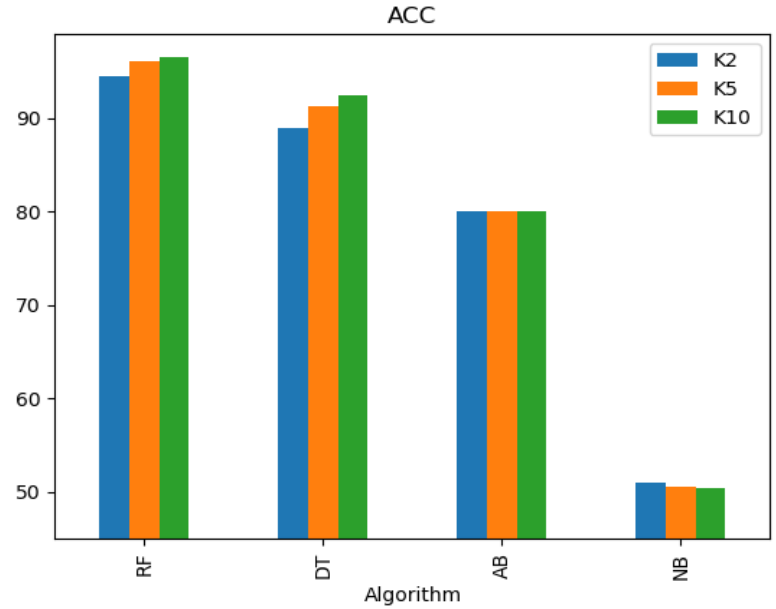
Figure: Naïve Bayes

Results

Performance measurement criteria

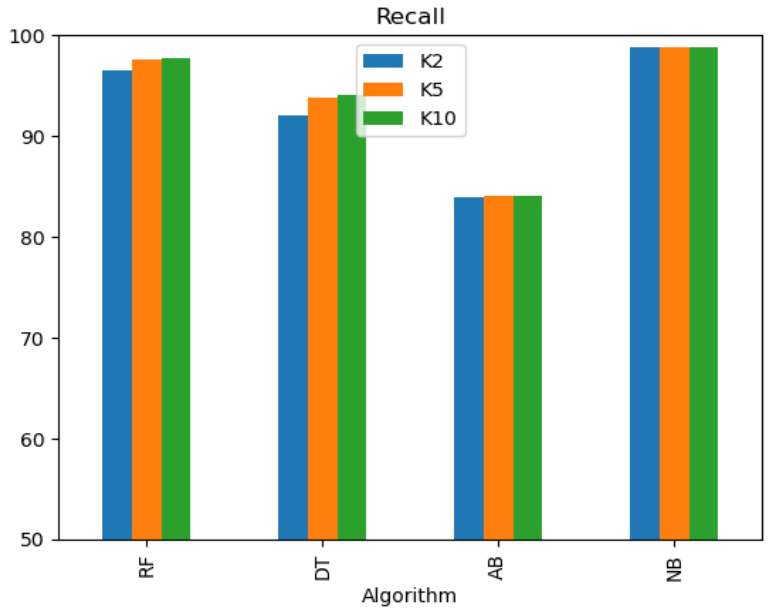
- Accuracy and AUC

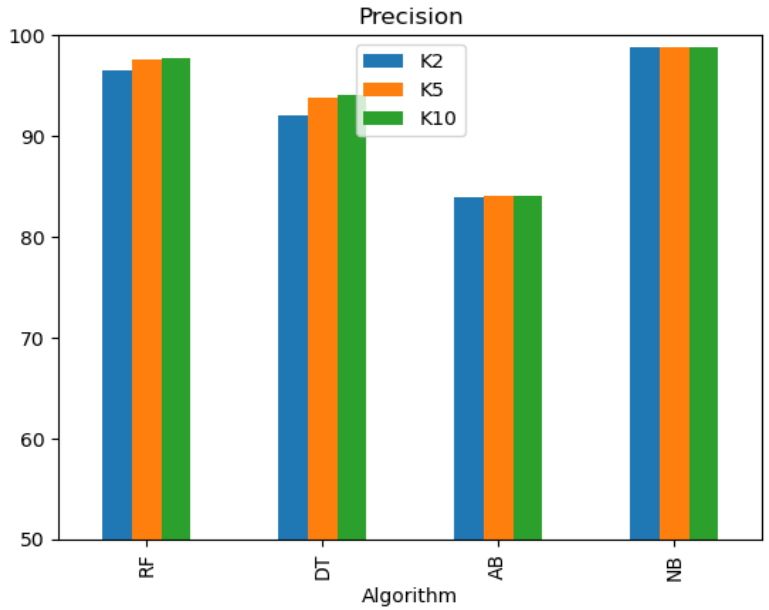




Performance measurement criteria

- Recall and Precision





Performance measurement criteria

- ROC

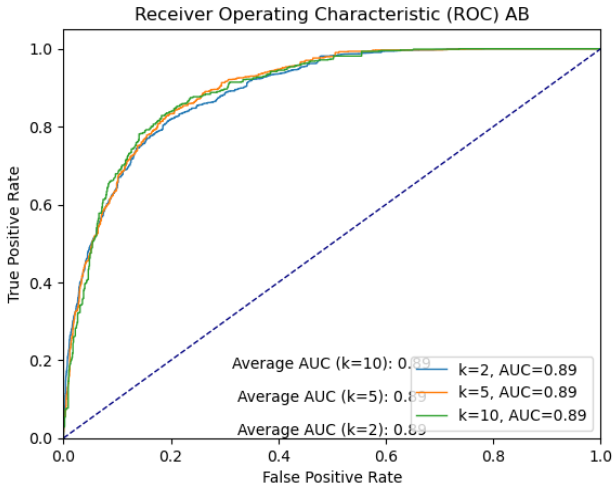


Figure: Ada

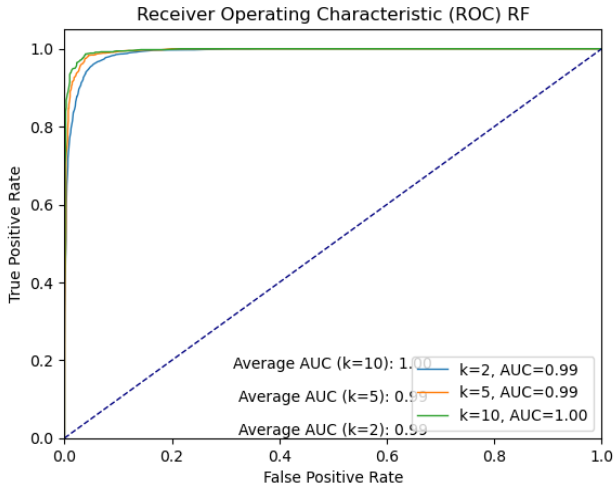


Figure: RF

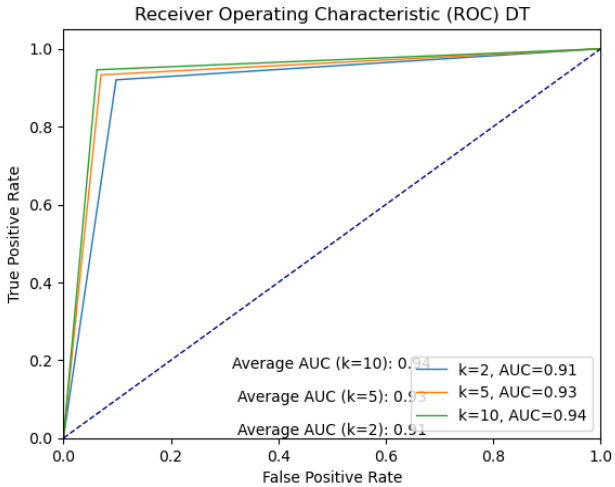


Figure: DT

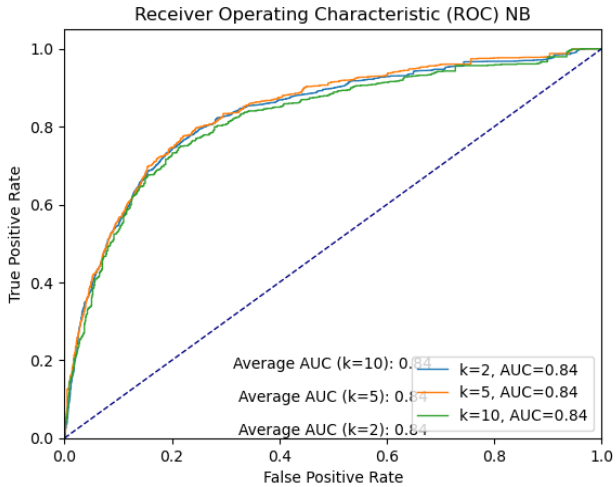
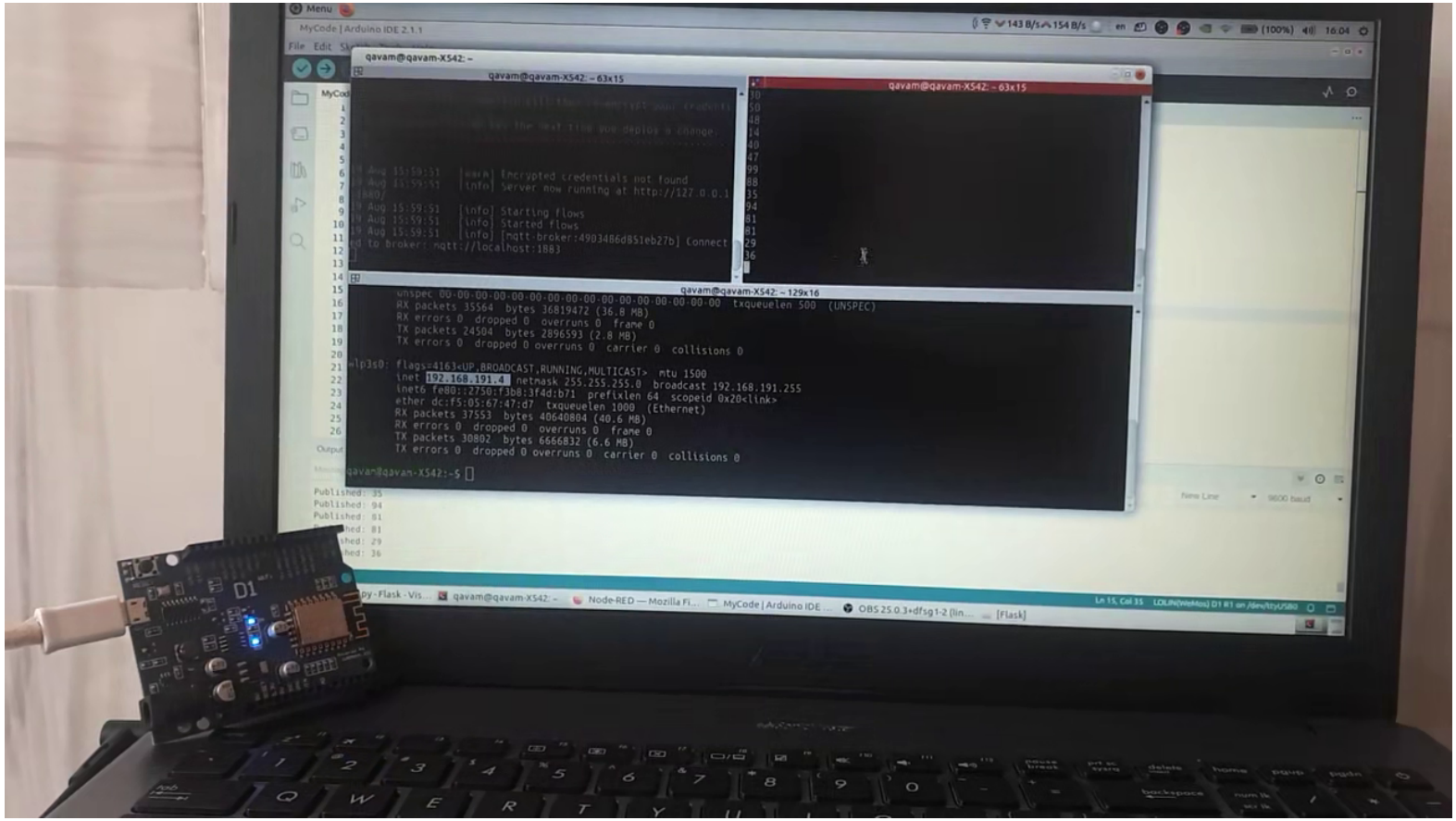


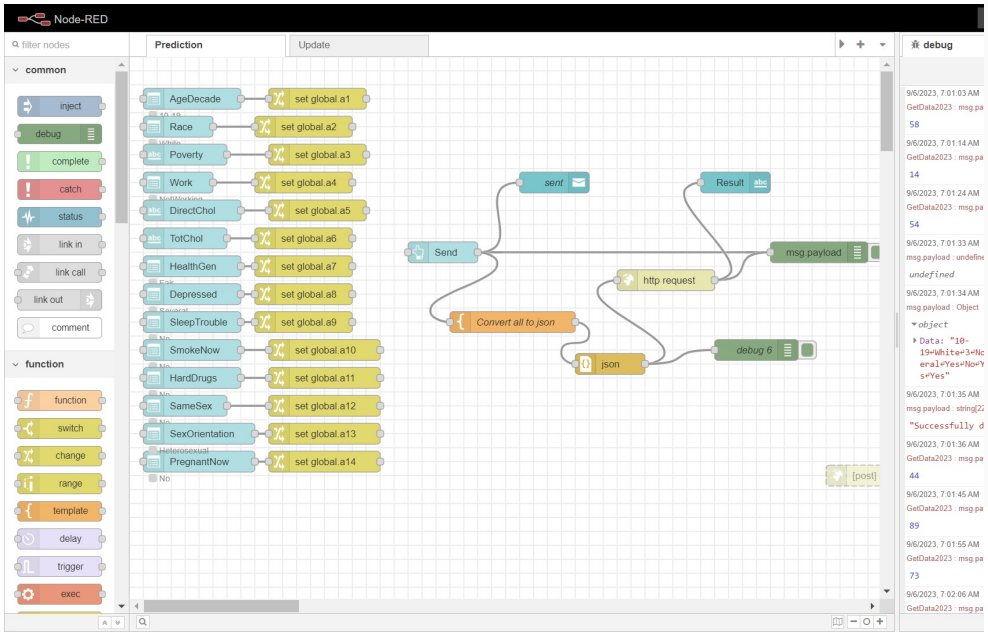
Figure: NB

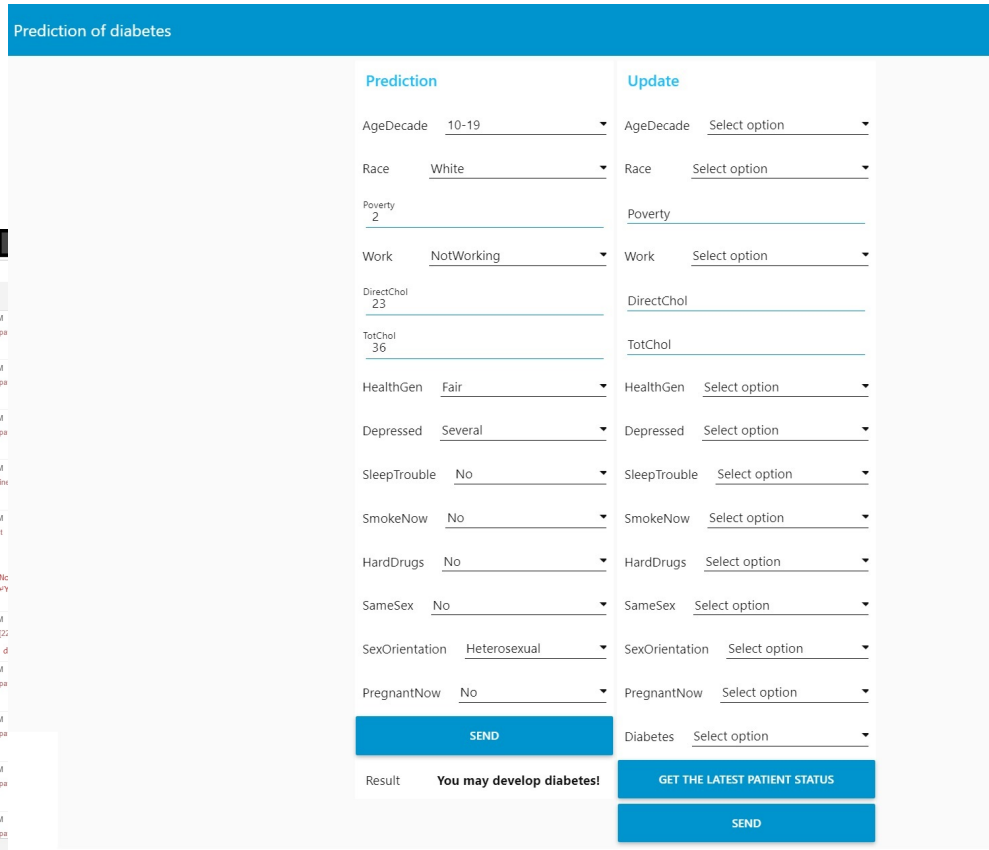
IOT

- Arduino


qavam@qavam-X542: ~
qavam@qavam-X542: ~ 63x15
qavam@qavam-X542: ~ 129x16
[info] Starting flows
[info] Started flows
RX packets 35564 bytes 36819472 (36.8 MB)
TX packets 24504 bytes 2896593 (2.8 MB)
inet 192.168.191.4 netmask 255.255.255.0 broadcast 192.168.191.255
RX packets 37553 bytes 40640804 (40.6 MB)
TX packets 30802 bytes 6666832 (6.6 MB)
D1

- Node-red





- Flask

Any Questions?