

Classification and prediction of diabetes disease using machine learning algorithms

Qavamodin Soleimani

Advisor: Dr. Fereshteh Dehghani

September 13, 2023

Contents

1 Introduction

2 Data mining methods

3 Visualization

4 Data modeling

5 Results

6 IOT

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.

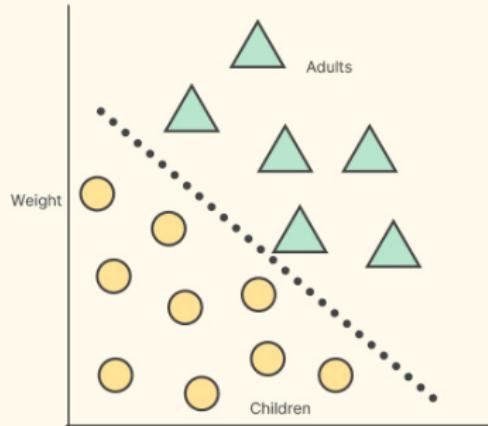
Predicting diabetes

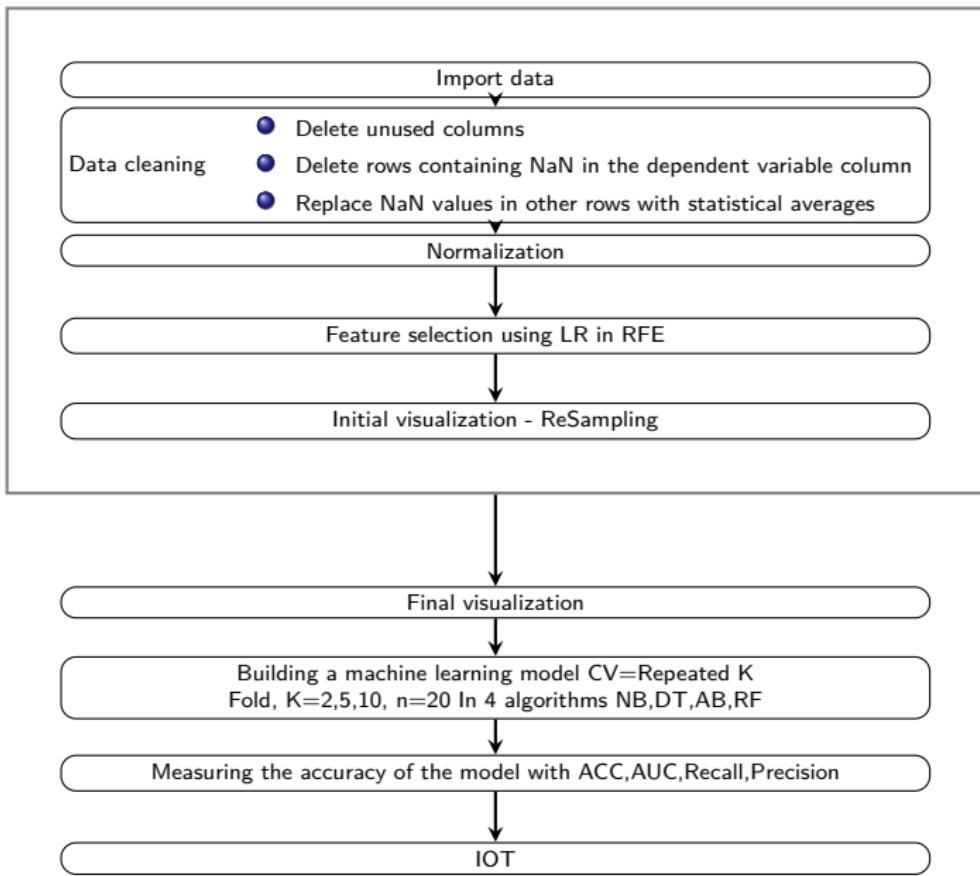
Scientists found that BMI, age, systolic and diastolic blood pressure, and a family history of diabetes were the most significant predictive features for prediabetes (Lama et al., 2021)

Data mining methods

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

Classification vs Clustering





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']

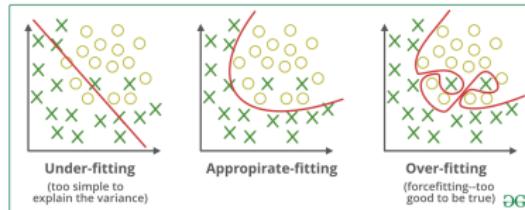
Column 'PregnantNow':
Number of unique values: 3
Unique values: ['No' 'Unknown' 'Yes']

```

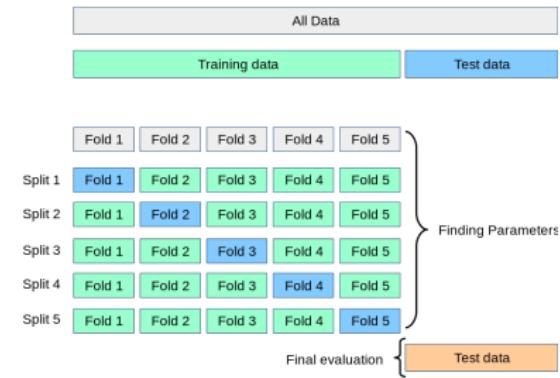
	count	mean	std	min	25%	50%	75%	max
Poverty	9858.0	2.810078	1.615772	0.00	1.32	2.810078	4.54	5.00
DirectChol	9858.0	1.365029	0.370114	0.39	1.11	1.365029	1.53	4.03
TotChol	9858.0	4.878875	0.996462	1.53	4.22	4.878875	5.40	13.65

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

	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

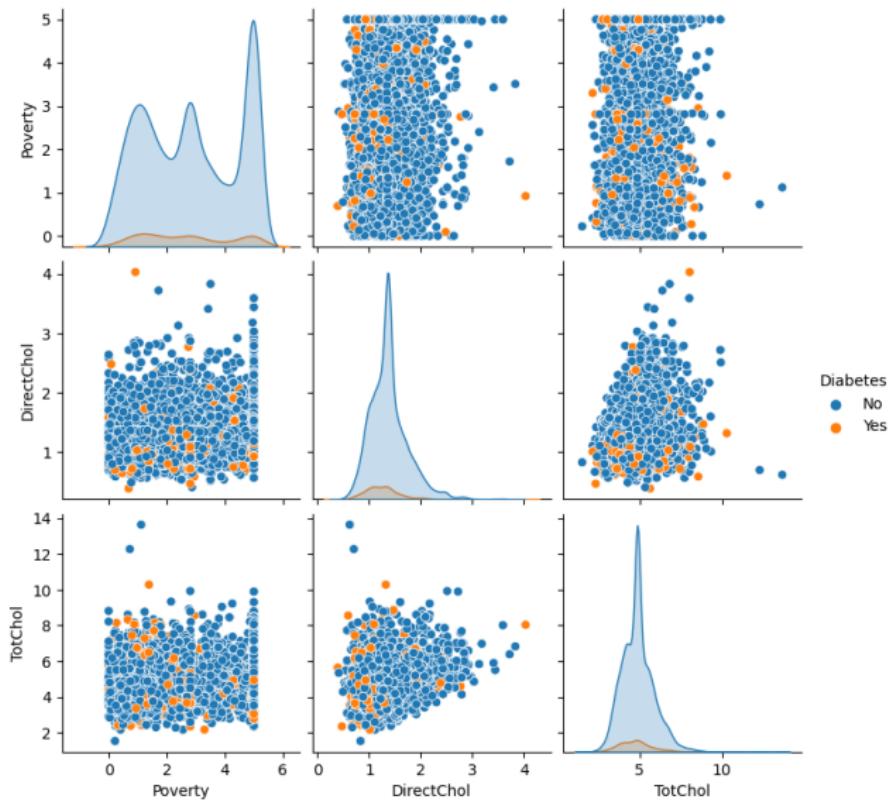


Figure: Pairplot (Not resampled)

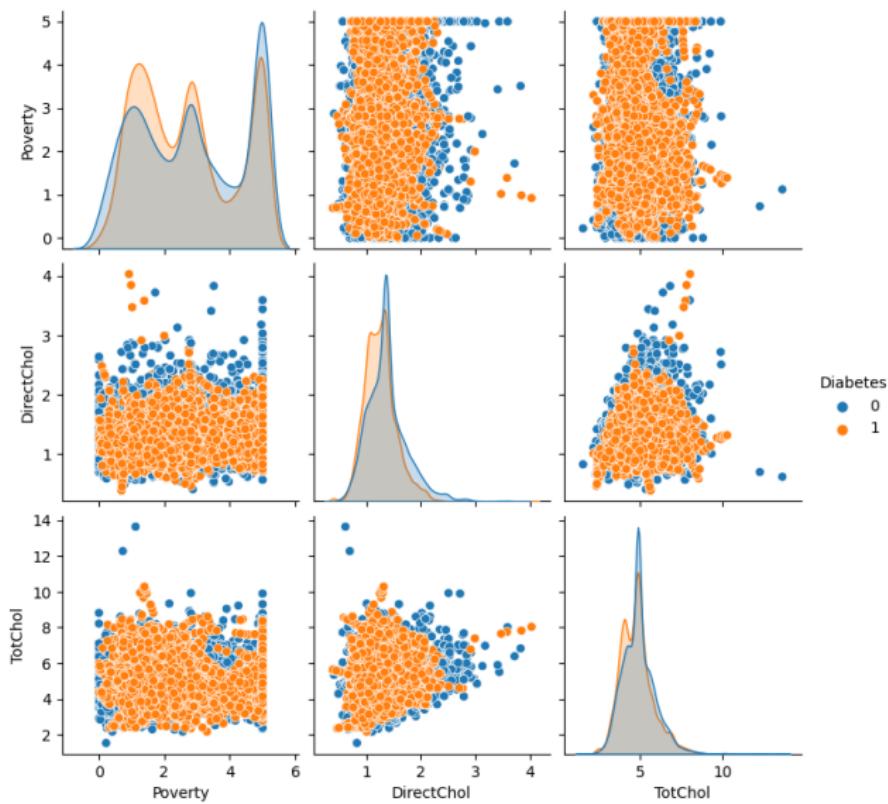


Figure: Pairplot (Resampled)

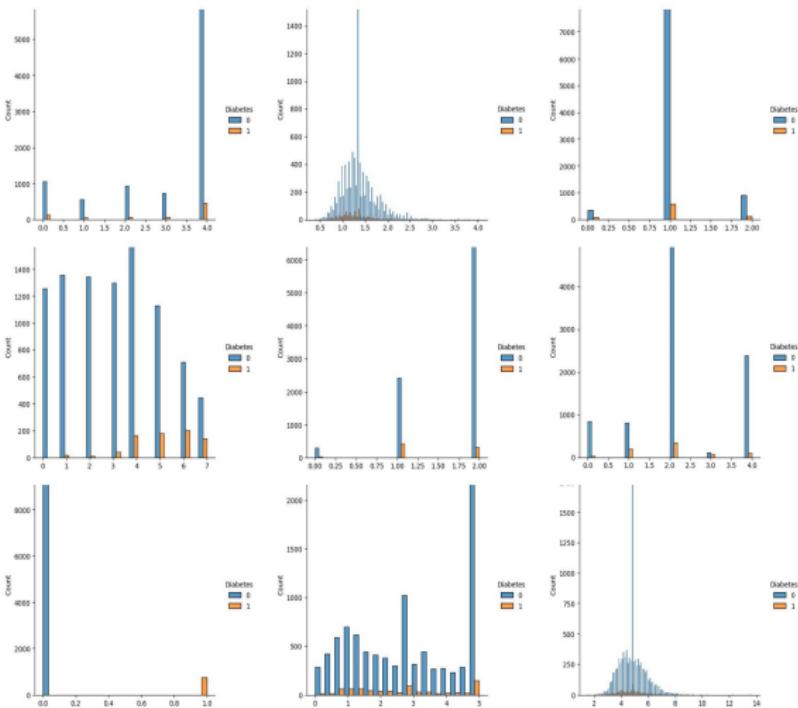


Figure: Bar chart (Not resampled)

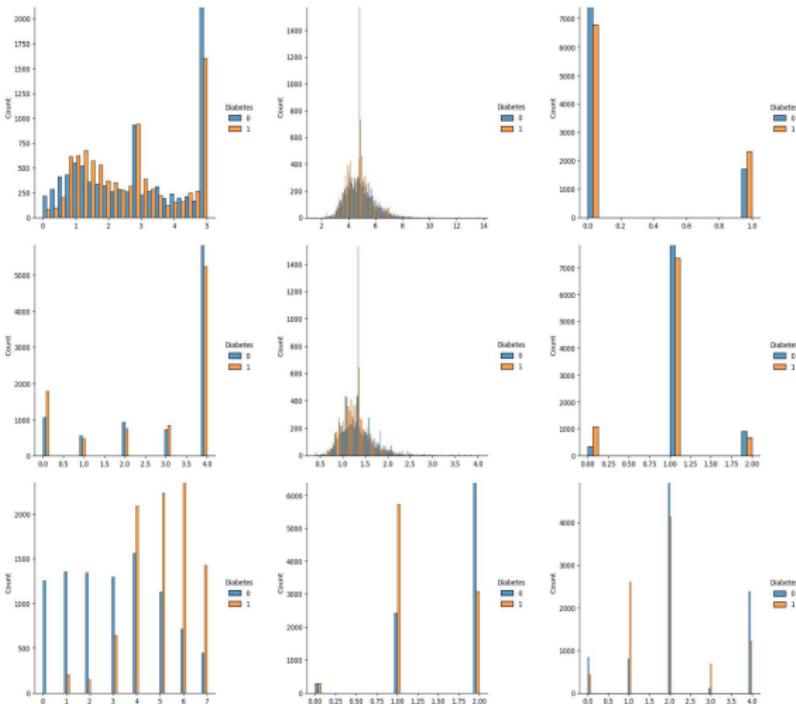


Figure: Bar chart (Resampled)

Data modeling

Algorithms

- Logistic regression for feature selection

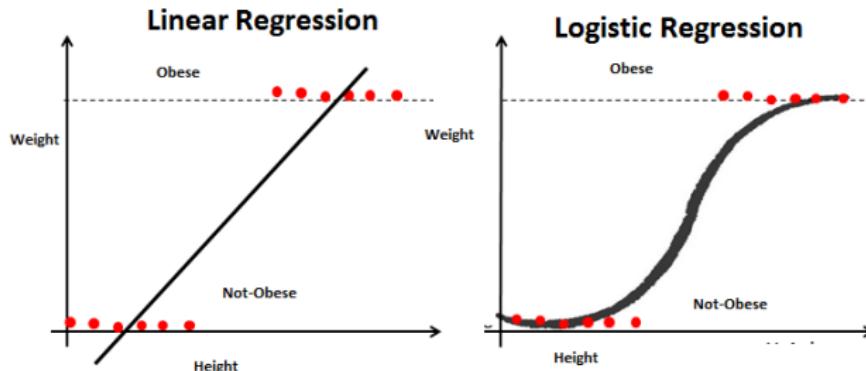


Figure: LR model

Algorithms

- Decision tree

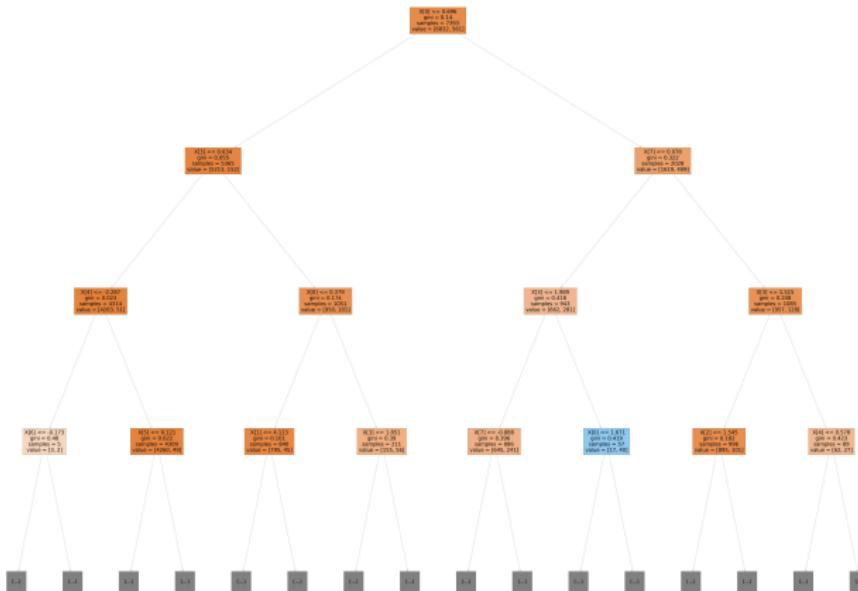


Figure: Decision tree

Algorithms

- Random forest

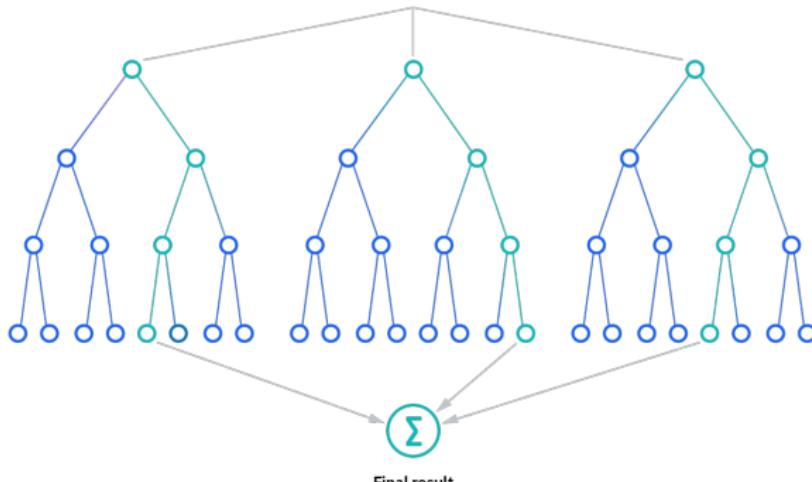


Figure: Random forest

Algorithms

- AdaBoost

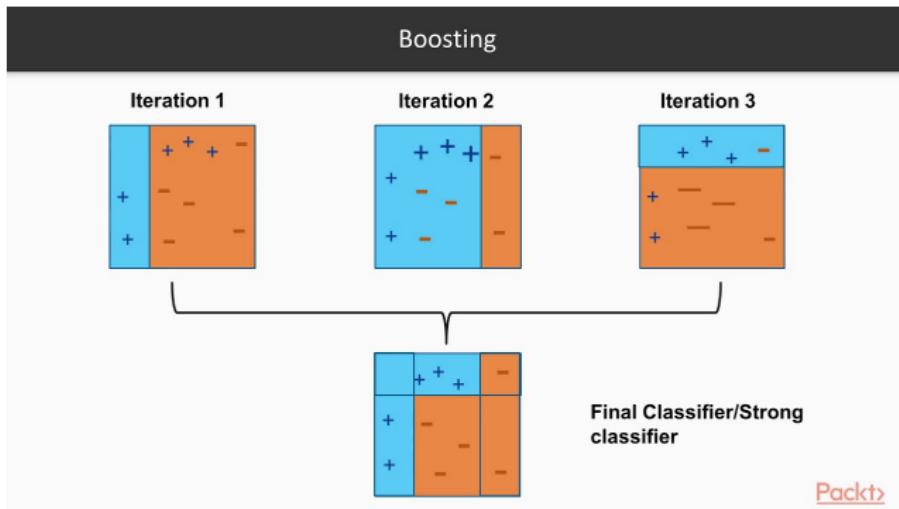


Figure: AdaBoost

Algorithms

- Naive Bayes

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True

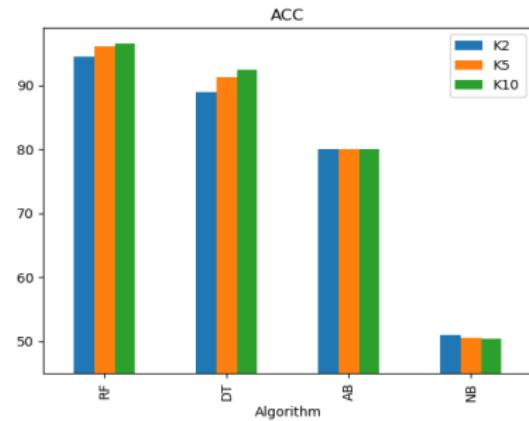
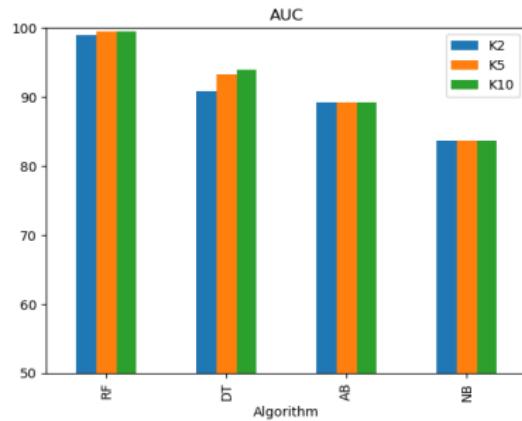
The diagram illustrates the Naive Bayes formula with colored annotations. The formula is $P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$. The term $P(E|H)$ is labeled 'Likelihood of the Evidence given that the Hypothesis is True' in orange. The term $P(H)$ is labeled 'Prior Probability of the Hypothesis' in red. The term $P(E)$ is labeled 'Prior Probability that the evidence is True' in green. The term $P(H|E)$ is labeled 'Posterior Probability of the Hypothesis given that the Evidence is True' in blue.

Figure: Naïve Bayes

Results

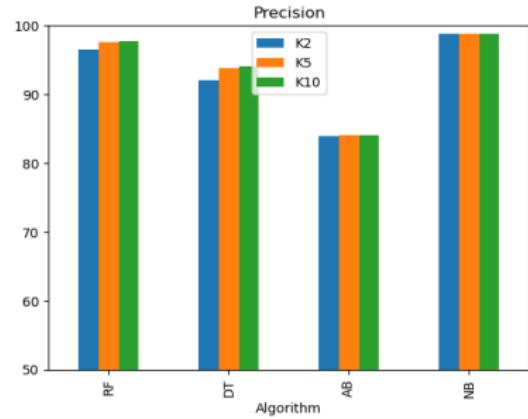
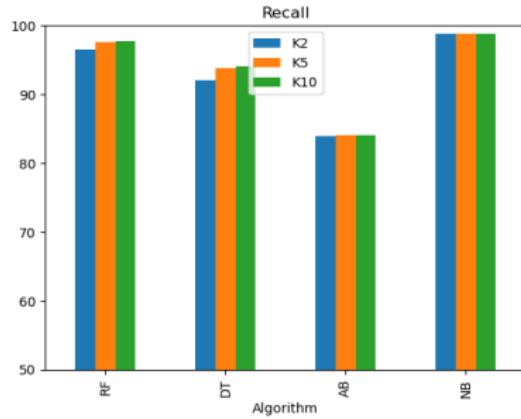
Performance measurement criteria

- Accuracy and AUC



Performance measurement criteria

- Recall and Precision



Performance measurement criteria

- ROC

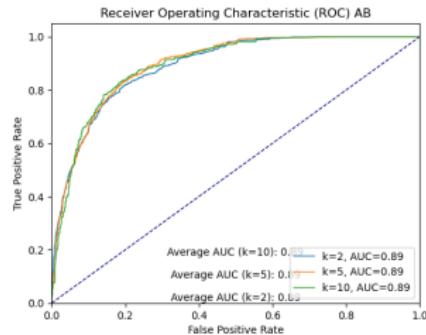


Figure: Ada

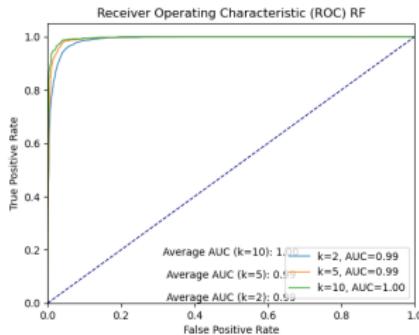


Figure: RF

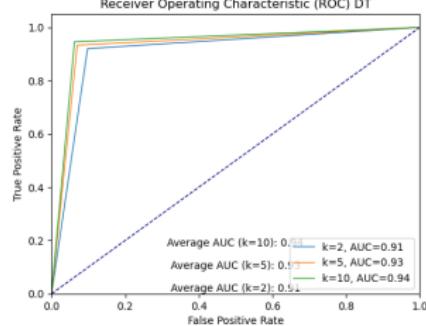


Figure: DT

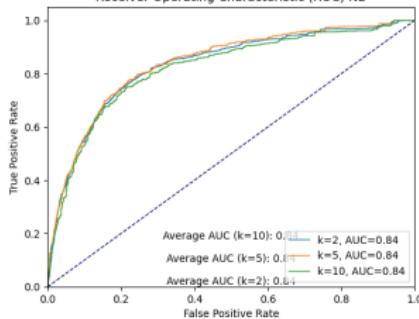
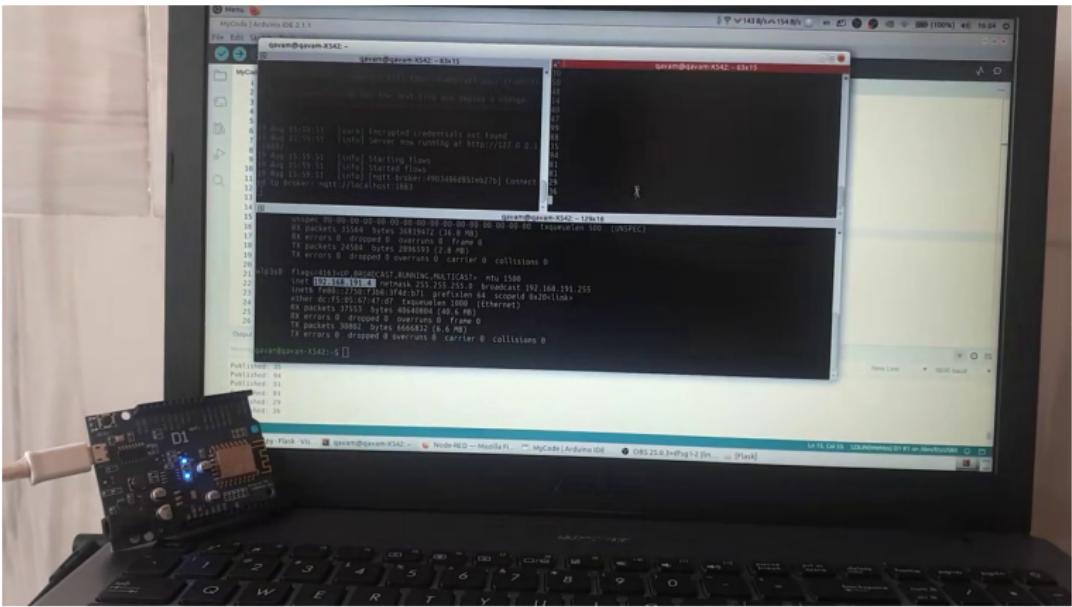


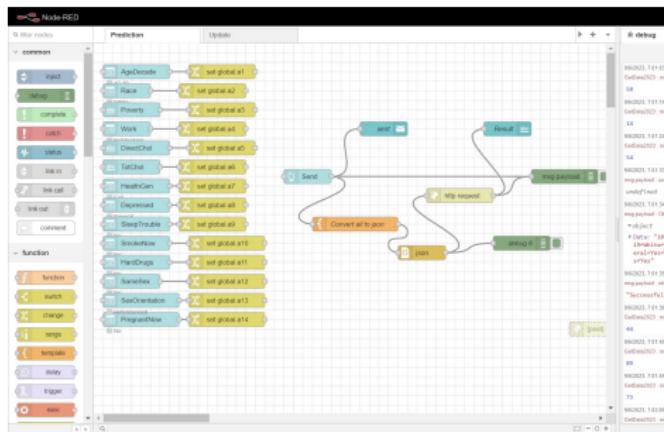
Figure: NB

IOT

• Arduino



- Node-red



Prediction of diabetes

Prediction		Update	
AgeDecade	10-19	AgeDecade	Select option
Race	White	Race	Select option
Poverty	Z	Poverty	
Work	NotWorking	Work	Select option
DirectChol	23	DirectChol	
TotChol	36	TotChol	
HealthGen	Fair	HealthGen	Select option
Depressed	Several	Depressed	Select option
SleepTrouble	No	SleepTrouble	Select option
SmokeNow	No	SmokeNow	Select option
HardDrugs	No	HardDrugs	Select option
SameSex	No	SameSex	Select option
SexOrientation	Heterosexual	SexOrientation	Select option
PregnantNow	No	PregnantNow	Select option
SEND		Diabetes Select option	
Result: You may develop diabetes!		GET THE LATEST PATIENT STATUS	
		SEND	

- Flask



Any Questions?