# Diabetes prediction system via ML algorithms

Qavamodin Soleimani

Advisor: Dr. Fereshteh Dehghani

August 30, 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?

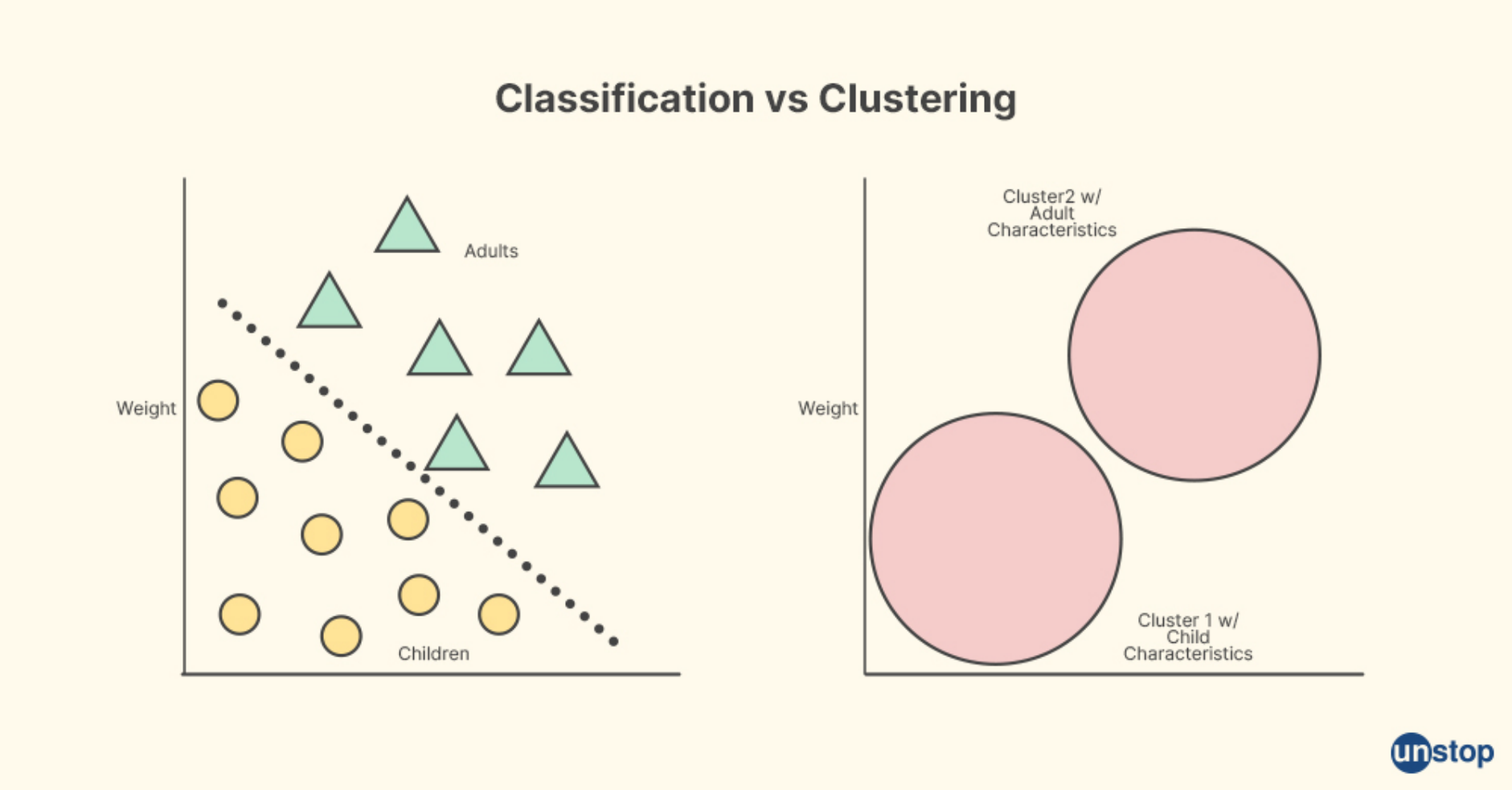
Methods of preventing diabetes.

## 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)

# Data mining methods

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



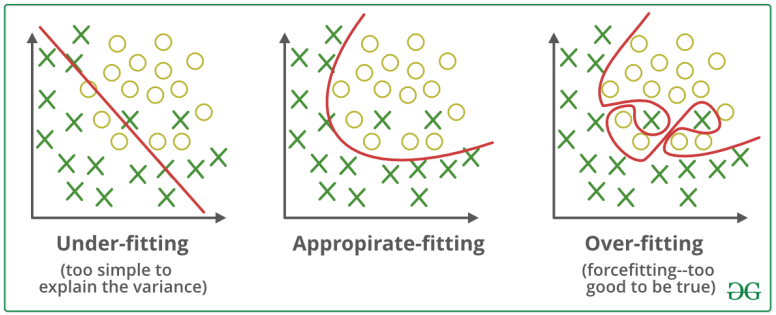
# Data and processing

- Dataset : NHANES 2009-2012
- Preprocessing (Data cleaning) : Missing data (Statistical averages, Record deletion)

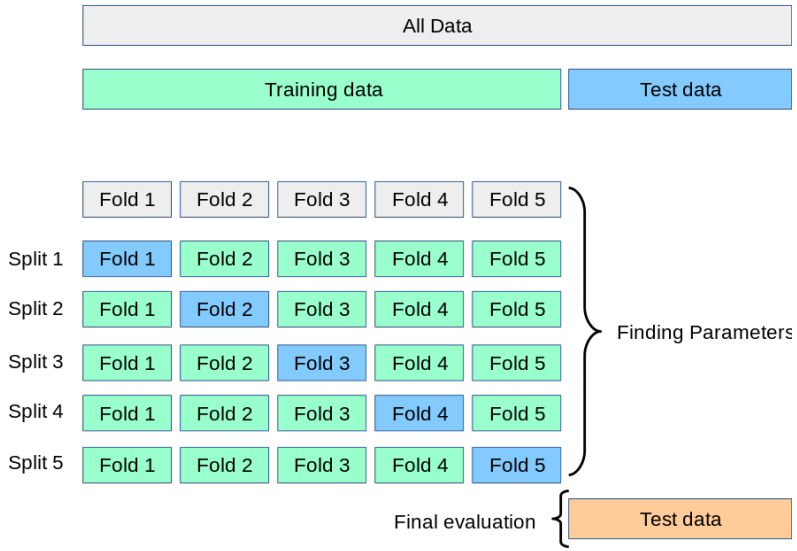
| # | Column |
|---|---|
| 0 | Gender |
| 1 | Age |
| 2 | Race1 |
| 3 | Education |
| 4 | MaritalStatus |
| 5 | Work |
| 6 | Weight |
| 7 | Height |
| 8 | BMI |
| 9 | BPSysAve |
| 10 | BPDiaAve |
| 11 | DirectChol |
| 12 | TotChol |
| 13 | PhysActive |
| 14 | Diabetes |

# Data and processing

- Overfitting and underfitting of the data




- Data splitting - Cross validation (Repeated K Fold)



# Data and processing

- Dummy variables
- Scaling

# Visualization

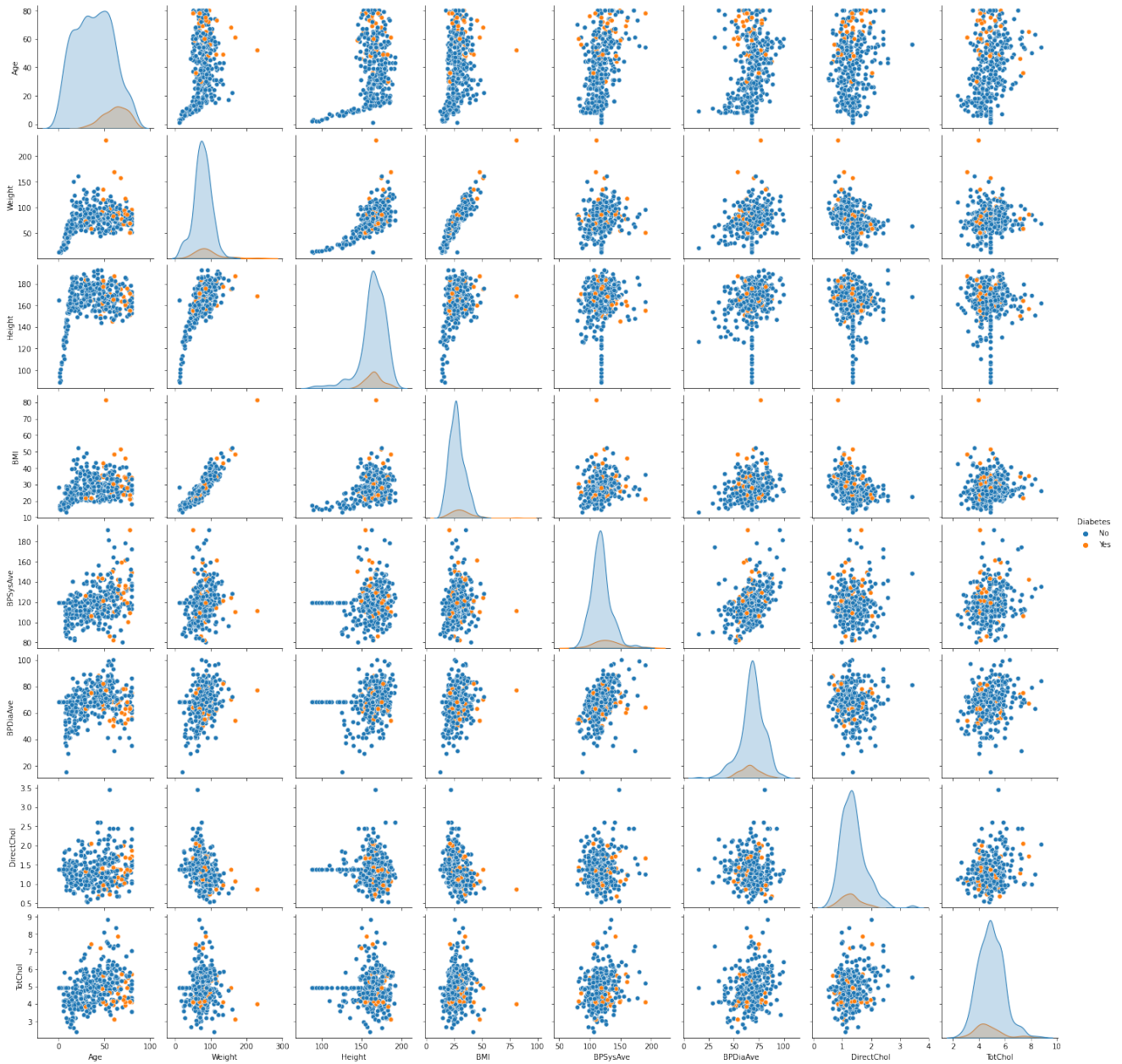



Figure: Pairplot

| | Age | Weight | Height | BMI | BPSysAve | BPDiaAve | DirectChol | TotChol |
|---|---|---|---|---|---|---|---|---|
| Age | 1.000000 | 0.485982 | 0.448773 | 0.396693 | 0.436330 | 0.192314 | 0.087136 | 0.279955 |
| Weight | 0.485982 | 1.000000 | 0.722345 | 0.870981 | 0.211238 | 0.239157 | -0.256111 | 0.109253 |
| Height | 0.448773 | 0.722345 | 1.000000 | 0.434615 | 0.099839 | 0.156478 | -0.091657 | 0.056475 |
| BMI | 0.396693 | 0.870981 | 0.434615 | 1.000000 | 0.231158 | 0.213611 | -0.268621 | 0.130599 |
| BPSysAve | 0.436330 | 0.211238 | 0.099839 | 0.231158 | 1.000000 | 0.426362 | 0.004474 | 0.202014 |
| BPDiaAve | 0.192314 | 0.239157 | 0.156478 | 0.213611 | 0.426362 | 1.000000 | -0.019679 | 0.250050 |
| DirectChol | 0.087136 | -0.256111 | -0.091657 | -0.268621 | 0.004474 | -0.019679 | 1.000000 | 0.221467 |
| TotChol | 0.279955 | 0.109253 | 0.056475 | 0.130599 | 0.202014 | 0.250050 | 0.221467 | 1.000000 |

Figure: Heatmap

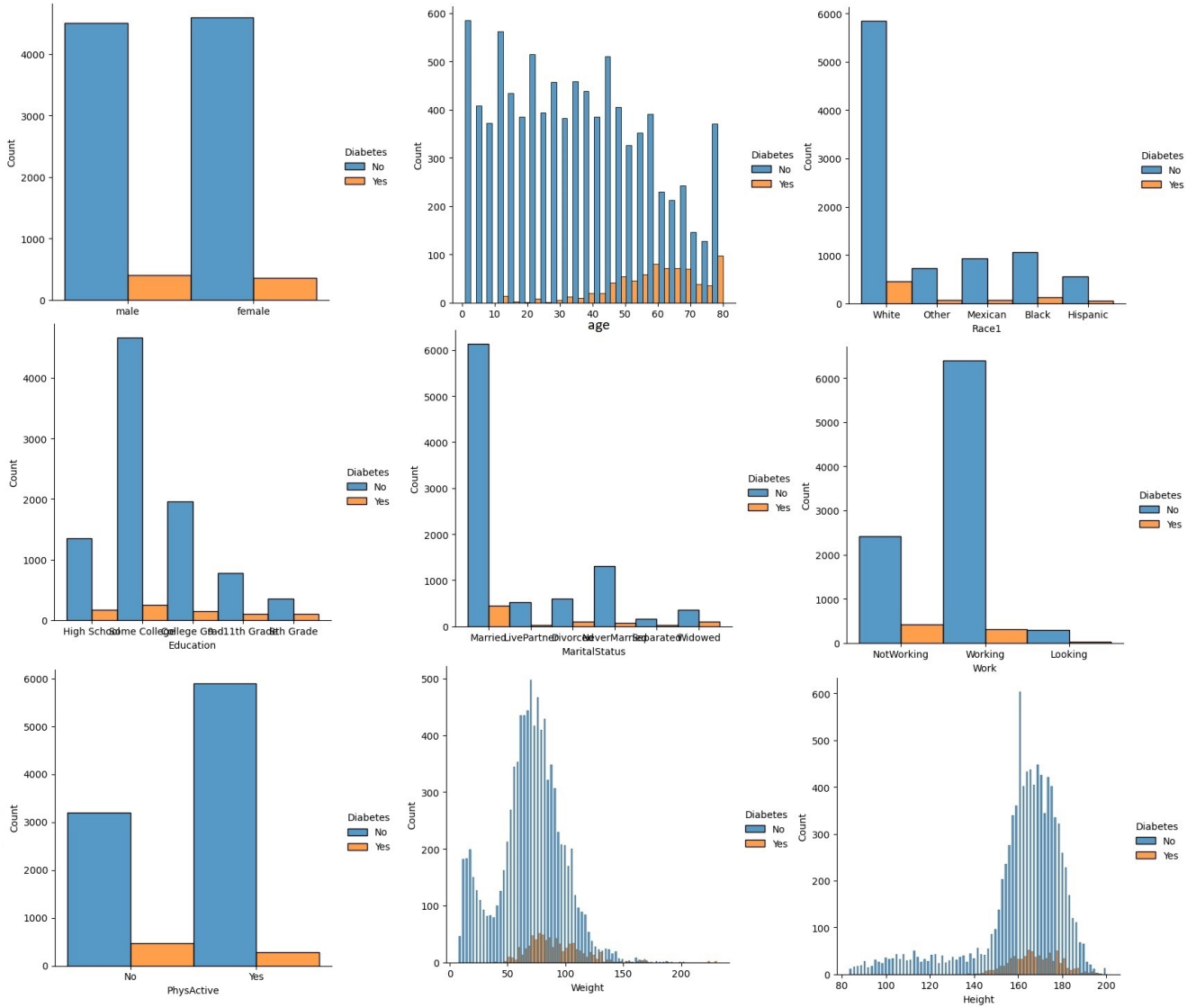



Figure: Bar chart

# Data modeling

# Algorithms

- Logistic regression

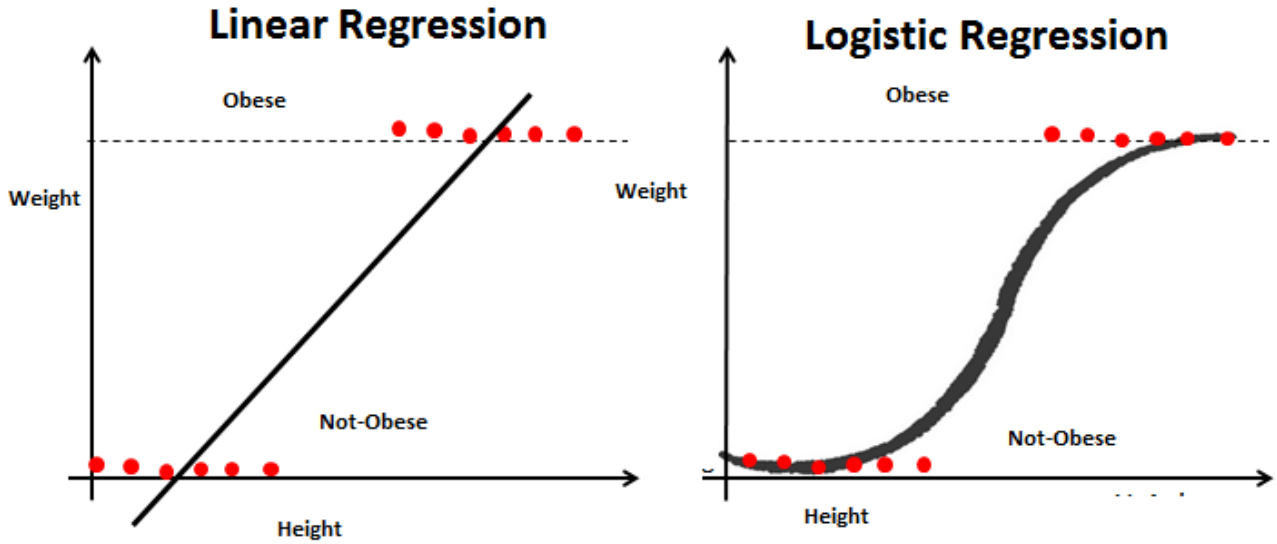



Figure: LR model

# Algorithms

- Decision tree

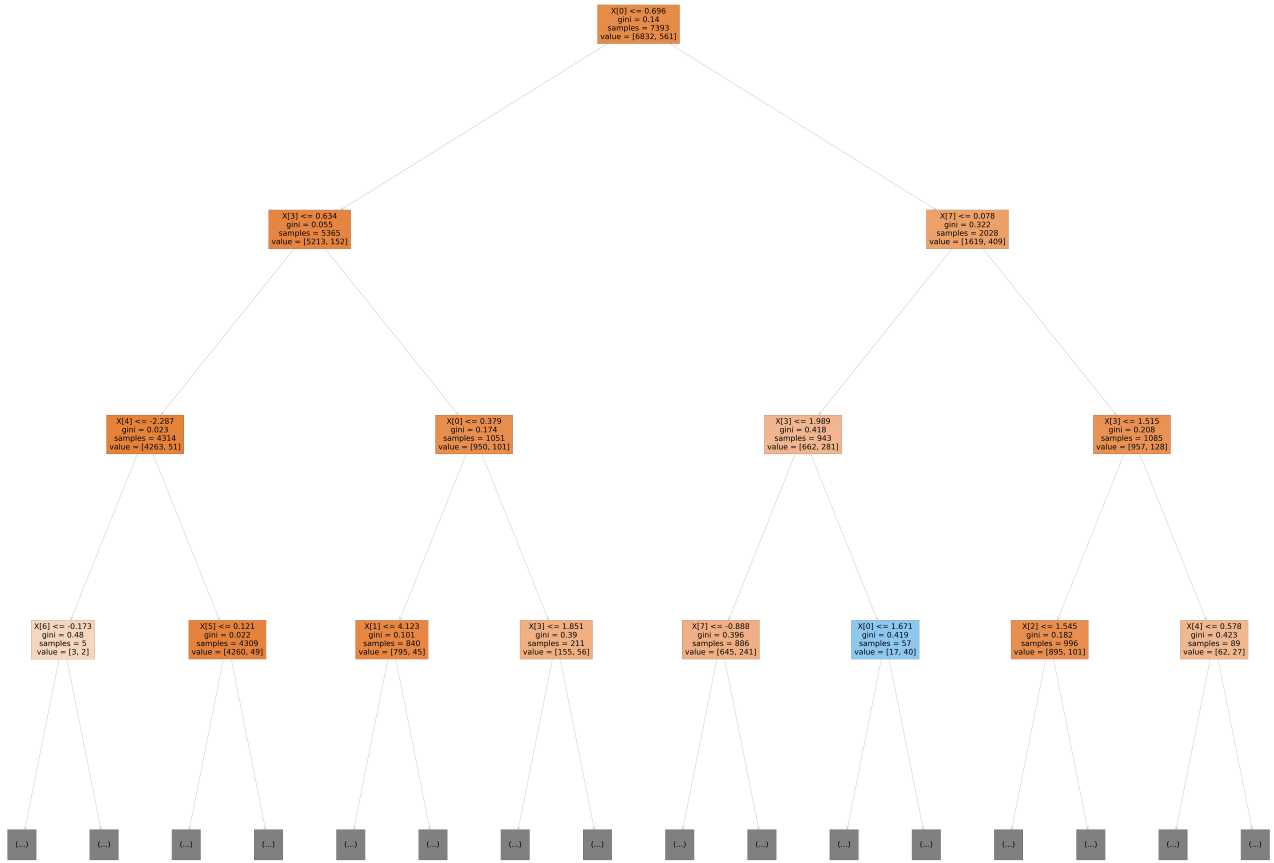



Figure: Decision tree

# Algorithms

- Random forest

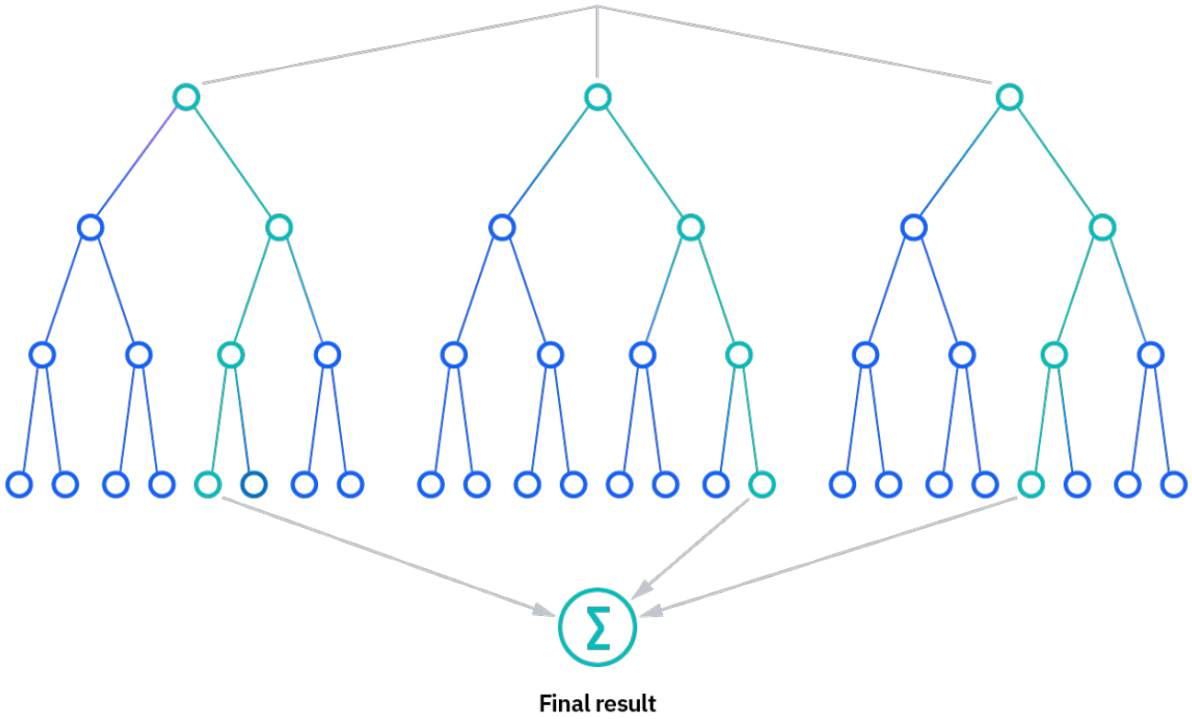



Figure: Random forest

# Algorithms

- AdaBoost

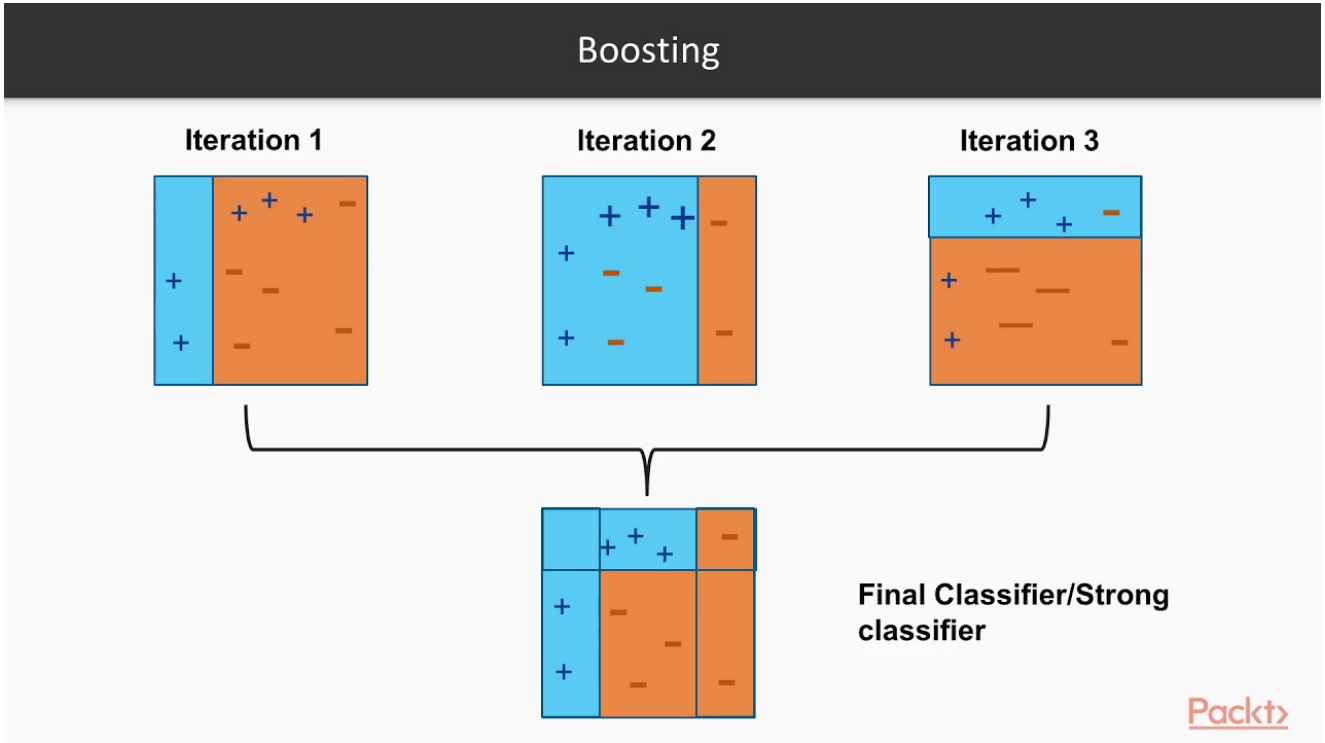



Figure: AdaBoost

# Algorithms

- Naive Bayes

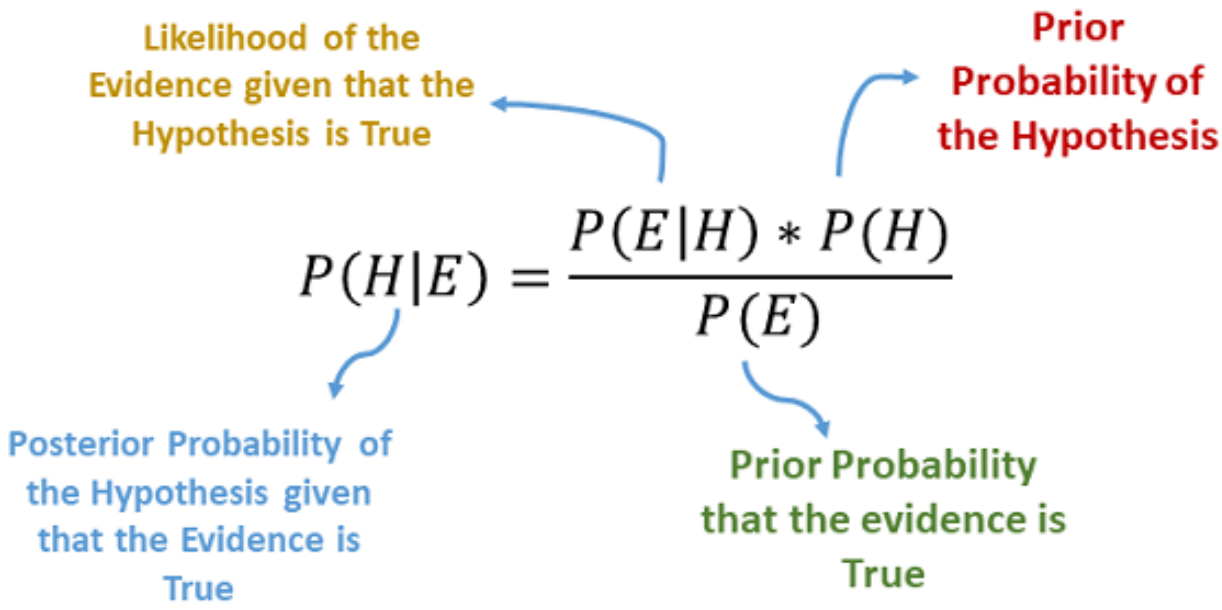
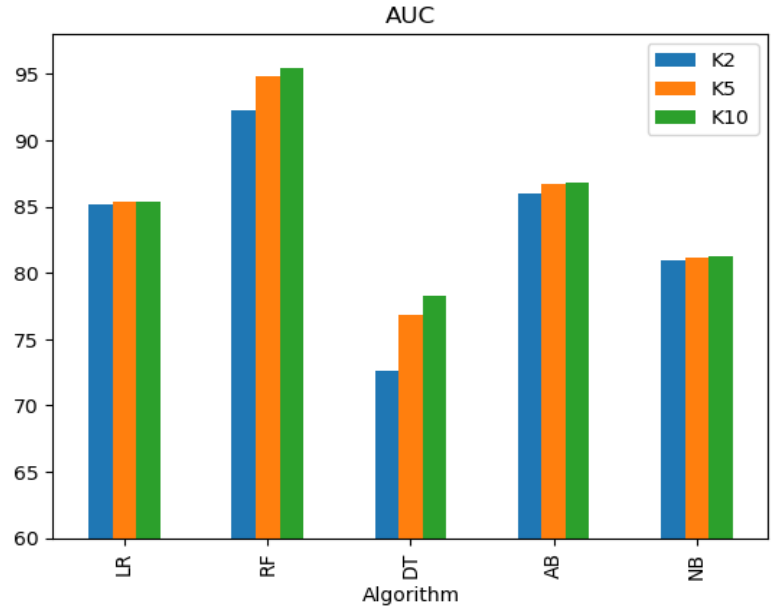


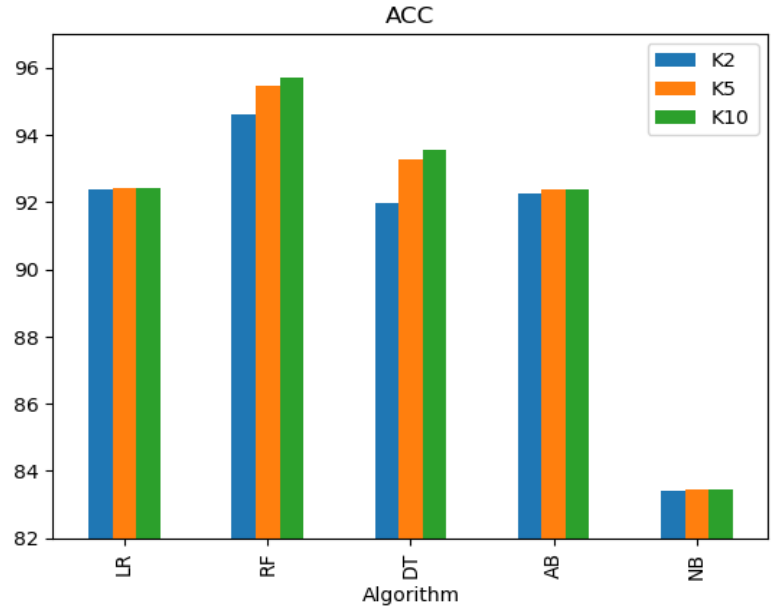

Figure: Naïve Bayes

# Results

# Performance measurement criteria

- Accuracy and AUC



# Performance measurement criteria

- ROC

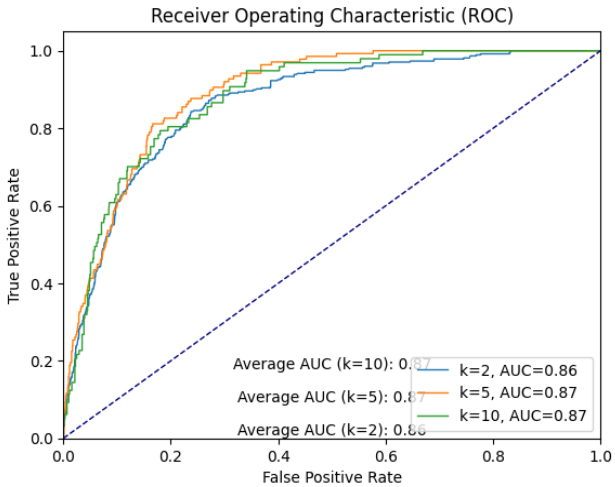



Figure: Ada

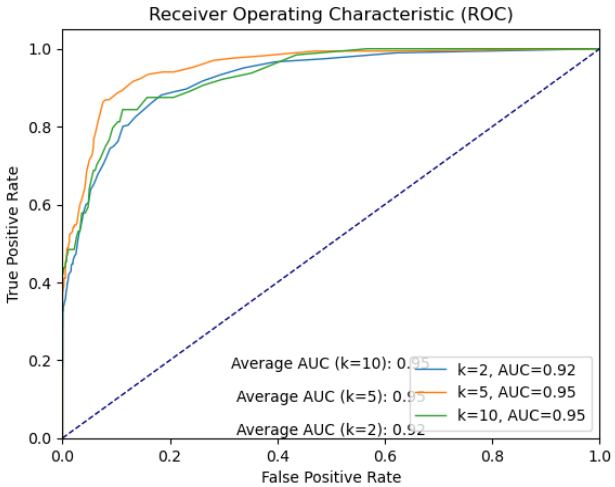



Figure: RF

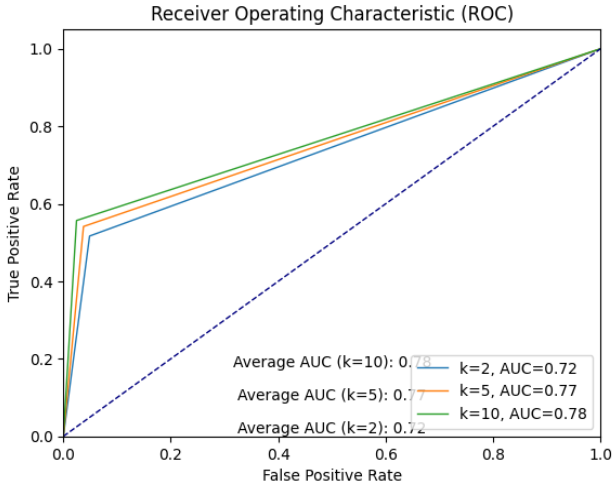



Figure: DT

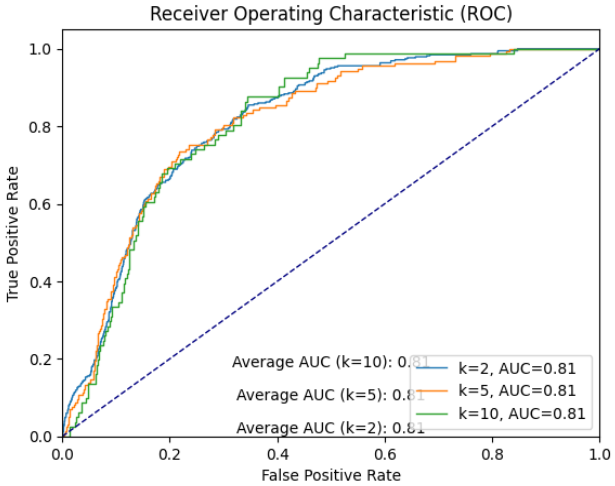



Figure: NB

# Performance measurement criteria

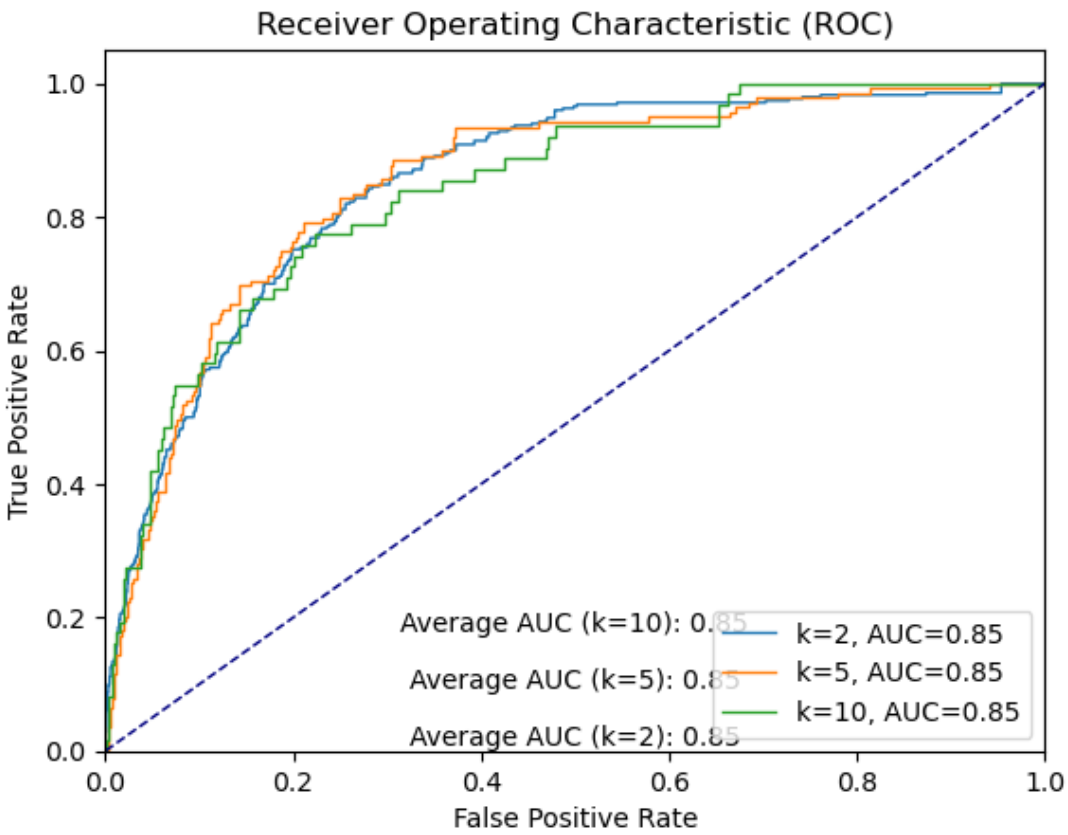
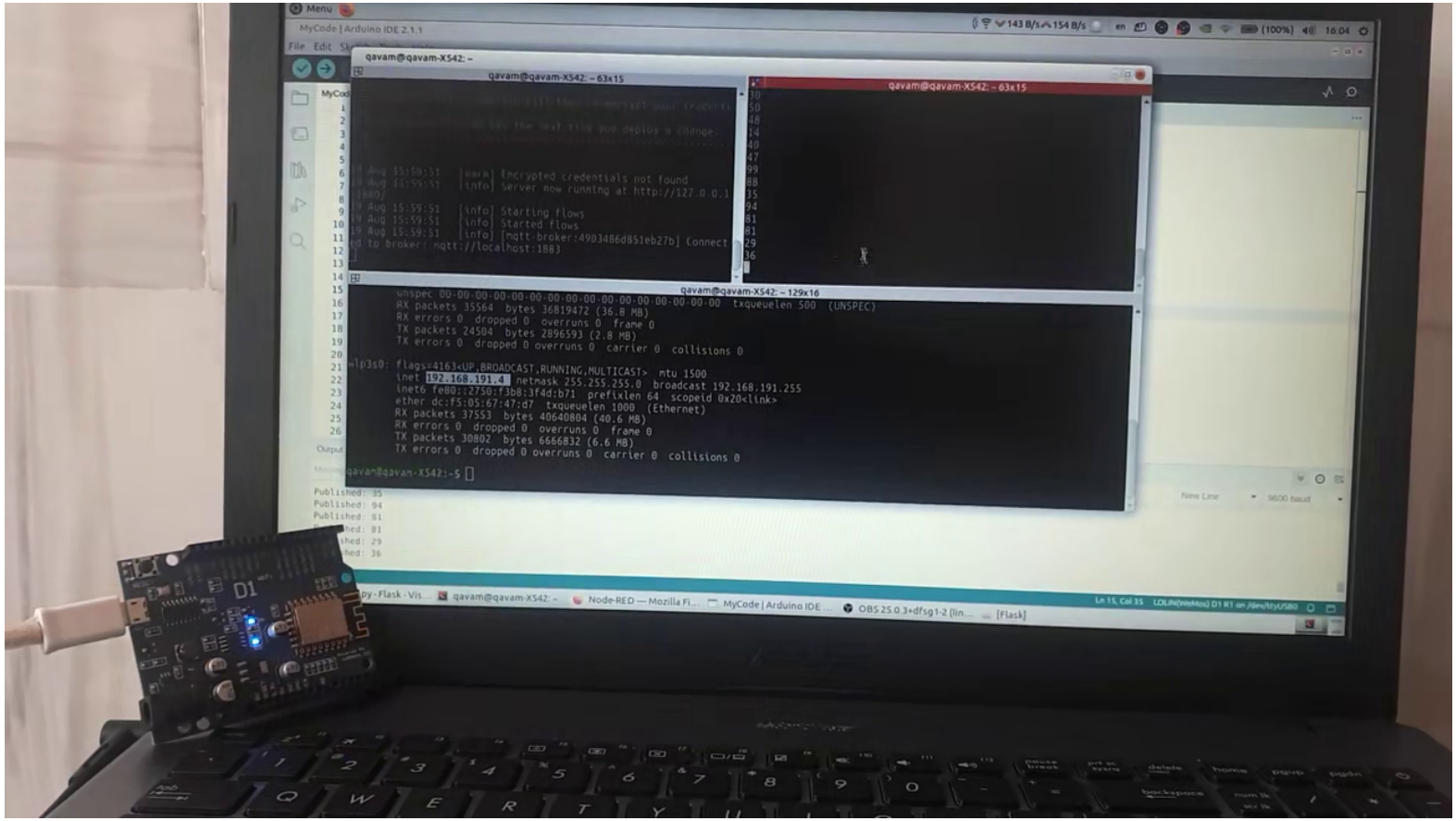


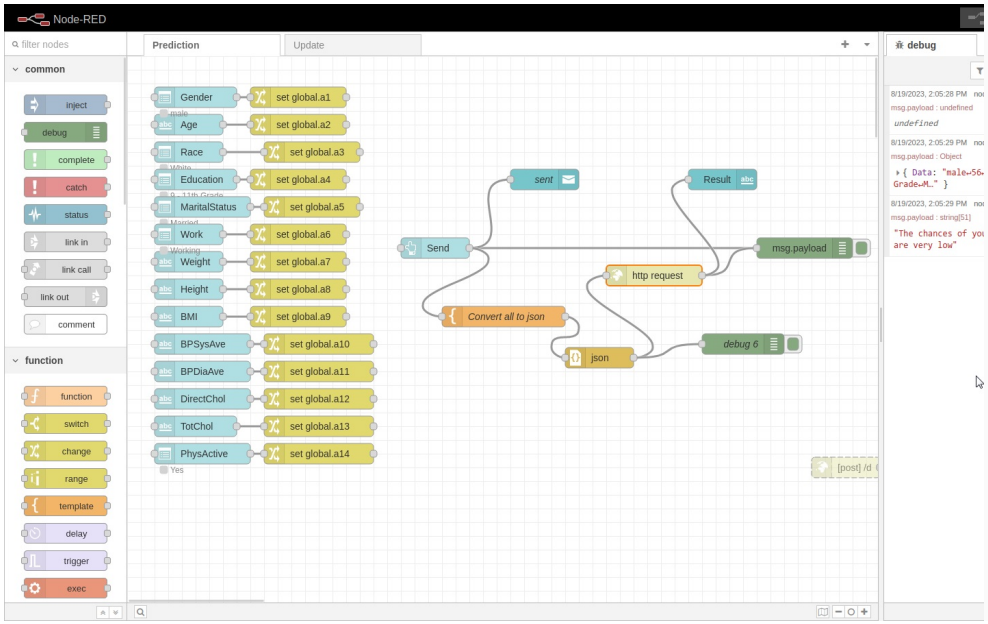

Figure: LR

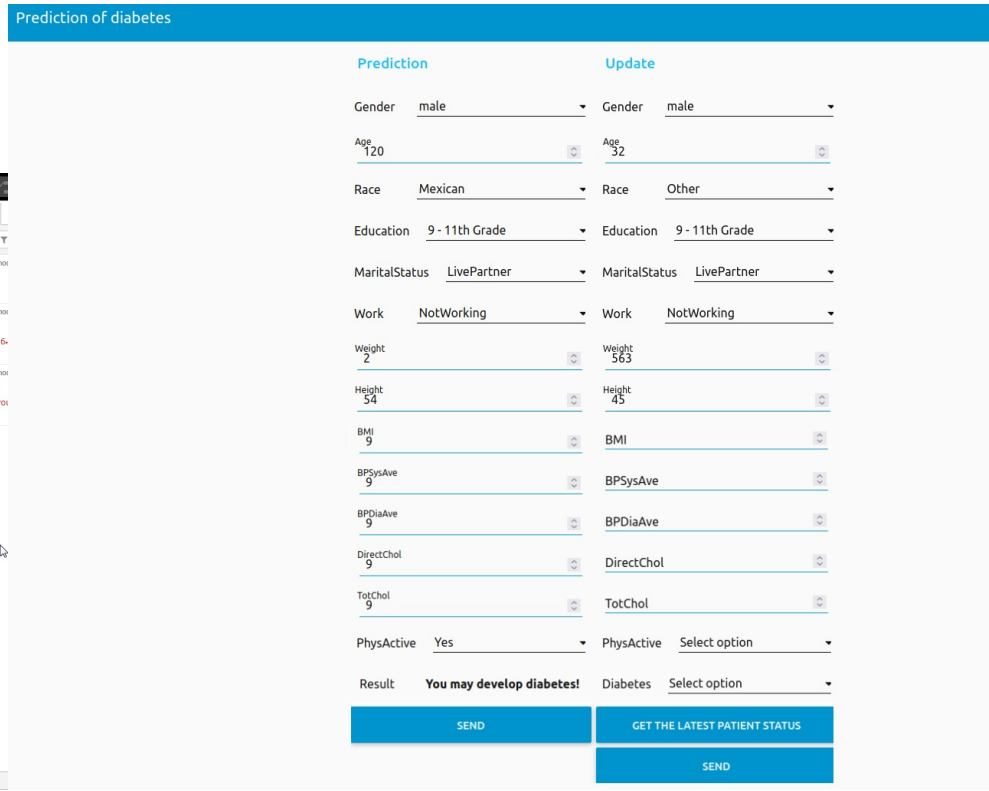
# IOT

- Arduino


qavam@qavam-X542: ~
qavam@qavam-X542: ~ 63x15
qavam@qavam-X542: ~ 129x16
[info] Starting flows
[info] Started flows
[info] [mqtt-broker:4903486d851eb27b] Connect
to broker: mqtt://localhost:1883
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?