Faculty of Engineering **Systems & Biomedical Engineering Department**

Cairo University

Artificial Intelligence in Medicine- Final Project
Diabetes detection

Done by:

Name	Sec	Bn
Muhammed Magdy Abdel Hady	2	18
Kareem Mustafa Mahmoud	2	6
Omar Essam Muhammed	2	3
Aya Emad Fuad	1	18

Contents

Introduction

Diabetes is characterized by abnormally high levels of sugar; after a meal, the amount of the blood glucose in the blood increase which release the insulin hormone, which stimulates the muscle and cells to use the blood glucose causing the blood sugar level to decrease to normal levels.

The Diabetes can be classified as type I and type II:

- Type I: the type I considered to be 5% of the diabetes patients which is a chronic condition in which the pancreas produces little or no insulin.
- Type II: the type II considered to be 95% of the diabetes patients which is associated with obesity, where the body can't regulate the use of the blood sugar normally.

There is another type of diabetes named Gestational Diabetes that develops during pregnancy and usually disappears after giving birth.

In out project we are focusing on classifying whether you have diabetes or not based on some main features the features:

- General Health
- Blood Pressure
- Body Mass Index
- Difficulty walking
- High cholesterol
- Age
- Heart Attack
- Physical Health
- Had Stroke or not
- Income
- Education
- Physical Activity
- Mental Health

Data Preparation

The previously mentioned features are selected from 330 features in the original dataset and dropped the records with missing values; the original data had three classes, 0 for no diabetes, 1 for pre-diabetes and 2 for diabetes, however we binarized the data to be 0 for no diabetes and 1 for pre-diabetes and diabetes at the same time.

The data used consists of 253680 records.

Data preparation process:

1. Binarize the data to be only labeled as 0 and 1.
2. Choosing the top contributing features through SelectKBest tool provided by sklearn library as it uses the co-matrix to determine what is the best features to use
3. Normalizing the data

After the data preparation we split the data 70-30% to start train the model and test the performance of the model.

Scientific Papers' Models Comparison

Henock and Intaek

Model	Accuracy %
Logistic Regression	71
Reinforcement	73
XGBoost	72
SVM	73
CIM	73

Sisodiaa

Model	Accuracy %
Naïve Bayes	76.30

Yunzhen Ye et al

Model	Accuracy %
Random forest	80.84

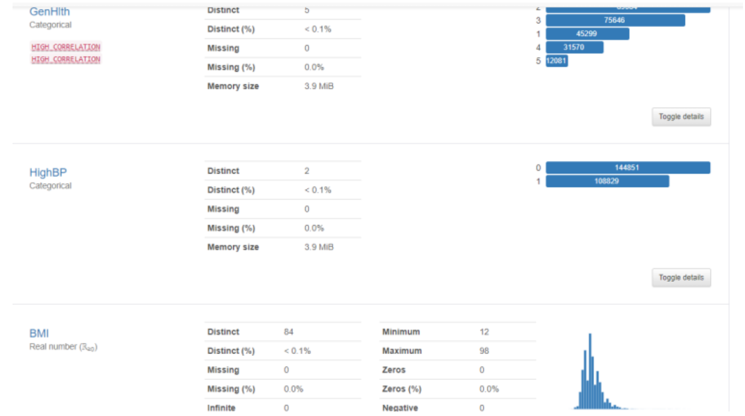
Amelec Viloria

Model	Accuracy %
SVM	95.36

Mujumdar

Model	Accuracy %
Decision Tree	86
Random forest	91
AdaBoost	93
LR	96
Bagging	90
KNN	90

Data Visualization



GenHlth
Categorical
HIGH CORRELATION
HIGH CORRELATION

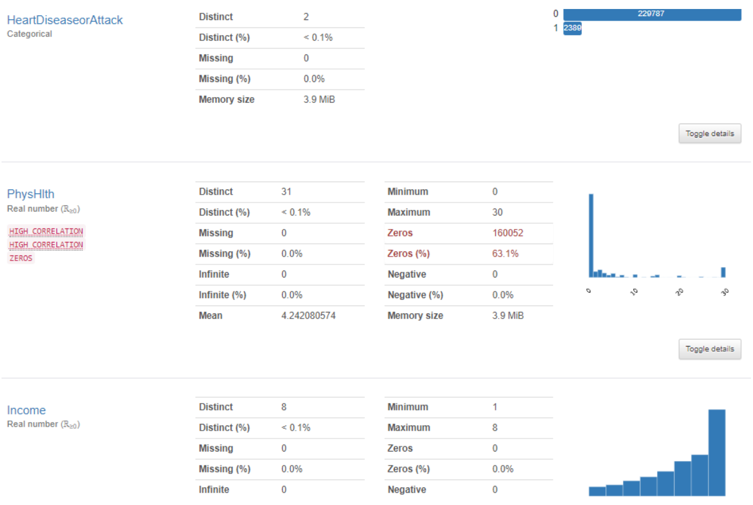
Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.9 MiB

HighBP
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.9 MiB

BMI
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	84	Minimum	12
Distinct (%)	< 0.1%	Maximum	98
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0



HeartDiseaseorAttack
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.9 MiB

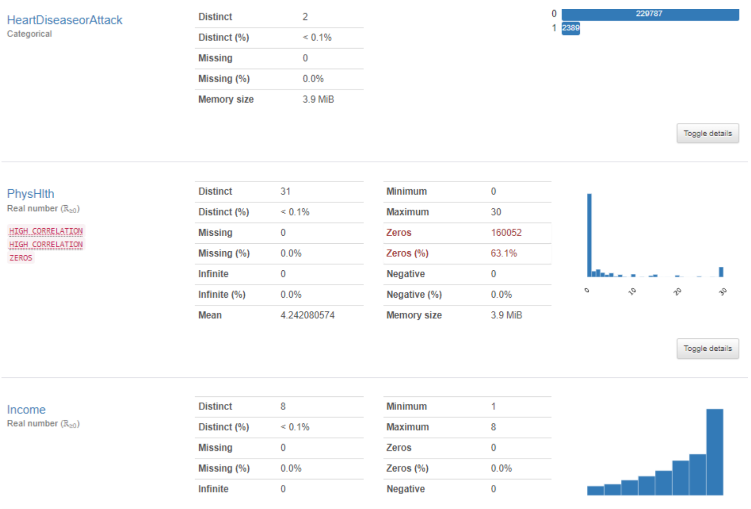
PhysHlth
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	31	Minimum	0
Distinct (%)	< 0.1%	Maximum	30
Missing	0	Zeros	160052
Missing (%)	0.0%	Zeros (%)	63.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	4.242080574	Memory size	3.9 MiB

Income
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	8	Minimum	1
Distinct (%)	< 0.1%	Maximum	8
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0



HeartDiseaseorAttack
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.9 MiB

0 229787
1 2389
Toggle details

PhysHlth
Real number ($\mathbb{R}_{\geq 0}$)

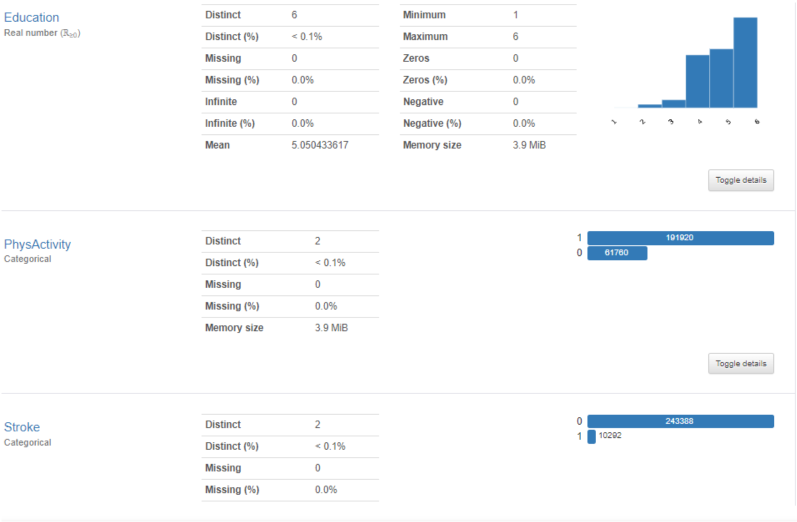
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	31	Minimum	0
Distinct (%)	< 0.1%	Maximum	30
Missing	0	Zeros	160052
Missing (%)	0.0%	Zeros (%)	63.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	4.242080574	Memory size	3.9 MiB

0 10 20 30
Toggle details

Income
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	8	Minimum	1
Distinct (%)	< 0.1%	Maximum	8
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0



Education
Real number ($\mathbb{R}_{\geq 0}$)

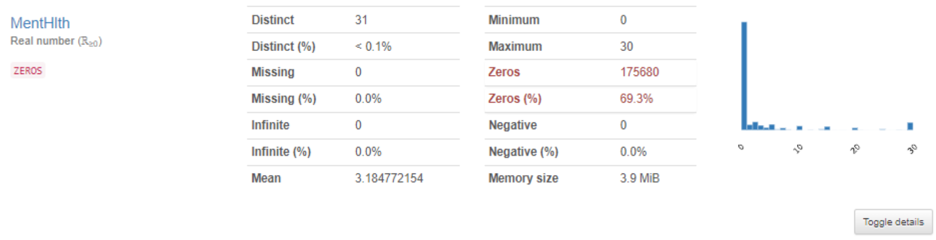
Distinct	6	Minimum	1
Distinct (%)	< 0.1%	Maximum	6
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	5.050433617	Memory size	3.9 MiB

PhysActivity
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.9 MiB

Stroke
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

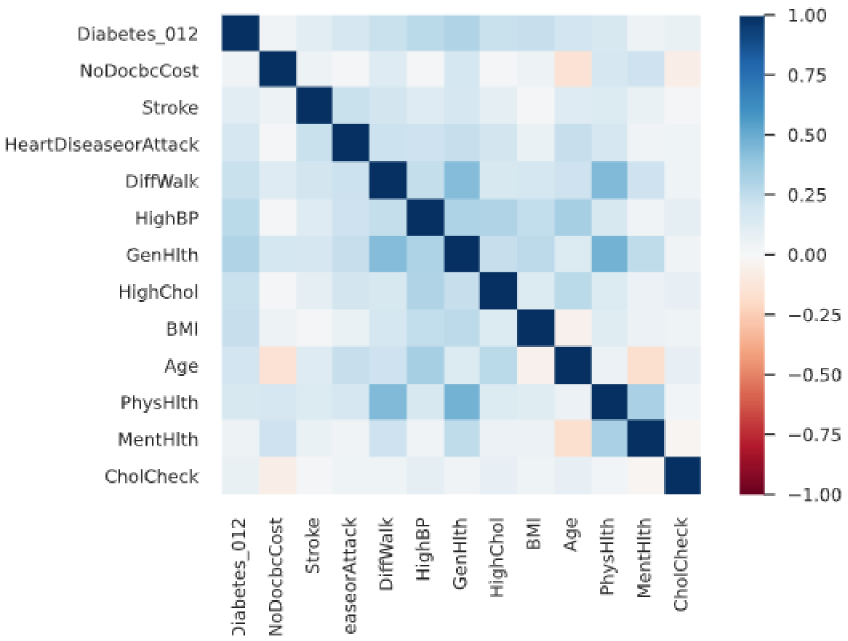


MentHlth
Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct	31	Minimum	0
Distinct (%)	< 0.1%	Maximum	30
Missing	0	Zeros	175680
Missing (%)	0.0%	Zeros (%)	69.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	3.184772154	Memory size	3.9 MiB

Correlation



Results

We get an accuracy 85% using more than 250000 records

There is a way we can oversample with a better accuracy 89% so the model can learn a lot about both labels

Conclusion