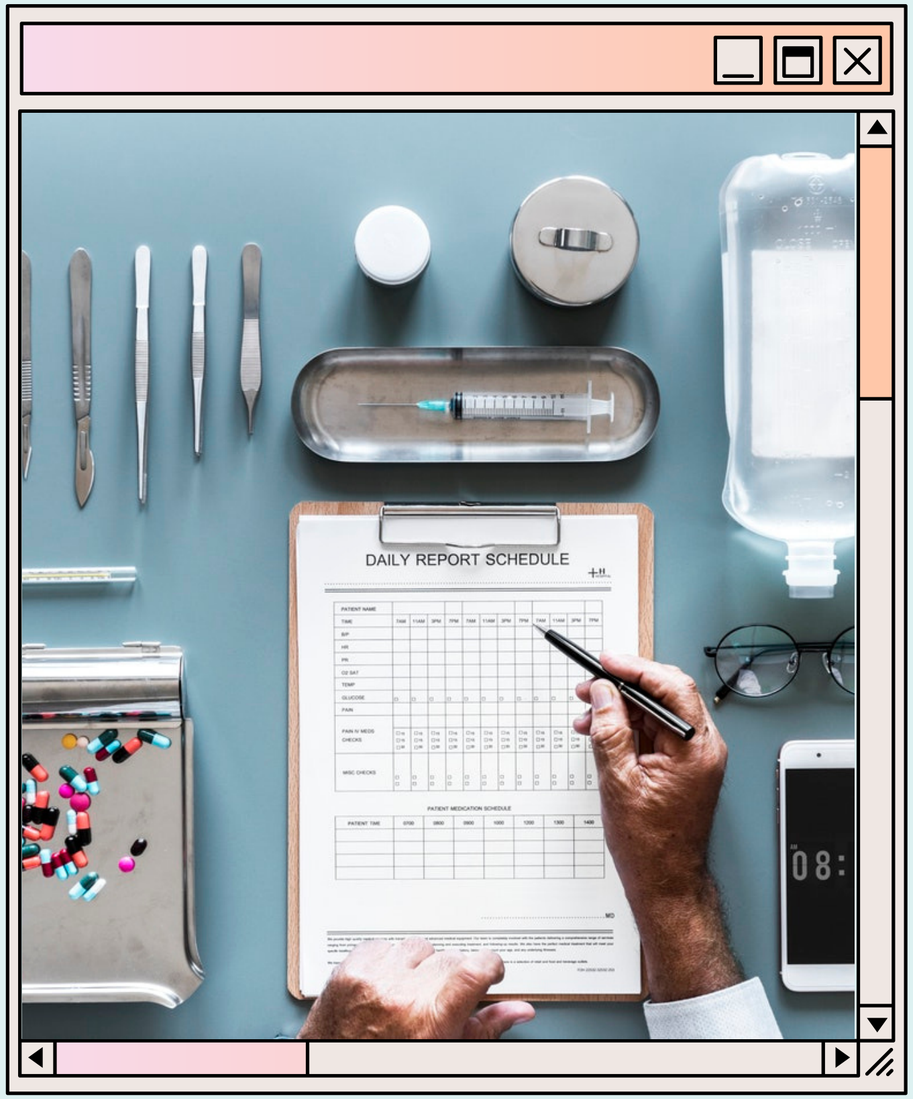

NURUL LIYANA

Project Proposal DS105

VIEW MORE →

NURUL LIYANA

New Tab

Diabetes Health Indicator Dataset

I've chosen this dataset as I would like to know the risk factors getting diabetic and what can be prevented with the risk factors known.

VIEW MORE →

NURUL LIYANA

Script Runner 2.0

Script Language: HyperType

Contents

Run Script

VIEW MORE →

About the dataset

Source of data: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_012_health_indicators_BRFSS2015.csv

Diabetes Health Indicators Dataset

Data Code (40) Discussion (5)

210 New Notebook Download (6 MB)

About Dataset

Context

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of

Usability
10.00

License
CC0: Public Domai

Expected updat
Never

NURUL LIYANA

Size of dataset

Number of rows:

253680

Number of columns:

22

VIEW MORE →

Records found in dataset

1.DIABETES_012	Currently have diabetes
2.HIGHBP	Currently have high blood pressure
3.HIGHCHOL	Currently have high cholesterol
4.CHOLCHECK	Did their cholesterol check in 5 years
5.BMI	Body Mass Index
6.SMOKER	Smokes at least 100 cigarettes
7.STROKE	Ever encounter have stroke
8.HEARTDISEASEORATTACK	Coronary heart disease or myocardial infarction
9.PHYSACTIVITY	Done physical activity in the past 30 days
10.FRUITS	Consume Fruits per day
11.VEGGIES	Consume Vegetables per day
12.HVYALCOHOLCONSUMP	Heavy drinkers
13.ANYHEALTHCARE	Any healthcare coverage
14.NODOCBCCOST	Didn't see doctor because of cost
15.GENHLTH	General health
16.MENTHLTH	Mental health
17.PHYSHLTH	Physical health
18.DIFFWALK	Serious difficulty walking or climbing stairs
19.SEX	Gender
20.AGE	13-level age category 18 < x >= 80
21.EDUCATION	Education level (Scale 1-6)
22.INCOME	Annual income (scale 1-8)

Important Columns

DIABETES_012	To separate the non/pre-diabetic and diabetic
GENHLTH MENTHLTH PHYSHLTH	To see one overall health in general, mental and physical.
AGE EDUCATION INCOME	To look at which part of position they are sitting in life.

VIEW MORE →

Sample of data

pd.read_csv('diabetes_012_health_indicators_BRFSS2015.csv')

diabetes_012_health_indicators_BRFSS2015

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	1.0	10.0	6.0	8.0
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	14.0	0.0	0.0	9.0	6.0	7.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	11.0	4.0	4.0
2.0	1.0	1.0	1.0	30.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	5.0	30.0	30.0	1.0	0.0	9.0	5.0	1.0
0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	8.0	4.0	3.0
2.0	0.0	0.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	1.0	13.0	6.0	8.0
0.0	1.0	1.0	1.0	34.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	30.0	1.0	0.0	10.0	5.0	1.0
0.0	0.0	0.0	1.0	26.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	15.0	0.0	0.0	7.0	5.0	7.0
2.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	4.0	6.0
0.0	0.0	1.0	1.0	33.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	4.0	30.0	28.0	0.0	0.0	4.0	6.0	2.0

Challenges/difficulties faced

There is class imbalance in this dataset	1
Smaller amount of people to focus on the health problem faced	2
Minimal knowledge on diabetes but wanting to know the risk factors	3

ERROR!

We
In

Th
on

It
pro

To

< Back

sub-goals

1. IS THERE ANY NON-DIABETIC AND DIABETIC INDIVIDUALS THAT HAS THE SAME RISK FACTORS WHICH COULD LED TO BEING DIABETIC

2. WHAT RISK FACTORS ARE MOST PREDICTIVE OF DIABETES RISK

3. CAN I CONCLUDE WITH THE LITTLE AMOUNT OF DIABETIC INDIVIDUALS FOUND WITH THE RISK FACTORS TO PREDICT ACCURATELY IF ONE HAS DIABETES

4. DOES PHYSICAL AND MENTAL HEALTH PLAYS A MAJOR PART IN DIABETIC INDIVIDUALS

5. DOES INCOME HAS AN EFFECT FOR THE COST TO SEE A DOCTOR

6. THOSE INDIVIDUAL WHO DO NOT EAT FRUITS AND VEGGIES ARE DIABETIC

7. CAN DIABETIC INDIVIDUALS BE AT THE LOWER RANGE AND HIGHER RANGE OF BMI OR EITHER ONE OF THE RANGE

Goals

To build a machine learning model that is able to help an individual to measure their health condition

To provide suggestions to non/pre-diabetic as to not fall into the diabetic category

To enjoy life while being in a healthy state

NURUL LIYANA

Classification Problem

Chosen problem/approach

Will be using these algorithm to see which is the best outcome

1. Logistic Regression

2. K-Nearest Neighbours

3. Decision tree

Logistic Regression algorithm which can be use for Binary classification as most of the columns are numerical/binary.

K-Nearest Neighbour algorithm could display a different scene for individuals to observe if they are more prone to be diabetic due to the other health issues they are facing or mainly the food and lifestyle

Decision tree algorithm could promote to new individuals on try seeing if the tree could lead to being pre-diabetic to diabetic

THANK YOU!