Roshith K R - RVCE20MCA009

Akshay T N – RVCE20MCA069

Phase 1 Machine Learning:

1) **Problem Definition:**

Diabetes is a Chronic disease characterized by hyperglycemia. It is characterized by high blood glucose levels resulting from defects in insulin production, insulin action or may be both. based on their blood pressure, and other we will predict which is more beneficial/important to get the accurate output.

our dataset consists of pregnancies woman getting infected to diabetes along with glucose, insulin level, skin-thickness and BMI. we have further classified into age and diabetes pedigree function to get the accurate result he/she is diabetic.

2) **What, Why and How:**

What – assuming that we are a disease prediction center, on what particular aspect the person get infected to disease will find out here, so that we will get the outcomes of the positive diabetic patient.

Why – solving this problem will help to sort or figure out which persons will get disease diabetes in what stage or reason.

How – we can solve this issue/medical problem using previous set of medical data, and using an KNN-Algorithm.

3) **Data Set and Features Considered:**

Data Set – Diabetic Prediction Data set scraped from National Institute of Diabetes.

Features Considered - Glucose, blood-sugar, Age, BMI, Diabetes Pedigree function, Thickness, Pregnancies, Skin-thickness and outcome.

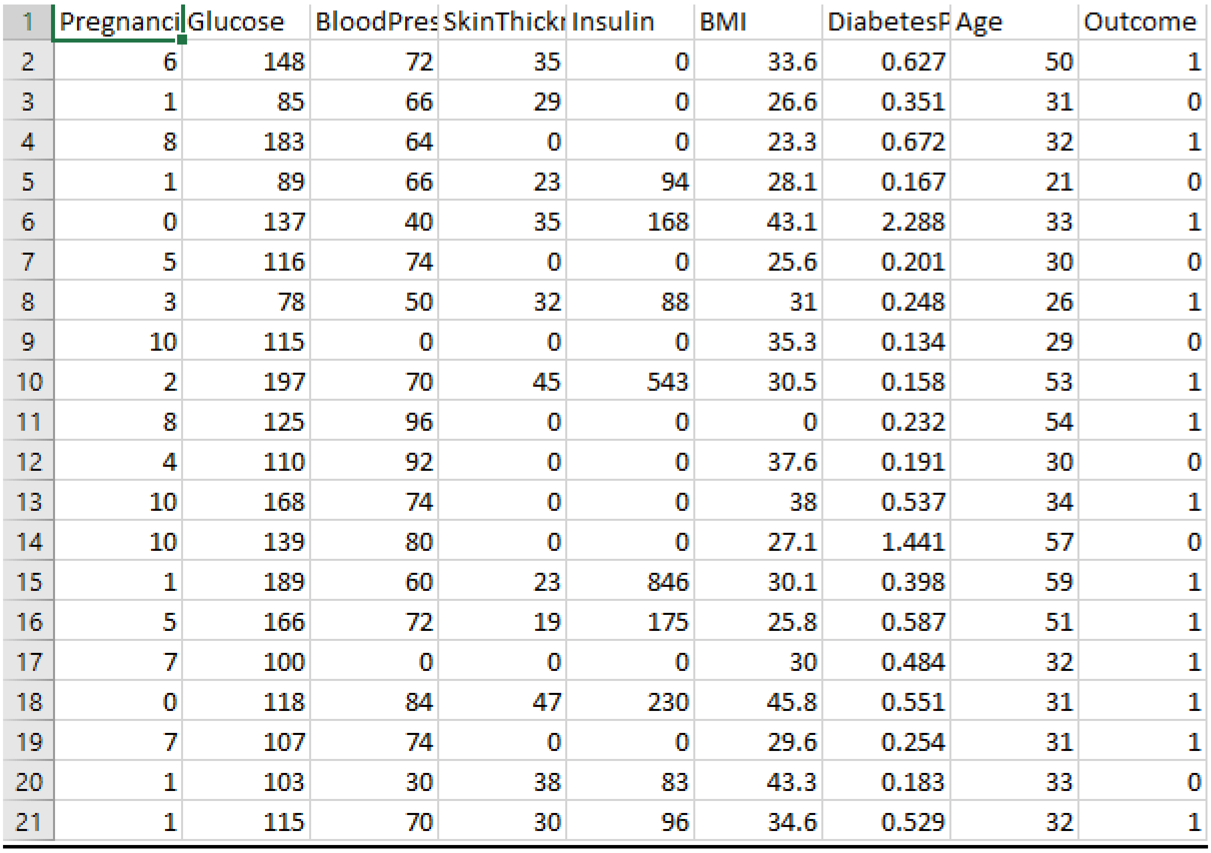
Dataset has about 769 Patient Records.

4) **Type of Algorithm to be used and Why!?:**

Type - K-Nearest Algorithm (KNN) Based on Supervised Learning Technique.

Why – Since we will be predicting diabetic or not, we will use KNN because it uses data to build a model and then uses that model to predict the outcome in 0 or 1 and which makes the disease risk prediction model more effective than other classifiers.

5) **<u>Sample Dataset With Explanation:</u>**



1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1

<u>NOTE:</u>

Glucose: The Glucose level of the person and it is the main sugar found in our blood of the person.

Blood-pressure: The blood pressure range or level of the person which may be high or low in the person.

Age: The ages of a particular person will get infected. here, the particular person age is available.

BMI: BMI will give a body mass index to measure the body fat mass of the particular person.

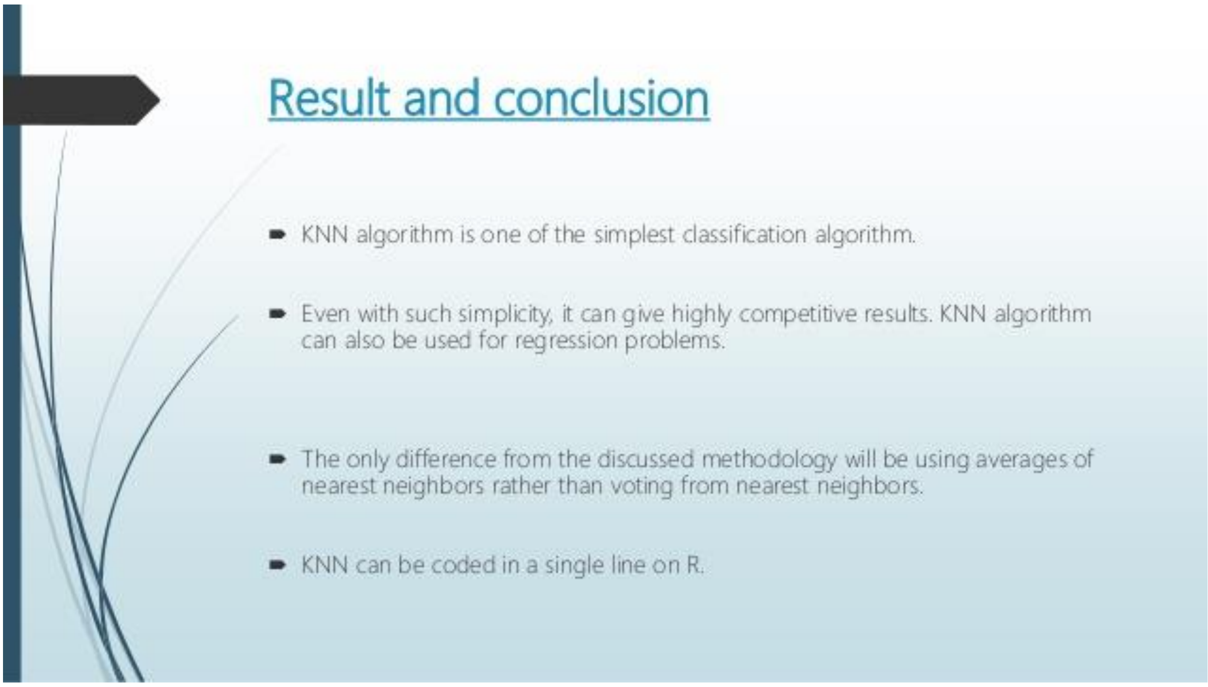
Diabetes Pedigree Function: It is a function which gives scores/values based on the family history of whether a patient is diabetic or not in bit like 0 or 1.

Skin Thickness: It gives thickness of the skin when the person gets diabetic. Automatically skin get thickness.

Pregnancies: The number of pregnancy woman getting high blood sugar levels in pregnancy time and also chances of getting infected to diabetes. here, number of persons data given.

Outcome: It will give the final result as whether the person is diabetic or not in bit values like 0 or 1. If 0 means he/she is non diabetic and 1 means he/she is diabetic. It is a target variable.

6) **KNN-Algorithm:**



Result and conclusion

- KNN algorithm is one of the simplest classification algorithm.
- Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems.
- The only difference from the discussed methodology will be using averages of nearest neighbors rather than voting from nearest neighbors.
- KNN can be coded in a single line on R.