HEALTH CARE DIABETES PROJECT

PROBLEM STATEMENT

Build a model to accurately predict whether the patients in the dataset have diabetes or not?

DATA

In this project we are provided with the **health care diabetes.csv** dataset containing 768 data points and 9 features.

The features are described below:

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age (years)
Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

The features Glucose to DiabetesPedigreeFunction have float data type and others are integers. The target variable ('Outcome') is binary

METHODS

Data Exploration

- Load data into Python environment and check dimensionality
- View the data
- Visually explore the data using histograms
- Performing descriptive analysis before cleaning
- Check for missing values
- Clean the missing values in the diabetes_data by filling mean and median
- Perform descriptive analysis after cleaning

Data Modeling

The data has 768 samples (>50 samples). Outcome is the target variable, so this is supervised learning problem.

We predicting a category, so this is a classification problem.

The data is imbalanced with 268 outcomes being classified as 0 and 500 outcomes being classified as 1.

We try variety of algorithms. This can be especially beneficial for imbalanced data

The data has 768 samples(<100K samples), we can use LinearSVC

If LinearSVC is not working, use KNeighbors Classifier.

If KNeighbors is not working, use tree-based algorithms, e.g., Random Forest Classifier.

Decision trees often perform well on imbalanced datasets.

Accuracy is not the best metric when evaluation imbalanced data sets as it can be very misleading

Metrics that can provide better insight include: confusion matrix, precision, recall f1: score

Data Reporting:

A dashboard was created in tableau by choosing appropriate chart types and metrics useful for the business.

The dashboard entails the following:

a. Pie chart to describe the diabetic or non-diabetic population

b. Scatter charts between relevant variables to analyze the relationships

c. Histogram or frequency charts to analyze the distribution of the data

d. Heatmap of correlation analysis among the relevant variables

e. A bubble chart to analyze different variables for these age brackets : 20-25, 25-30, 30-35, etc.

Link:https://public.tableau.com/profile/thobela4031#!/vizhome/Healthcaredashboard_15855990973450/Healthcaredashbord

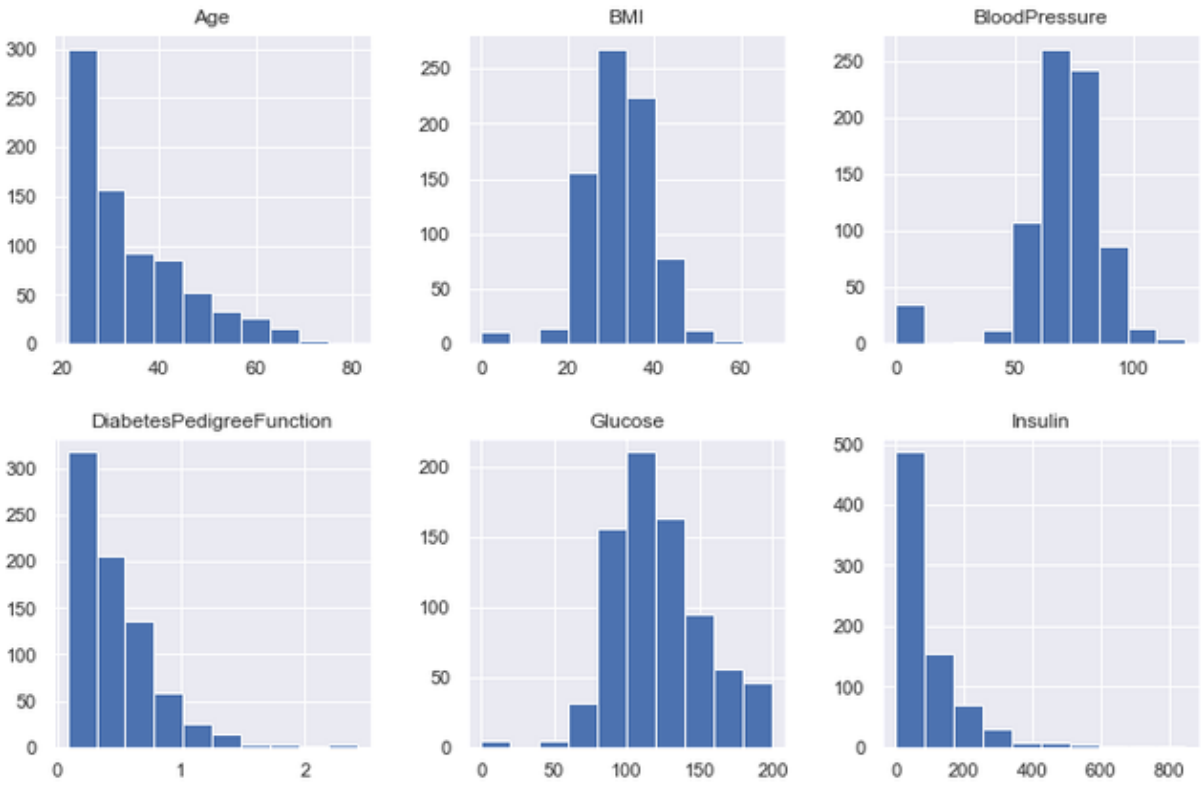
RESULTS

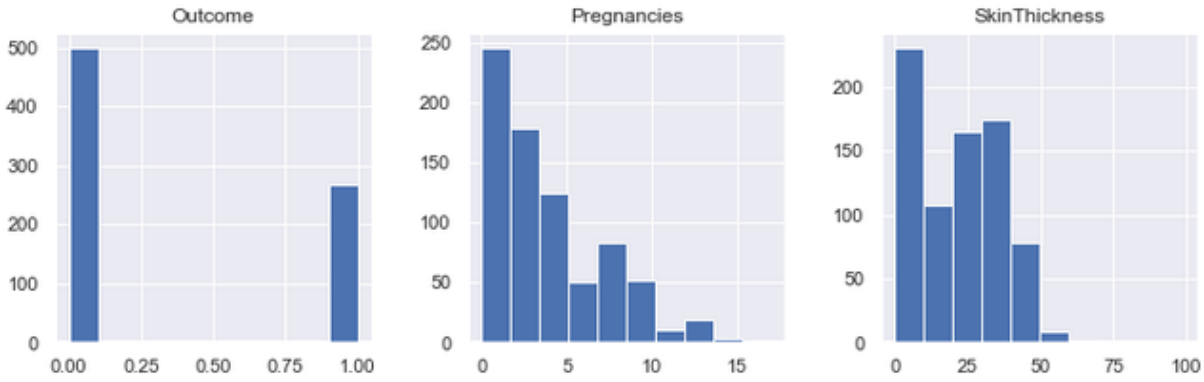
Data Exploration

Table 1: The health care diabetes dataset showing the features and the first 5 rows of data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Distribution of the data in each column





Descriptive analysis

Table 2: Descriptive analysis of the data before cleaning

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Relationships between the pair of variables in the data

Table 3: Descriptive analysis of the data after cleaning

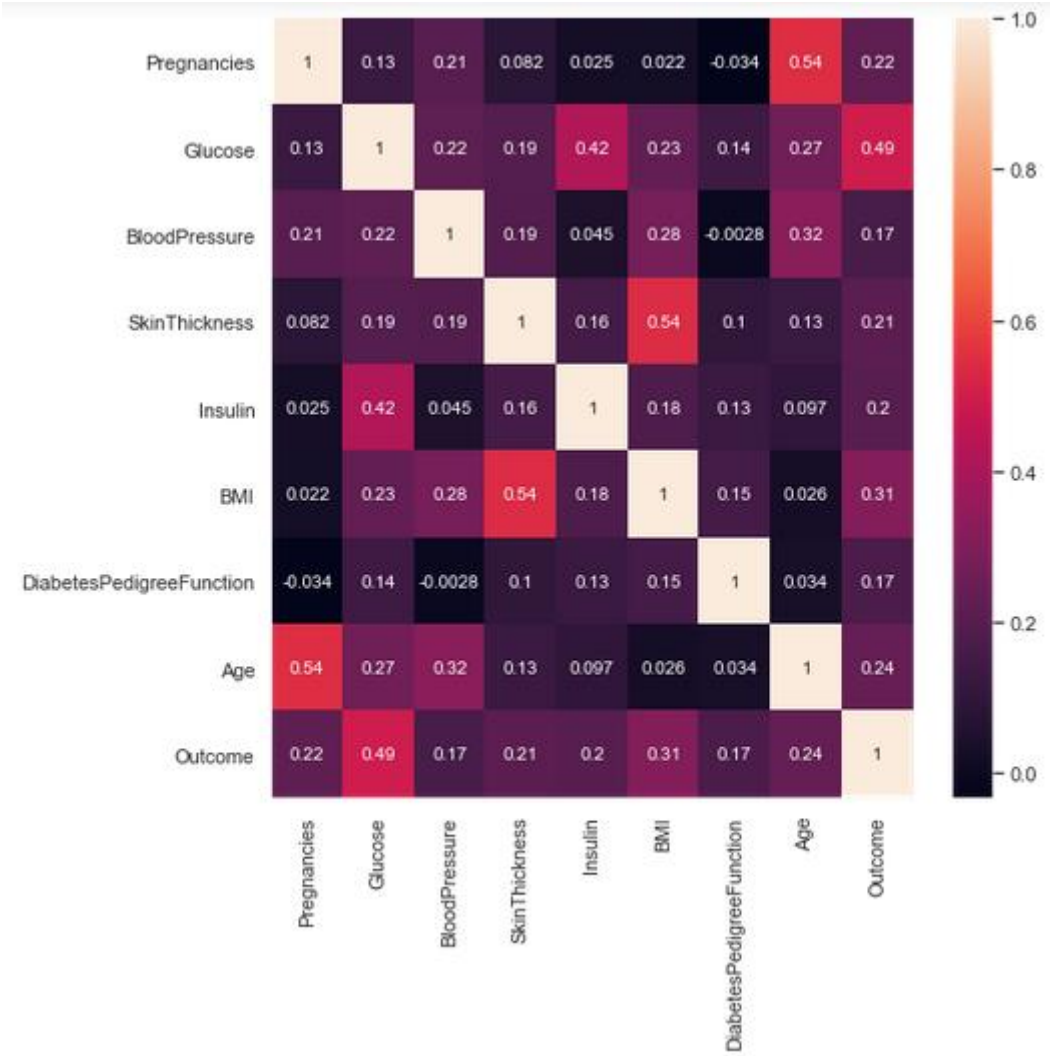
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.686763	72.405184	29.108073	140.671875	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.435949	12.096346	8.791221	86.383060	6.875151	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.000000	125.000000	32.400000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Correlation analysis

Table 4: correlation analysis of the data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127911	0.208522	0.081770	0.025047	0.021565	-0.033523	0.544341	0.221898
Glucose	0.127911	1.000000	0.218367	0.192686	0.419064	0.230941	0.137060	0.266534	0.492928
BloodPressure	0.208522	0.218367	1.000000	0.191853	0.045087	0.281268	-0.002763	0.324595	0.166074
SkinThickness	0.081770	0.192686	0.191853	1.000000	0.155610	0.543162	0.102188	0.126107	0.214873
Insulin	0.025047	0.419064	0.045087	0.155610	1.000000	0.180170	0.126503	0.097101	0.203790
BMI	0.021565	0.230941	0.281268	0.543162	0.180170	1.000000	0.153400	0.025519	0.311924
DiabetesPedigreeFunction	-0.033523	0.137060	-0.002763	0.102188	0.126503	0.153400	1.000000	0.033561	0.173844
Age	0.544341	0.266534	0.324595	0.126107	0.097101	0.025519	0.033561	1.000000	0.238356
Outcome	0.221898	0.492928	0.166074	0.214873	0.203790	0.311924	0.173844	0.238356	1.000000

Figure 3: Visually exploring correlation using a heat map.



Data modelling:

LinearSVC

Table 5: Confusion matrix

```
[[137  20]
 [ 35  39]]
```

Table 6 : Classification report

	precision	recall	f1-score	support
0	0.80	0.87	0.83	157
1	0.66	0.53	0.59	74
accuracy			0.76	231
macro avg	0.73	0.70	0.71	231
weighted avg	0.75	0.76	0.75	231

KNN

Table 7 : Confusion matrix

```
[[137  20]
 [ 30  44]]
```

Table 8 : Classification report

	precision	recall	f1-score	support
0	0.82	0.87	0.85	157
1	0.69	0.59	0.64	74
accuracy			0.78	231
macro avg	0.75	0.73	0.74	231
weighted avg	0.78	0.78	0.78	231

RandomForestClassifier

Table 9 : Confusion matrix

```
[[134  23]
 [ 32  42]]
```

Table 10 : Classification report

	precision	recall	f1-score	support
0	0.81	0.85	0.83	157
1	0.65	0.57	0.60	74
accuracy			0.76	231
macro avg	0.73	0.71	0.72	231
weighted avg	0.76	0.76	0.76	231

ANALYSIS

Data Exploration

Figure 1 shows how the data is distributed in each. For Glucose, BloodPressure and BMI, data is not skewed. For Insulin and SkinThickness columns, the data is skewed to th left. It appears as if there are no missing values in the data. But zeros on the columns, Glucose, BloodPressure, SkinThickness, Insulin and BMI, are meaningless and thus indicates missing value.There are 5 missing values in the Glucose column, 35 in the Bloodpressure column ,227 in the SkinThickness column, 374 in the Insulin column and 11 in the BMI column. For those columns where the data is not skewed, the mean values of that column is used to fill the missing values in it. For those columns where the data is skewed, the median is used. The relationship between pair of variables in the data is shown in Figure 2. There is a positive linear relationship between BMI and SkinThickness, Insulin and Glucose, BloodPressure and Age, BMI and BloodPressure and between Glucose and Age. This means that an increase in one variable in a pair positively affect another variable.

Data Modeling

KNN (shown in Table 8) has the highest accuracy score, precision, recall and f1-score compared to Linear SVC (shown in Table 6) and Random Forest (shown in Table 10). Thus KNN is the best estimator for this problem. Table 7 show the confusion matrix of KNN. About 137 patients being correctly predicted as not diabetic and 20 being falsely predicted as not diabetic. And 30 being falsely predicted diabetic, while 44 are correctly predicted as diabetic. Table 8 shows the classification report of KNN. 82 % of the time the classifier correctly label 0 as 0, and 69 % of the time the classifier correctly label 1 as 1. 87% of the time the classifier is able to find all the 0 samples, and 59% of the time the classifier is able to find all the 0 samples.