Diabetes

Detect whether a person is diabetic or not?

Dataset Description

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Age	Outcome
count	768.000000	768.000000	763.000000	768.000000	768.000000	768.000000	768.000000
mean	120.894531	69.105469	20.520315	79.799479	31.992578	33.240885	0.348958
std	31.972618	19.355807	15.966929	115.244002	7.884160	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	21.000000	0.000000
25%	99.000000	62.000000	0.000000	0.000000	27.300000	24.000000	0.000000
50%	117.000000	72.000000	23.000000	30.500000	32.000000	29.000000	0.000000
75%	140.250000	80.000000	32.000000	127.250000	36.600000	41.000000	1.000000
max	199.000000	122.000000	99.000000	846.000000	67.100000	81.000000	1.000000

Data Preprocessing

SkinThickness column has 5 missing values (Nan)

```
[142] dataset.isnull().sum()
```

```
Glucose          0
BloodPressure    0
SkinThickness    5
Insulin          0
BMI              0
Age              0
Outcome          0
dtype: int64
```

Fill the missing values with the **most frequent** value

Classification Algorithms

Supervised learning algorithms

- KNN (k-nearest neighbors) Algorithm
- Naïve Bayes Algorithm

KNN: K-nearest neighbors Algorithm

Step-1: Select the number K of the neighbors **(K = 5)**

Step-2: Calculate the distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated by distance function.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Naïve Bayes Algorithm

Step 1: Convert the given dataset into frequency tables.

Step 2: Generate Likelihood table by finding the probabilities of given features.

Step 3: Use Bayes theorem to calculate the posterior probability.

Bayes theorem: $P(Y|X) = P(X|Y) * P(Y) / P(X)$

Classification Evaluation metrics

Accuracy = (TP+TN)/(TP+FP+FN+TN)

Precision = (TP)/(TP+FP)

Recall = (TP)/(TP+FN)

F1 = 2 * (precision * recall) / (precision + recall)

KNN vs Naïve Bayes

KNN behavior

accuracy score: **0.69**

precision score: **0.55**

recall score: **0.55**

F1 score: **0.55**

Naïve Bayes behavior

accuracy score: **0.72**

precision score: **0.6**

recall score: **0.59**

F1 score: **0.59**

Naïve Bayes is better than KNN

Team Members

Asmaa Mahmoud

Alaa Ahmed

Amr Mahmoud

Mustafa Mahmoud

Yousef Shereen

Omar Mohamed