DIABETES PREDICTION

In this diabetes prediction model, we used a dataset containing various health measurements of individuals, including blood sugar levels, body mass index (BMI), and age, among others. The goal was to predict whether a person is diabetic based on these measurements.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

DATA PREPARATION

We loaded the diabetes dataset into a Pandas DataFrame. We separated the features (input data) from the labels (output, or whether the person is diabetic). The dataset was split into training and testing sets, with 20% of the data reserved for testing.

```
data = pd.read_csv('/content/diabetes.csv')
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
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Next steps: Generate code with `data` | View recommended plots | New interactive sheet

```
data.isnull().sum()
```

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

dtype: int64

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
```

```
 ---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

```
data.shape
```

```
(768, 9)
```

```
data.describe()
```

	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

```
data['Outcome'].value_counts()
```

Outcome	count
0	500
1	268

dtype: int64

```
data.groupby('Outcome').mean()
```

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

```
x = data.drop(columns='Outcome',axis=1)
x
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63
764	2	122	70	27	0	36.8	0.340	27
765	5	121	72	23	112	26.2	0.245	30
766	1	126	60	0	0	30.1	0.349	47
767	1	93	70	31	0	30.4	0.315	23

768 rows × 8 columns

Next steps: Generate code with x | View recommended plots | New interactive sheet

```
y = data['Outcome']
y
```

	Outcome
0	1
1	0
2	1
3	0
4	1
...	...
763	0
764	0
765	0
766	1
767	0

768 rows × 1 columns

dtype: int64

DATA STANDARDIZATION

We standardized the feature data to ensure that all measurements are on a similar scale. This helps the model learn more effectively since SVM is sensitive to the scale of the data.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(x)
```

```
StandardScaler()
```

```
standard_data = scaler.transform(x)
standard_data
```

```
array([[ 0.63994726,  0.84832379,  0.14964075, ...,  0.20401277,
         0.46849198,  1.4259954 ],
       [-0.84488505, -1.12339636, -0.16054575, ..., -0.68442195,
        -0.36506078, -0.19067191],
       [ 1.23388019,  1.94372388, -0.26394125, ..., -1.10325546,
```

```
              0.60439732, -0.10558415],
            ...,
            [ 0.3429808 ,  0.00330087,  0.14964075, ..., -0.73518964,
             -0.68519336, -0.27575966],
            [-0.84488505,  0.1597866 , -0.47073225, ..., -0.24020459,
             -0.37110101,  1.17073215],
            [-0.84488505, -0.8730192 ,  0.04624525, ..., -0.20212881,
             -0.47378505, -0.87137393]])
```

```
x = standard_data
x
```

```
array([[ 0.63994726,  0.84832379,  0.14964075, ...,  0.20401277,
         0.46849198,  1.4259954 ],
       [-0.84488505, -1.12339636, -0.16054575, ..., -0.68442195,
        -0.36506078, -0.19067191],
       [ 1.23388019,  1.94372388, -0.26394125, ..., -1.10325546,
         0.60439732, -0.10558415],
       ...,
       [ 0.3429808 ,  0.00330087,  0.14964075, ..., -0.73518964,
        -0.68519336, -0.27575966],
       [-0.84488505,  0.1597866 , -0.47073225, ..., -0.24020459,
        -0.37110101,  1.17073215],
       [-0.84488505, -0.8730192 ,  0.04624525, ..., -0.20212881,
        -0.47378505, -0.87137393]])
```

```
y = data['Outcome']
y
```

	Outcome
0	1
1	0
2	1
3	0
4	1
...	...
763	0
764	0
765	0
766	1
767	0

768 rows × 1 columns

dtype: int64

DATA TRAINING

```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,stratify=y,random_state=2)
```

```
x_train.shape
```

```
(614, 8)
```

```
y_train.shape
```

```
(614,)
```

```
x_test.shape
```

```
(154, 8)
```

```
y_test.shape
```

```
(154,)
```

DATA MODEL TRAINING

We used a Support Vector Machine (SVM) classifier with a linear kernel to train the model on the training data.


```
from sklearn import svm

classifier = svm.SVC(kernel='linear')
classifier.fit(x_train,y_train)
```

```
SVC
SVC(kernel='linear')
```

Why SVM Was Used Support Vector Machine (SVM) was chosen for this model for several reasons:

Effective in High Dimensions: SVM is particularly powerful for classification problems in high-dimensional spaces, which is often the case with medical datasets where many features can be present.

Robust to Overfitting: Especially with a linear kernel, SVM can avoid overfitting when the dataset is not too large. This makes it a good choice for smaller datasets like the diabetes dataset.

Clear Margin of Separation: SVM finds the hyperplane that best separates the classes (diabetic vs. non-diabetic), which can lead to good classification performance.

Flexibility: By using different kernels (like linear, polynomial, or RBF), SVM can be adapted to a variety of classification problems.

```
x_pred = classifier.predict(x_train)
x_pred
```

```
array([0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0,
       1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1,
       0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0,
       0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1,
       0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0,
       0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
       0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0])
```

```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(x_pred,y_train)
print('Accuracy - ',accuracy)
```

```
Accuracy -  0.7866449511400652
```

MAKING A PREDICTIVE SYSTEM

```
df = (7,244,64,46,88,24.6,0.34,45)

df_np = np.asarray(df)
df_np
```

```
array([  7.  , 244.  ,  64.  ,  46.  ,  88.  ,  24.6 ,   0.34,  45.  ])
```

```
df_r = df_np.reshape(1,-1)
df_r
```

```
array([[  7.  , 244.  ,  64.  ,  46.  ,  88.  ,  24.6 ,   0.34,  45.  ]])
```

```
std= scaler.transform(df_r)
std
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but StandardScaler was fi
  warnings.warn(
array([[ 0.93691372,  3.85284975, -0.26394125,  1.5972786 ,  0.07120427,
        -0.93826044, -0.39828208,  1.00055664]])
```

```
pred = classifier.predict(std)
pred
```

```
array([1])
```

```
if pred[0] == 0:
  print('Non-Diabetic')
else:
  print('Diabetic')
```

```
Diabetic
```

we successfully built a diabetes prediction model using a Support Vector Machine classifier. After standardizing the data and training the model, we were able to predict whether an individual is likely to be diabetic based on their health measurements.

The use of SVM allowed us to effectively handle the classification task, leveraging its strengths in high-dimensional data and its robust performance. The model's predictions can help in identifying individuals at risk of diabetes, leading to timely interventions and better health outcomes.

By continually refining the model, perhaps through hyperparameter tuning or experimenting with different kernels, we could further improve its accuracy and reliability.