AI Based Diabetes Prediction System.

Loading and preprocessing the dataset: Phase_3

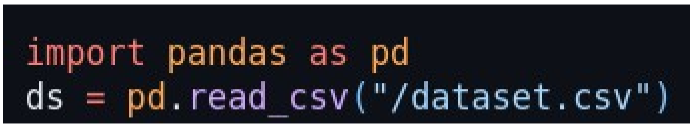
1 .Download the dataset:

Download the kaggle dataset given using the link https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

2 .Data exploration:

After downloading the dataset from kaggle.The dataset format is in the format .csv file.

Start exploring the dataset using the python libraries like pandas.



```
import pandas as pd
ds = pd.read_csv("/dataset.csv")
```

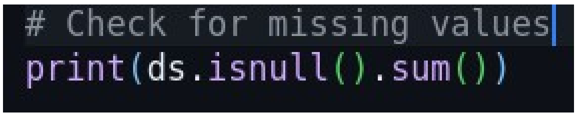
3 .Data preprocessing:

Perform the data-preprocessing using the previously explored data-set.

Preprocessing includes the handling of missing values, feature engineering and data scaling.

* Missing values-

Check for missing values and decide whether to impute them or remove rows/columns with missing data.

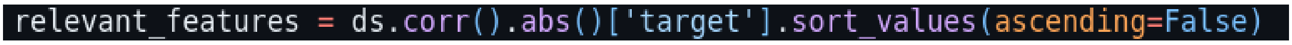


```
# Check for missing values
print(ds.isnull().sum())
```

* Feature selection-

Identify the features that are relevant for diabetes prediction. This might require domain knowledge or statistical analysis.

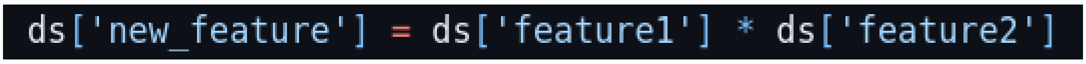
You can use techniques like correlation analysis to determine feature importance.



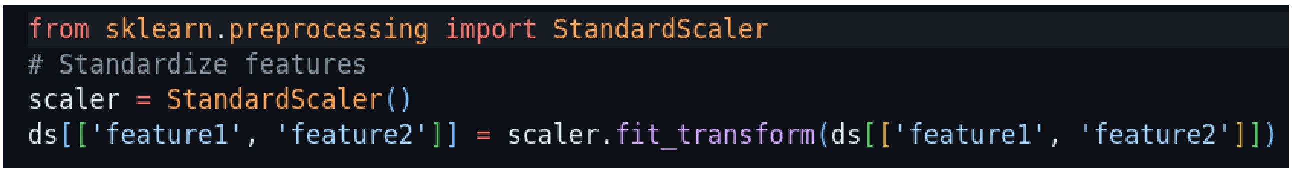
```
relevant_features = ds.corr().abs()['target'].sort_values(ascending=False)
```

* Feature engineering-

Create new features or transform existing ones that might be useful for the prediction task. This can involve mathematical operations or encoding categorical variables.



```
ds['new_feature'] = ds['feature1'] * ds['feature2']
```

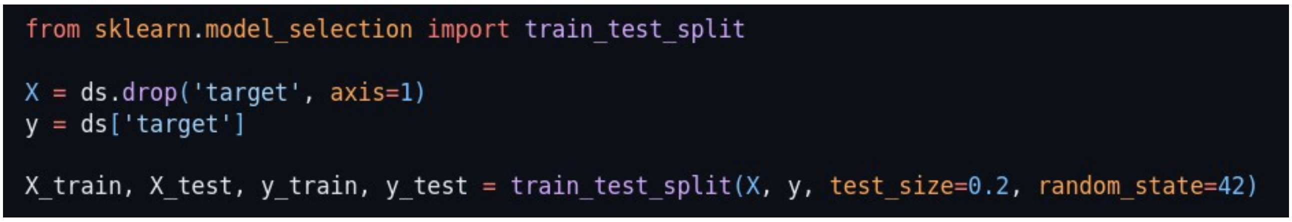


```
from sklearn.preprocessing import StandardScaler
# Standardize features
scaler = StandardScaler()
ds[['feature1', 'feature2']] = scaler.fit_transform(ds[['feature1', 'feature2']])
```

4 .Data splitting:

Slit the data into training and testing sets for model evaluation .

The typical split ratio is 70-80% for training & 20-30% for testing.



```
from sklearn.model_selection import train_test_split

X = ds.drop('target', axis=1)
y = ds['target']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```