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

from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, export_text
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix


from google.colab import files
uploaded = files.upload()

df = pd.read_csv("KaggleV2-May-2016.csv")
df.head()
```

Choose Files KaggleV2-May-2016.csv

- **KaggleV2-May-2016.csv**(text/csv) - 10739535 bytes, last modified: 9/20/2019 - 100% done

```
Saving KaggleV2-May-2016.csv to KaggleV2-May-2016.csv
```

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0

```
print(df.shape)
print(df.info())
print(df.head())

df.columns = df.columns.str.strip()
```

```
(110527, 14)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64  
 2   Gender          110527 non-null  object 
 3   ScheduledDay    110527 non-null  object 
 4   AppointmentDay  110527 non-null  object 
 5   Age             110527 non-null  int64  
 6   Neighbourhood   110527 non-null  object 
 7   Scholarship     110527 non-null  int64  
 8   Hipertension    110527 non-null  int64  
 9   Diabetes        110527 non-null  int64  
 10  Alcoholism      110527 non-null  int64  
 11  Handcap         110527 non-null  int64  
 12  SMS_received    110527 non-null  int64  
 13  No-show         110527 non-null  object 
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
None
      PatientId  AppointmentID Gender          ScheduledDay  \
0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z   
1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z   
2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z   
3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z   
4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z   

         AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1   
1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0   
2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0   
3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0   
4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1   
```

```
   Diabetes  Alcoholism  Handcap  SMS_received No-show
0         0           0        0             0      No
1         0           0        0             0      No
2         0           0        0             0      No
3         0           0        0             0      No
4         1           0        0             0      No
```

```
df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'])
df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])


df['LeadTime'] = (df['AppointmentDay'] - df['ScheduledDay']).dt.days
df = df[df['LeadTime'] >= 0]


df['No-show'] = df['No-show'].map({'Yes': 1, 'No': 0})


df['AppointmentWeekday'] = df['AppointmentDay'].dt.weekday
df['ScheduledWeekday'] = df['ScheduledDay'].dt.weekday

df = pd.get_dummies(df, columns=['Gender','Neighbourhood'], drop_first=True)

df.head()
```

	PatientId	AppointmentID	ScheduledDay	AppointmentDay	Age	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	...	Neighl
5	9.598513e+13	5626772	2016-04-27 08:36:51+00:00	2016-04-29 00:00:00+00:00	76	0	1	0	0	0	...	
6	7.336882e+14	5630279	2016-04-27 15:05:12+00:00	2016-04-29 00:00:00+00:00	23	0	0	0	0	0	...	
7	3.449833e+12	5630575	2016-04-27 15:39:58+00:00	2016-04-29 00:00:00+00:00	39	0	0	0	0	0	...	
9	7.812456e+13	5629123	2016-04-27 12:48:25+00:00	2016-04-29 00:00:00+00:00	19	0	0	0	0	0	...	
10	7.345362e+14	5630213	2016-04-27 14:58:11+00:00	2016-04-29 00:00:00+00:00	30	0	0	0	0	0	...	

5 rows × 95 columns

```
features = [col for col in df.columns if col not in ['PatientId','AppointmentID','ScheduledDay','AppointmentDay','No-show']]
X = df[features]
y = df['No-show']

print("Feature shape:", X.shape)
print("Target distribution:\n", y.value_counts())
```

```
Feature shape: (71959, 90)
Target distribution:
 No-show
0    51437
1    20522
Name: count, dtype: int64
```

```
# Step 6: Split dataset
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)

print("Train size:", X_train.shape)
print("Test size:", X_test.shape)
```

```
Train size: (50371, 90)
Test size: (21588, 90)
```

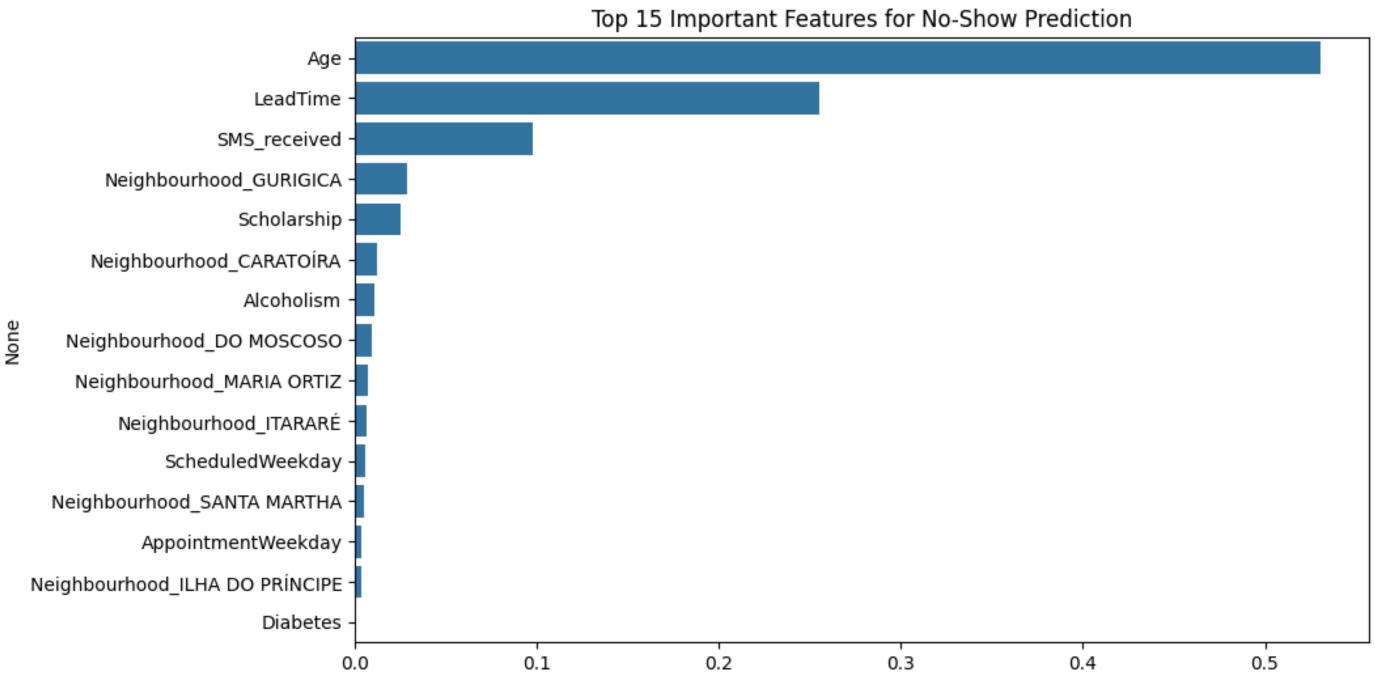
```
clf = DecisionTreeClassifier(max_depth=5, random_state=42, class_weight='balanced')
clf.fit(X_train, y_train)

print("Model trained successfully!")
```

```
Model trained successfully!
```

```
y_pred = clf.predict(X_test)

print("✅ Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

```
✅ Accuracy: 0.5669816564758199

Classification Report:
               precision    recall  f1-score   support

           0       0.77      0.57      0.65     15431
           1       0.34      0.57      0.43      6157

    accuracy                           0.57     21588
   macro avg       0.55      0.57      0.54     21588
weighted avg       0.65      0.57      0.59     21588


Confusion Matrix:
 [[8761 6670]
 [2678 3479]]
```

```
importances = pd.Series(clf.feature_importances_, index=X_train.columns).sort_values(ascending=False)[:15]

plt.figure(figsize=(10,6))
sns.barplot(x=importances.values, y=importances.index)
plt.title("Top 15 Important Features for No-Show Prediction")
plt.show()
```



```
print(export_text(clf, feature_names=list(X_train.columns)))
```

```
|   |   |   |   |   |   class: 1
|   |   |   |   |   |--- class: 1
|   |   |   |--- SMS_received >  0.50
|   |   |   |   |--- AppointmentWeekday <= 2.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- AppointmentWeekday >  2.50
|   |   |   |   |   |--- class: 0
|   |   |--- Age >  54.50
|   |   |   |--- LeadTime <= 0.50
|   |   |   |   |--- Neighbourhood_SANTA MARTHA <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- Neighbourhood_SANTA MARTHA >  0.50
|   |   |   |   |   |--- class: 0
|   |   |   |--- LeadTime >  0.50
|   |   |   |   |--- Age <= 89.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- Age >  89.50
|   |   |   |   |   |--- class: 1
|   |--- LeadTime >  6.50
|   |   |--- SMS_received <= 0.50
|   |   |   |--- Neighbourhood_GURIGICA <= 0.50
|   |   |   |   |--- Age <= 52.50
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- Age >  52.50
|   |   |   |   |   |--- class: 0
|   |   |   |--- Neighbourhood_GURIGICA >  0.50
|   |   |   |   |--- ScheduledWeekday <= 2.50
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- ScheduledWeekday >  2.50
|   |   |   |   |   |--- class: 0
|   |   |--- SMS_received >  0.50
|   |   |   |--- Age <= 59.50
|   |   |   |   |--- Scholarship <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- Scholarship >  0.50
|   |   |   |   |   |--- class: 1
|   |   |   |--- Age >  59.50
|   |   |   |   |--- Neighbourhood_DO MOSCOSO <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- Neighbourhood_DO MOSCOSO >  0.50
|   |   |   |   |   |--- class: 1
```

```
import joblib
joblib.dump(clf, "decision_tree_noshow.pkl")
print("Model saved as decision_tree_noshow.pkl")
```

```
Model saved as decision_tree_noshow.pkl
```

```
df.to_csv("cleaned_appointments.csv", index=False)

files.download("cleaned_appointments.csv")
```