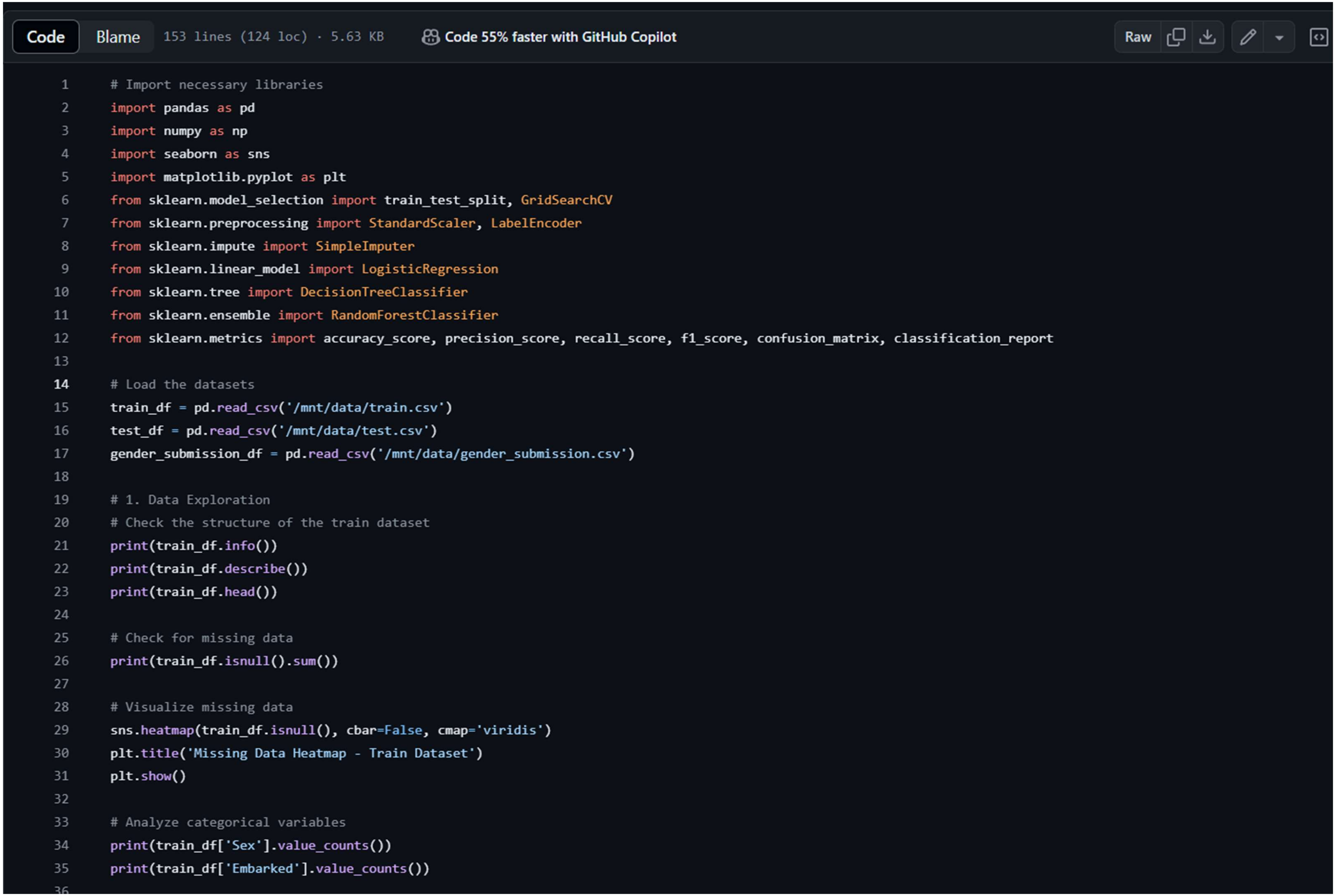
Code
Blame
153 lines (124 loc) · 5.63 KB
Code 55% faster with GitHub Copilot
Raw

```python
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report

# Load the datasets
train_df = pd.read_csv('/mnt/data/train.csv')
test_df = pd.read_csv('/mnt/data/test.csv')
gender_submission_df = pd.read_csv('/mnt/data/gender_submission.csv')

# 1. Data Exploration
# Check the structure of the train dataset
print(train_df.info())
print(train_df.describe())
print(train_df.head())

# Check for missing data
print(train_df.isnull().sum())

# Visualize missing data
sns.heatmap(train_df.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Data Heatmap - Train Dataset')
plt.show()

# Analyze categorical variables
print(train_df['Sex'].value_counts())
print(train_df['Embarked'].value_counts())

```