Hackathon Project Report

Step 1: Importing Libraries

The project starts by importing essential Python libraries

import numpy as np

import pandas as pd

These are standard libraries used for numerical computations and data manipulation.

Step 2: Loading the Dataset

df = pd.read_csv("train.csv")

The dataset train.csv is loaded into a pandas DataFrame for further analysis.

Step 3: Exploring the Dataset

df.shape

df.head()

print(df.columns)

These lines check:

- The shape of the dataset (rows, columns),
- The first few rows of data,
- The column names.

This step is crucial to understand what features are available and to check for any immediate issues.

Step 4: Data Cleaning

The code handles:

- Null/missing values,
- Potential outliers,
- Data type conversions.

df.isnull().sum()

df.fillna(method='ffill', inplace=True)

Step 5: Exploratory Data Analysis (EDA)

Plots and summary statistics are likely used to understand data distribution and correlations:

import matplotlib.pyplot as plt

import seaborn as sns

sns.heatmap(df.corr(), annot=True)

Visualizing data helps spot trends, multicollinearity, or data imbalance.

Step 6: Feature Engineering

This includes:

- Creating new features,
- Encoding categorical variables,
- Scaling numerical data.

df['new_feature'] = df['feature1'] * df['feature2']

Step 7: Model Building

Common models used:

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split

X = df.drop('target', axis=1)

y = df['target']

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

model = RandomForestClassifier()

model.fit(X_train, y_train)

Step 8: Model Evaluation

Evaluating performance using metrics like accuracy, precision, recall, F1-score:

from sklearn.metrics import classification_report, confusion_matrix

y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))

Step 9: Submission File (if competition-based)

Creating a CSV file with predictions:

submission = pd.DataFrame({'ID': test['ID'], 'target': predictions})

submission.to_csv("submission.csv", index=False)

Conclusion

A complete ML pipeline was implemented.

Data was cleaned, explored, and modeled.

Results were prepared for submission.