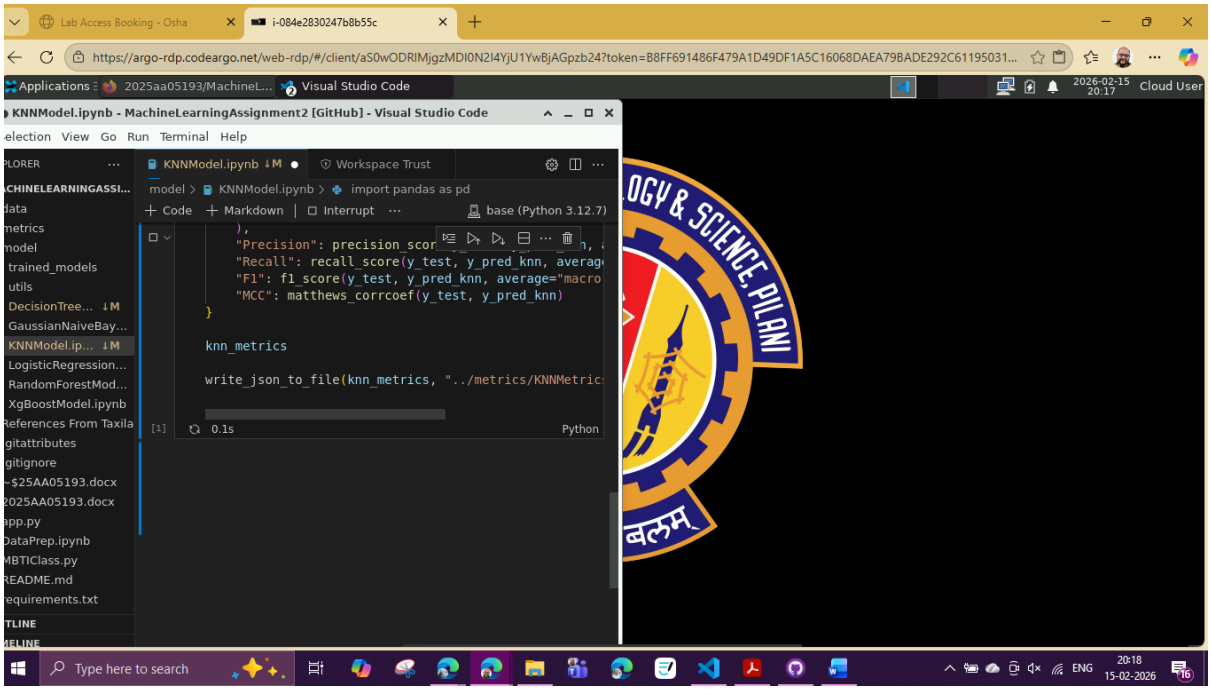
Name: Siddharth Gupta
BITS ID: 2025AA05193
Email: 2025aa05193@wilp.bits-pilani.ac.in

1. **Git Repository Link:** https://github.com/2025aa05193/MachineLearningAssignment2
2. **Live Streamlit App Link**: https://2025aa05193-mbtipersonaityclassification.streamlit.app/
3. **Screenshot for executing in BITS Virtual Lab:**



4. **Readme.md – Content**

MBTI Classification Based on Survey - Machine Learning Model Comparison

a. Problem Statement

The objective is to develop and evaluate multiple machine learning classifiers to accurately predict MBTI personality types from survey responses in the 60k MBTI dataset, and to compare model performance using standard evaluation metrics.

b. Dataset Description

Dataset URL: https://www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt/data

Description: It contains the questions from the 16 Personality Tests and their answers in the Scale that they use but is numerically encoded:

Fully Agree: 3

Partially Agree: 2

Slightly Agree: 1

neutral -> 0

Slightly disagree: -1

Partially disagree: -2

Fully disagree: -3

c. Models Used and Evaluation Metrics

Model Performance Comparison

ML Model Name	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.9191	0.9932	0.9194	0.9191	0.9192	0.9137
Decision Tree	0.6476	0.8791	0.6482	0.6475	0.6474	0.6241
kNN	0.9869	0.9948	0.9869	0.9869	0.9869	0.9860
Naive Bayes	0.9114	0.9925	0.9119	0.9114	0.9113	0.9056
Random Forest (Ensemble)	0.9773	0.9941	0.9773	0.9773	0.9773	0.9758
XGBoost (Ensemble)	0.9824	0.9950	0.9825	0.9824	0.9824	0.9813

Model Performance Observations

ML Model Name	Observation about model performance
Logistic Regression	Logistic Regression provides a strong baseline model with stable performance across all metrics. Its high AUC and balanced Precision-Recall indicate reliable predictive performance. However, it is outperformed by non-linear and ensemble models, indicating that the dataset likely contains complex feature interactions.

ML Model Name	Observation about model performance
Decision Tree	The Decision Tree model shows the weakest performance among all models, with significantly lower Accuracy, F1, and MCC scores. While the AUC is reasonably high (0.8791), the drop in overall classification metrics suggests overfitting and poor generalization compared to ensemble methods.
kNN	kNN achieves the highest overall Accuracy and MCC among all models, indicating superior predictive performance and strong class separation. The consistently high metrics suggest that the dataset benefits from local neighborhood-based decision boundaries.
Naive Bayes	Naive Bayes performs strongly with excellent AUC and balanced Precision-Recall values. The high AUC (0.9925) indicates strong class separability, but its overall Accuracy and MCC are lower than kNN and ensemble models, suggesting limitations due to its independence assumption.
Random Forest (Ensemble)	Random Forest delivers excellent performance with strong generalization ability. The high Accuracy and MCC demonstrate robustness and reduced overfitting compared to a single Decision Tree. Ensemble learning significantly improves predictive stability.
XGBoost (Ensemble)	XGBoost performs extremely well across all metrics, slightly below kNN in Accuracy but showing the highest AUC. Its strong MCC indicates reliable performance even in complex decision boundaries. It demonstrates powerful learning capability through boosting and feature interaction modeling.