MBTI Prediction

Chenrui Zhang

Department of Engineering

https://github.com/RickchenruiforGitHub/Data1030_Final

1.Introduction

Knowing someone's **MBTI (Myers-Briggs Type Indicator)** type helps people understand their preferences, behaviors, and motivations. This understanding can lead to better relationships, improved collaboration, and success in career path. The MBTI categorizes individuals into **16 personality types**:

- **Extraversion (E) vs. Introversion (I)**: Orientation of energy (external vs. internal focus).
- **Sensing (S) vs. Intuition (N)**: Information processing style (details vs. patterns).
- **Thinking (T) vs. Feeling (F)**: Decision-making process (logic vs. emotions).
- **Judging (J) vs. Perceiving (P)**: Lifestyle preference (structured vs. flexible).

Obviously, it's a multiclassification problem**.** The data set I used in my final project comes from**https://www.kaggle.com/datasets/stealthtechnologies/predict-people-personality-types/.**

There are a couple of predictions made by Kaggle users based on the same dataset. For example. User Joshua kab [1]achieved a 89.7% prediction accuracy using XGBoost; user Samanyu K[2] achieved 90.2% accuracy using Random Forest Classifier..

2. EDA

Before training the model, conducting exploratory data analysis (EDA) is a crucial step to gain deeper insights into the dataset. As part of this process, I visualized several key features to better understand their distributions and relationships.

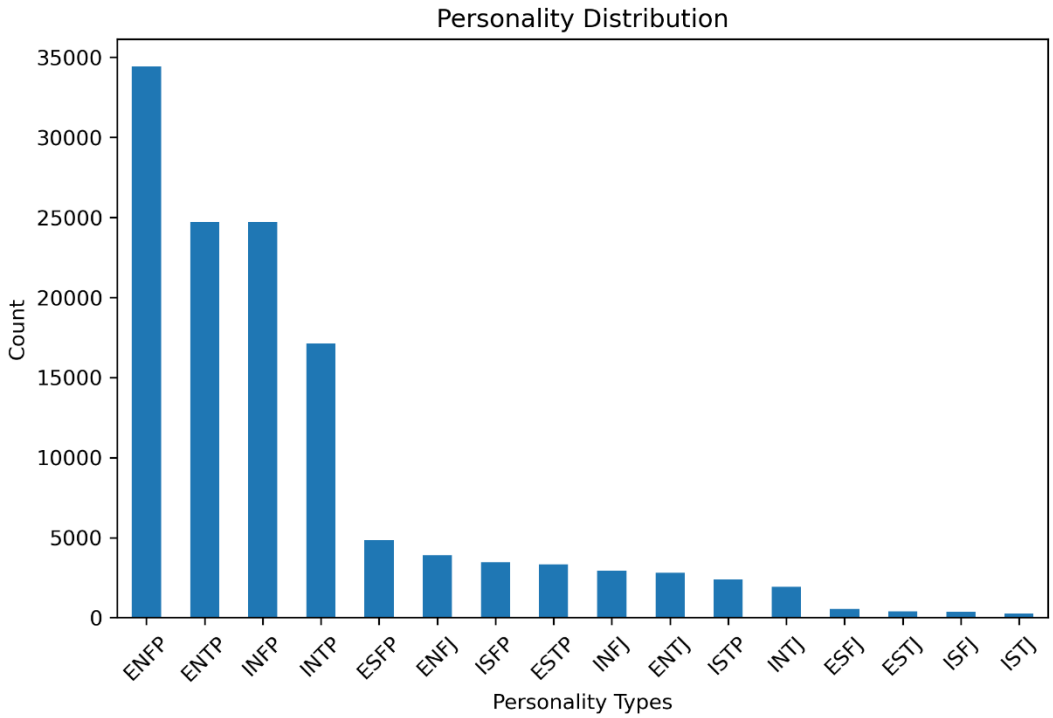

Fig 1. Personality Distribution

The figure shows that this is an unbalanced label. **ENFP**, **ENTP**, **INFP**, and **INTP** are the most frequent personality types, with **ENFP** being significantly higher than the others, exceeding 35,000 samples. Minority classes like **ISTJ** only occupies 0.2% of the whole samples. This imbalance needs a careful approach during the training process to ensure fair representation of all classes.

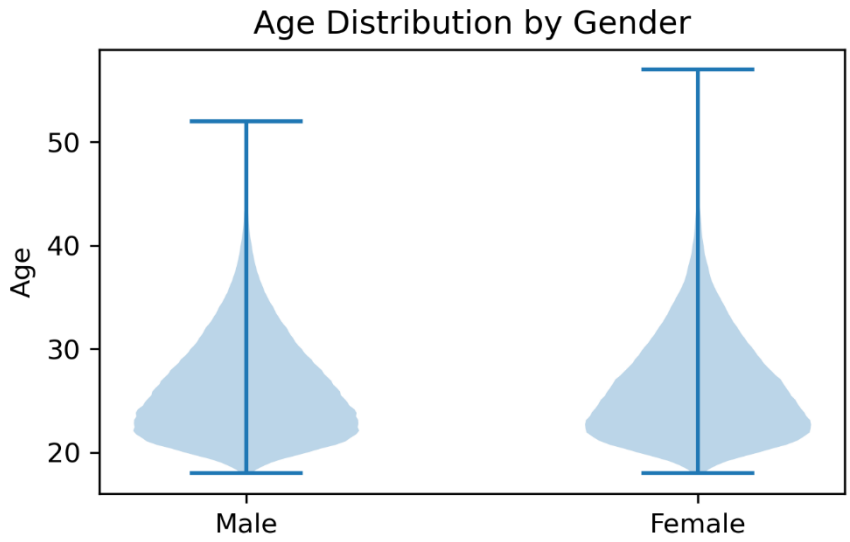

Fig 2: Age Distribution by Gender

The similarity in distribution suggests that age does not vary significantly by gender in this dataset.

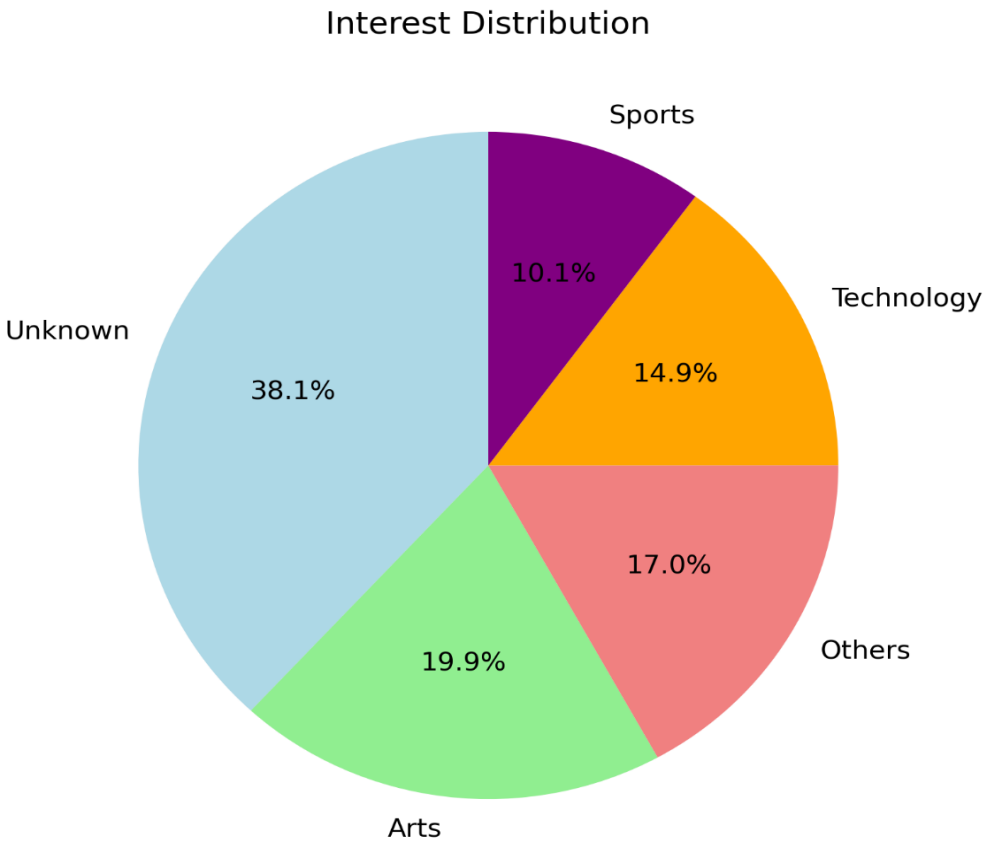

Fig3 . Interest Distribution

Feature “Interest” contains a missing value represented as”Unknown” with a proportion of 38.1% Below is the distribution of personality.

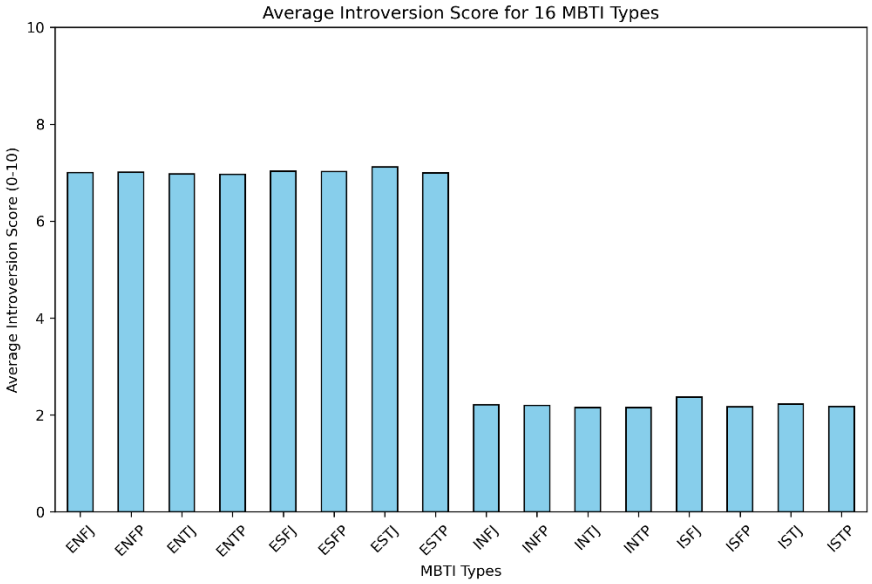

Fig 4. MBTI Introversion Score

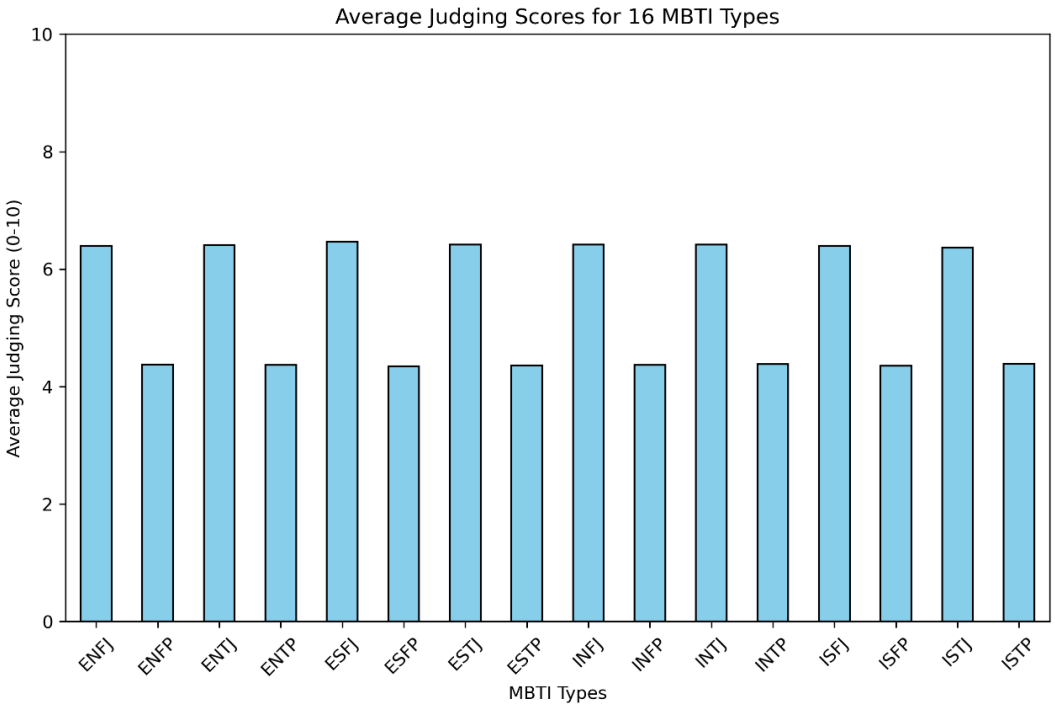

Fig 5. MBTI Judging Score

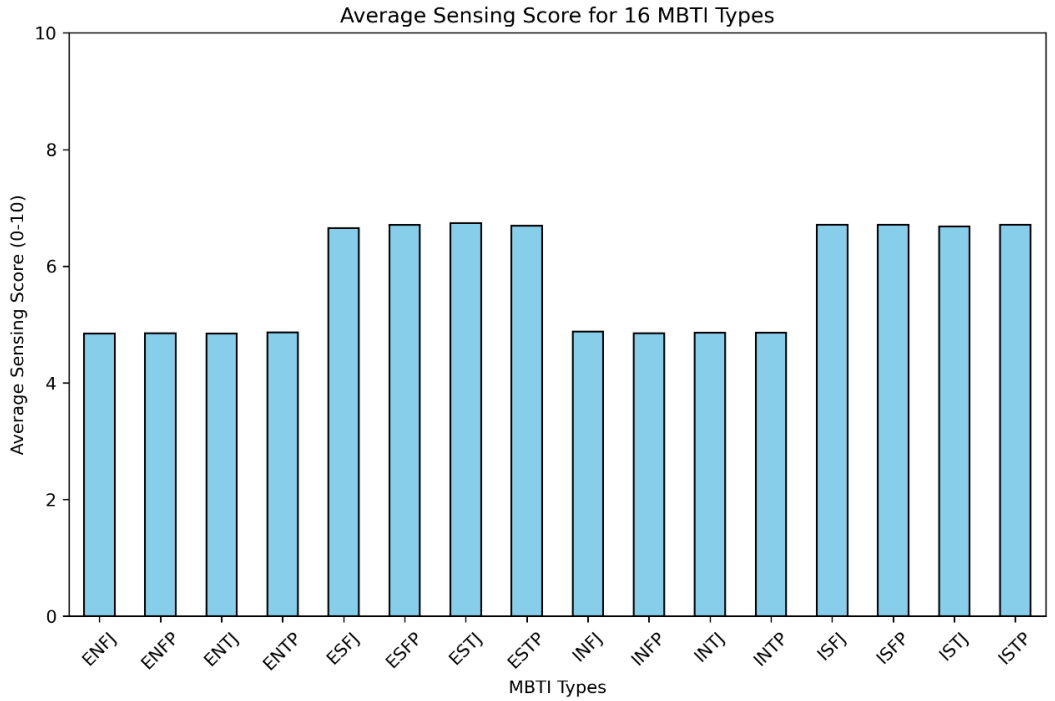

Fig 6. MBTI Sensing Score

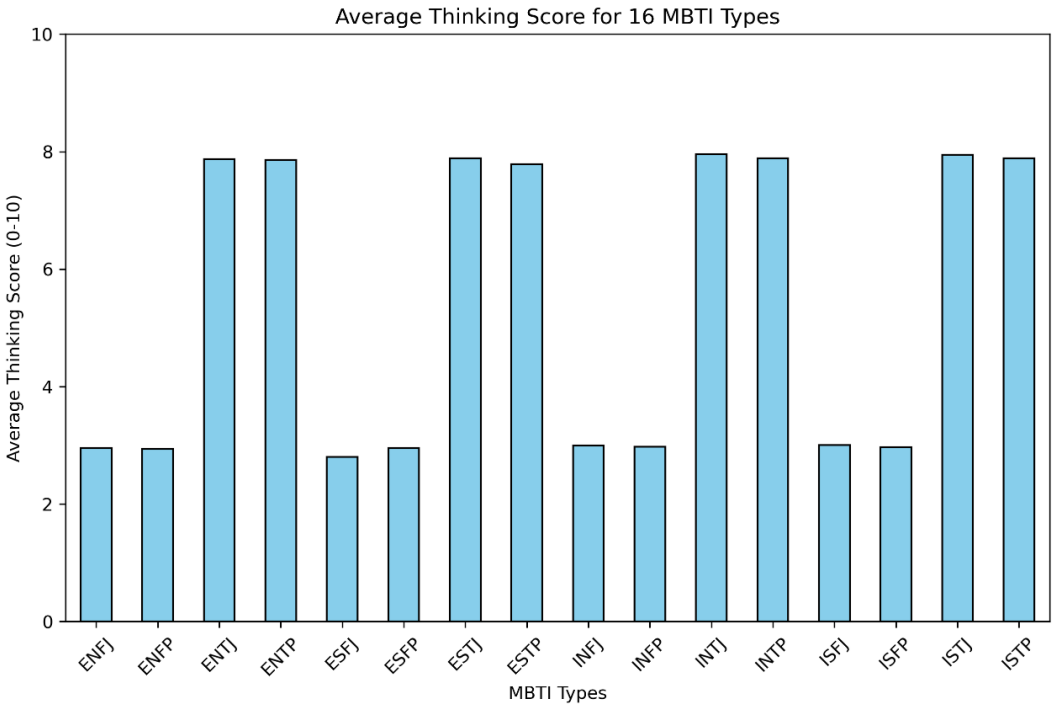

Fig 7. MBTI Thinking Score

The visualizations show how key MBTI traits vary across 16 personality types. **Introversion scores** clearly separate introverts (e.g., INFP, INTJ) from extroverts (e.g., ENFJ, ENFP). **Judging scores** are more evenly distributed. **Sensing scores** are higher for sensing types (e.g., ESFJ, ISTJ) and lower for intuitive types (e.g., INFP, INTJ). **Thinking scores** are highest for thinking types (e.g., INTJ, ISTP) and lowest for feeling types (e.g., INFP, ENFP), reflecting distinct personality traits.

Then I calculated Pearson correlation matrix:

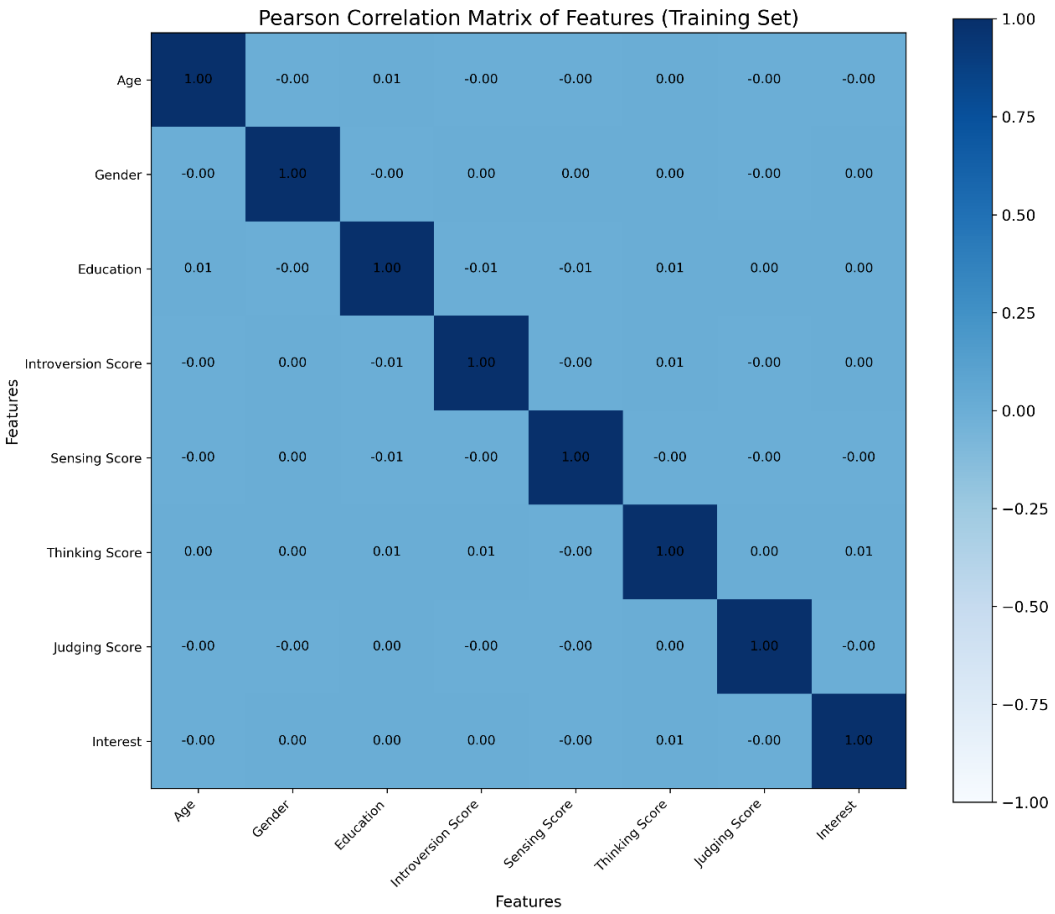

Fig 8 :Pearson Correlation Matrix

The figure indicates that multicollinearity is not a concern in this dataset, and the features likely provide independent information to the model.

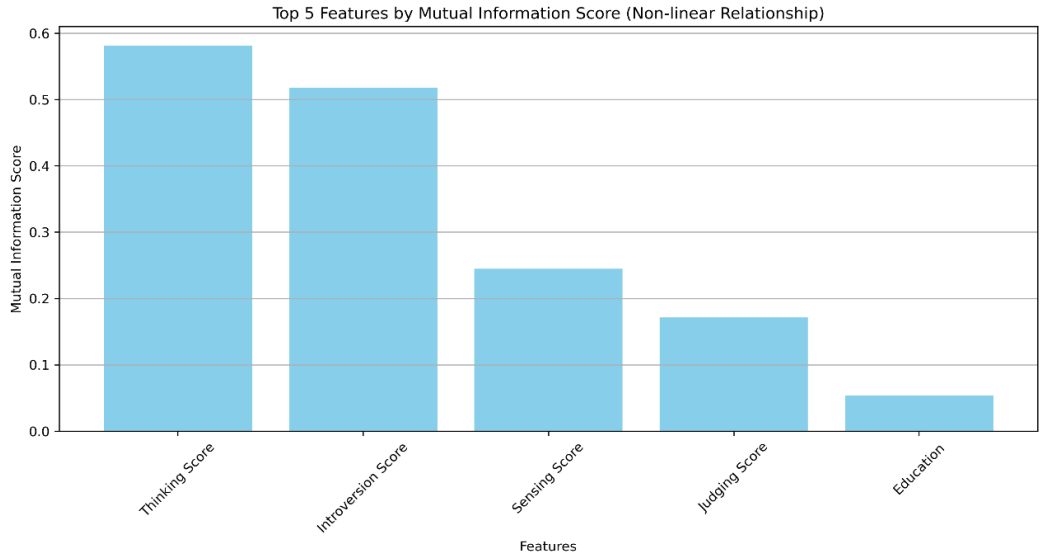

Fig 8 :Top Five Features Contributes to Mutual Information

This bar chart shows the top 5 features ranked by their Mutual Information Score. Thinking Score and Introversion Score are the most significant features, meaning they provide the most information

for predicting the targe

3.Methods

The strategy I used is StratifiedShuffleSplit, which ensures the dataset is split while preserving the class proportions in both the training and test sets, maintaining the overall class distribution. For cross-validation, I implemented k-fold StratifiedShuffleSplit with k=3.

Fig 9: Splitting Strategy

For the evaluation metric, I chose the **F1 weighted score** because it is suitable for imbalanced datasets. The F1 weighted score calculates the harmonic mean of precision and recall for each class and then weights thse scores according to the class distribution. This ensures that the metric accounts for both the model's ability to correctly classify the majority and minority classes.

$$Weighted\ F1 = \frac{\sum_{i=1}^{n} \omega_i \cdot F1_i}{\sum_{i=1}^{n} \omega_i} \quad (1)$$

I selected five machine learning algorithms to evaluate the dataset: **Random Forest**, **XGBoost**, **SVM**, **KNN**, and **Multinomial Logistic Regression**. For Random Forest, I tuned the number of trees (n_estimators) and tree depth (max_depth). For XGBoost, I explored tree depth, learning rate, and the number of estimators. SVM was tested with both linear and RBF kernels and regularization parameter (L2 regularization). KNN was evaluated by varying the number of neighbors (n_neighbors) and weighting schemes. Lastly, for Multinomial Logistic Regression, I optimized the L2 regularization parameter. These parameters were chosen to balance performance and model complexity effectively. These parameters are used for kfold cross validation to find the best model,

Algorithms	Hyper parameters
Random Forest	n_estimators: [50, 100, 200]
	max_depth: [None, 10, 20]
XGB	max_depth: [5, 10, 15, 20, 25]
	n_estimators: [50, 100, 200]
	Learning_rate: [0.01, 0.1, 0.2]
SVM	kernel: ['linear', 'rbf']
	L2 regularization: [0.01, 0.1, 1, 10, 100]
KNN	n_neighbors: [3, 5, 10, 15],

	weights: ['uniform', 'distance']
Multi-Logistic Classification	L2 regularization: [0.01, 0.1, 1, 10, 100]

Table 1: Model Parameters

To address the uncertainties in the evaluation metric and the non-deterministic nature of machine learning methods, I ran the model using five different random states(42-47), then calculated deviation

4.Result:

If we predict all the sample as ENFP, we would have an extremely low f1 score: 0.1138. Below is a summary of models I trained and their performance.

- **XGBoost** achieves the highest weighted F1 score (0.9066) and accuracy 0.9068, indicating it performs the best overall in handling imbalanced classes and achieving balanced precision and recall.
- **Random Forest** closely follows XGBoost with a weighted F1 score of 0.9051 and accuracy 0.9053, showing it is also effective for this dataset.
- **SVM** performs slightly worse than Random Forest with a score of 0.8882 but still delivers strong results.
- **Logistic Regression** achieves a weighted F1 score of 0.8611, showing it is reliable but less capable than the tree-based methods or SVM in this case.
- **KNN** performs the worst with a weighted F1 score of 0.8223, which is the worst.

Algorithms	Results
Random Forest	Mean Test Weighted F1 Score: 0.9051 Standard Deviation of Test Weighted F1 Score: 0.0009 Mean Test Accuracy: 0.9053 Standard Deviation of Test Accuracy: 0.0009
XGB	Mean Test Weighted F1 Score: 0.9066 Standard Deviation of Test Weighted F1 Score: 0.0014 Mean Test Accuracy: 0.9068 Standard Deviation of Test Accuracy: 0.0014
SVM	Mean Test Weighted F1 Score: 0.8882 Standard Deviation of Test Weighted F1 Score: 0.0013 Mean Test Accuracy: 0.8888 Standard Deviation of Test Accuracy: 0.0013
KNN	Mean Test Weighted F1 Score: 0.8223

	Standard Deviation of Test Weighted F1 Score: 0.0012 Mean Test Accuracy: 0.8266 Standard Deviation of Test Accuracy: 0.0011
Multi-Logistic Classification	Mean Test Weighted F1 Score: 0.8611 Standard Deviation of Test Weighted F1 Score: 0.0013 Mean Test Accuracy: 0.8641 Standard Deviation of Test Accuracy: 0.0013

Table 2: Training Results

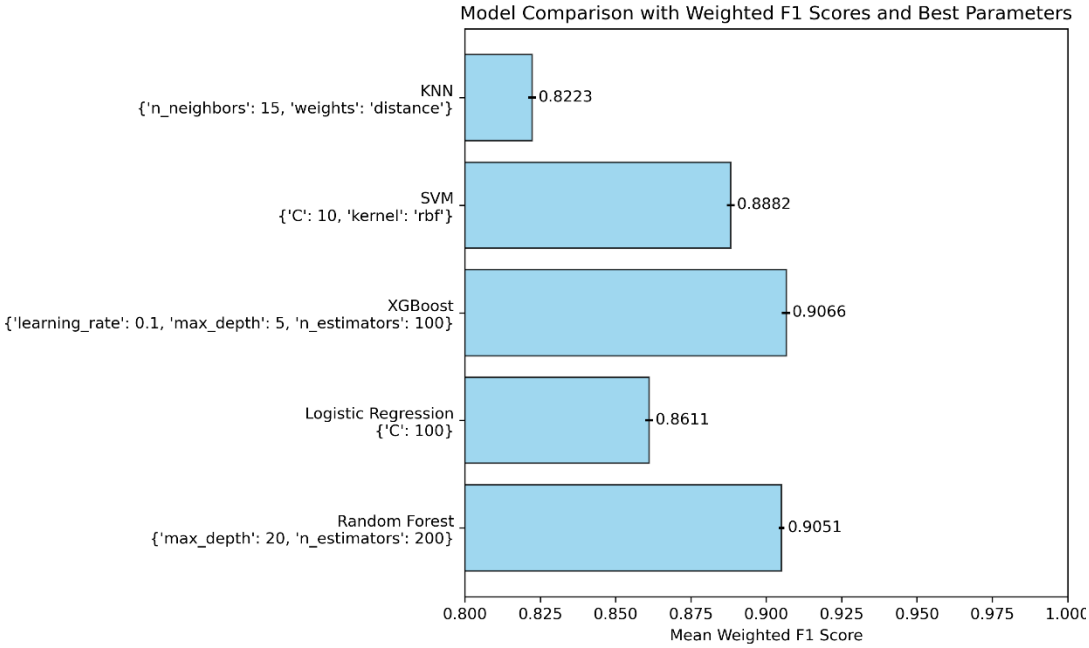

Fig 10: F1 scores comparison

I selected **XGBoost,** the most predictive model, for interpretation. **XGBoost** provides five built-in methods to visualize and analyze global feature importance effectively.

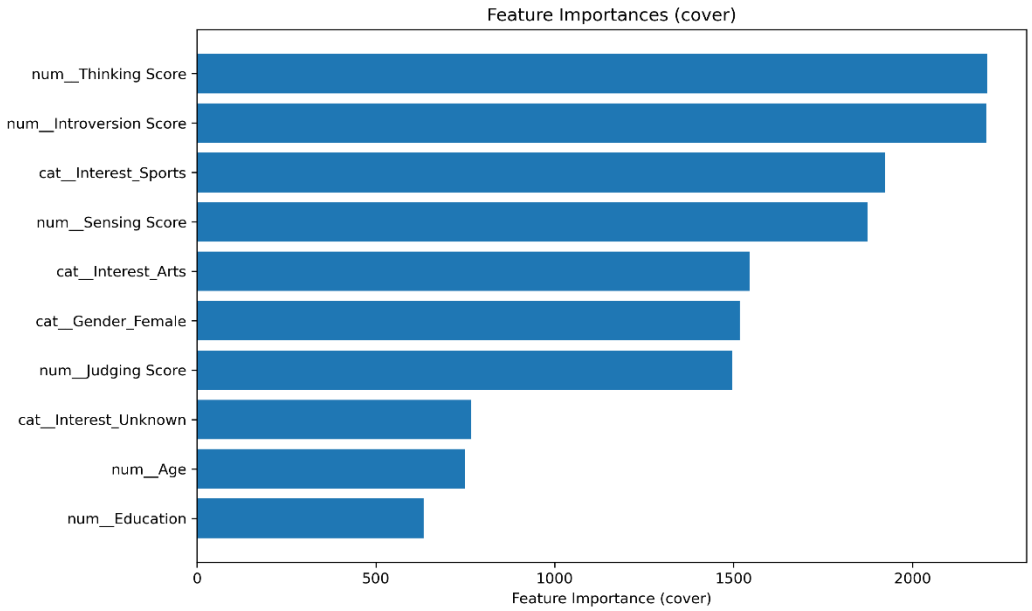

Fig 11: Feature importances(cover)

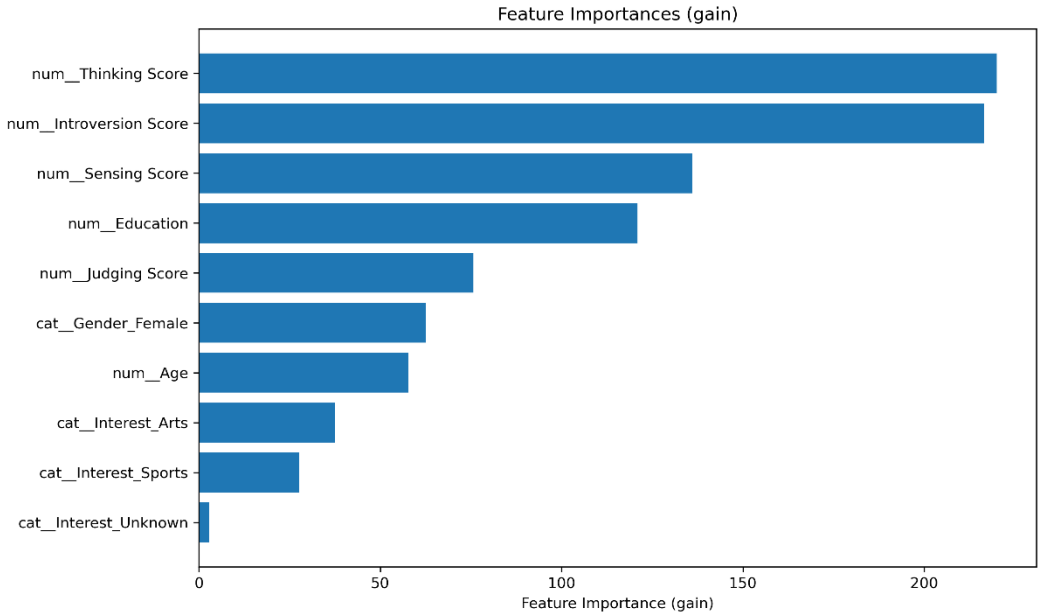

Fig 12: Feature importances(gain)

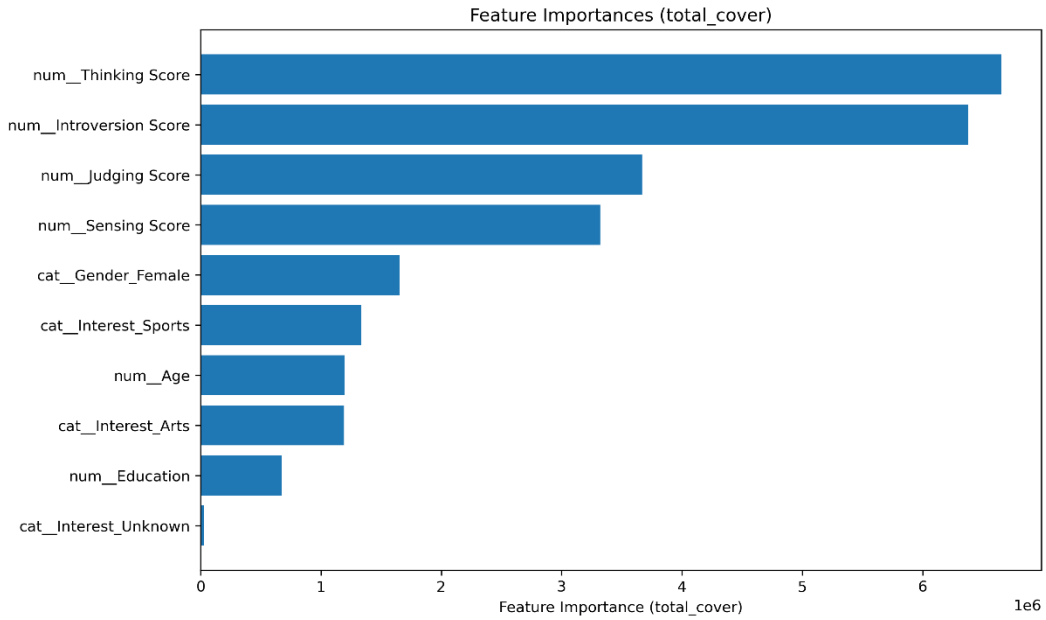

Fig 13: Feature importances(total_cover)

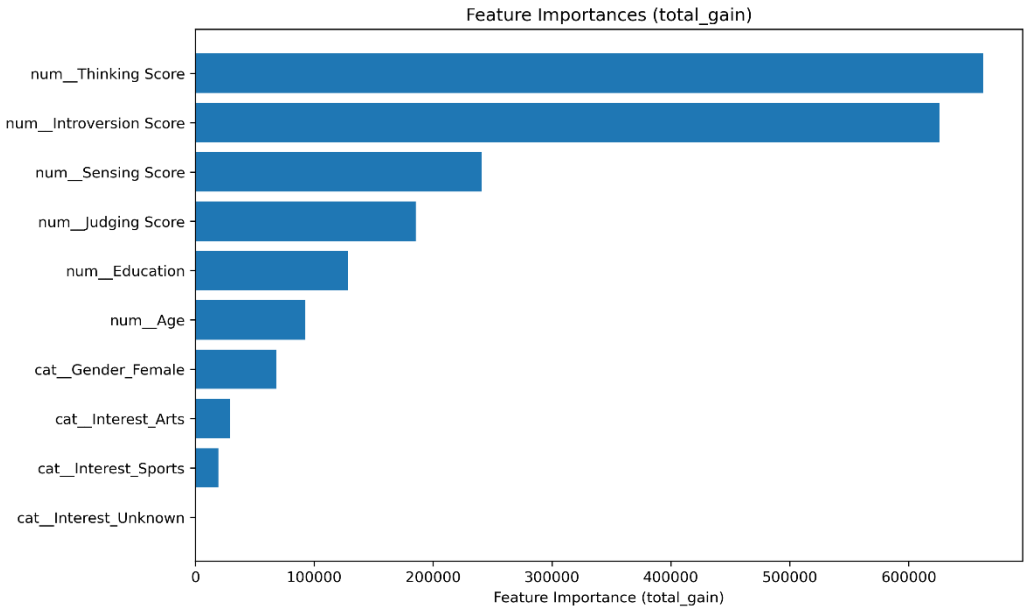

Fig 14: Feature importances(total_gain)

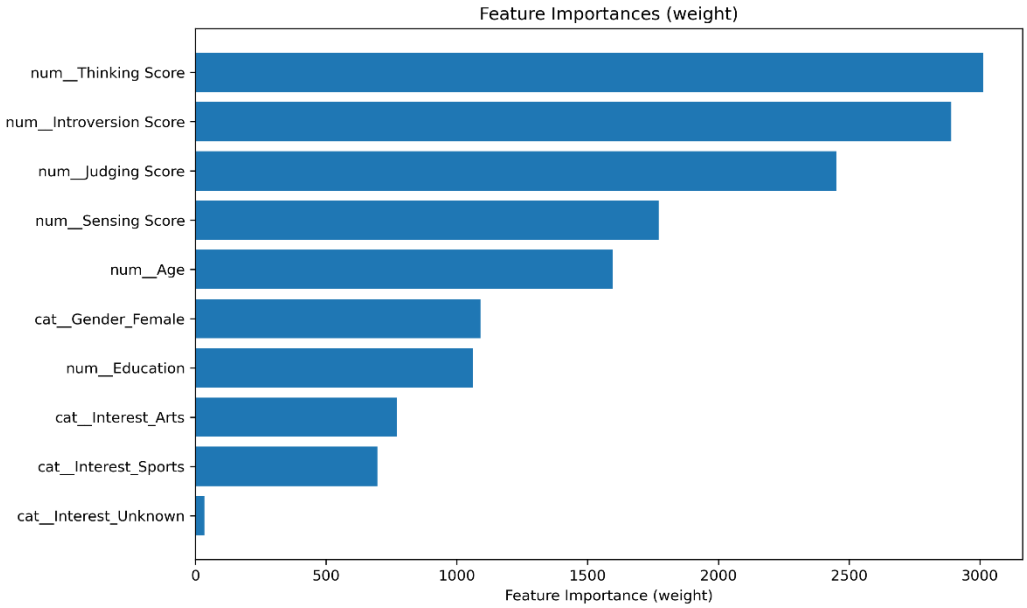

Fig 15: Feature importances(weight)

Across all metrics, the feature importance analysis highlights that Thinking Score and Introversion Score are consistently the most impactful features across all metrics, indicating their critical role in personality classification. Categorical variables like Interest_Sports, Interest_Arts, and Gender_Female also contribute significantly but are generally less influential than numerical features. Features like Sensing Score and Judging Score further support the model's performance, while less impactful features such as Interest_Unknown and Education may require additional data or representation to enhance their predictive value. Overall, the model's behavior aligns well with expectations, emphasizing cognitive and behavioral traits while capturing demographic and interest-based variations.

Then I calculated SHAP for local feature importance (feature 0) as well.

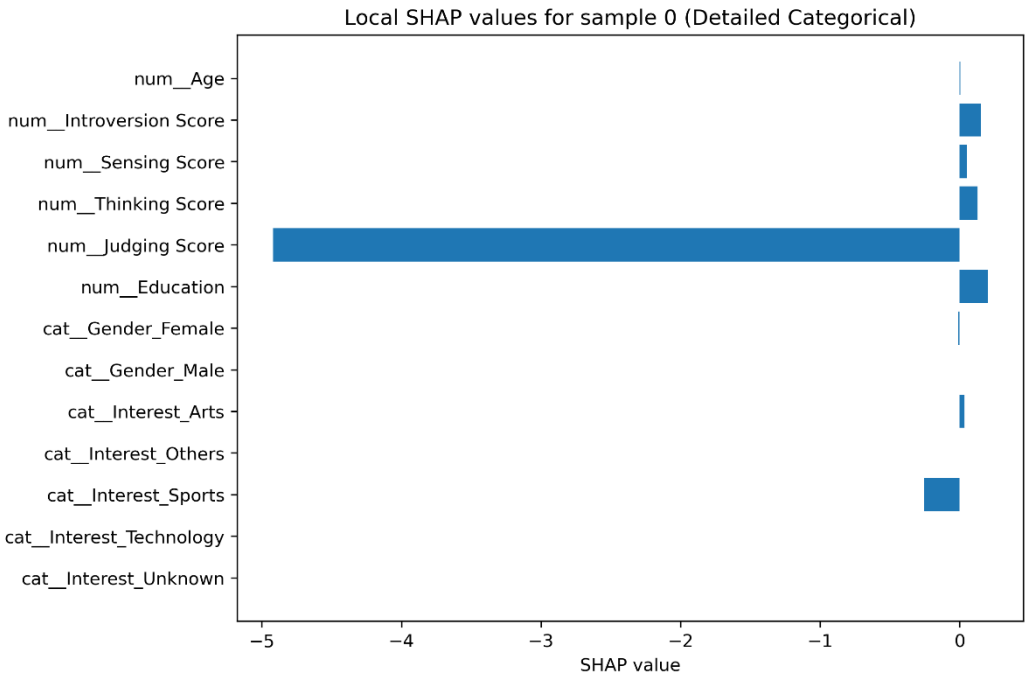

Fig 16: SHAP Values

The prediction for this sample is predominantly influenced by the **Judging Score**, highlighting its importance in determining the outcome.

Confusion matrix is a good method to evaluate the performance of model. The figure below shows strong performance for classes like ENFP and INFP but struggles with minority classes like ISTJ. The accuracy of ISTJ prediction is 84%, much lower than the total performance. What's more, there are misclassification problems between similar types (e.g., INFP as ENFP). Imbalanced data and overlapping features contribute to errors. Improvements include balancing classes, refining features, and gathering more data for underrepresented types.

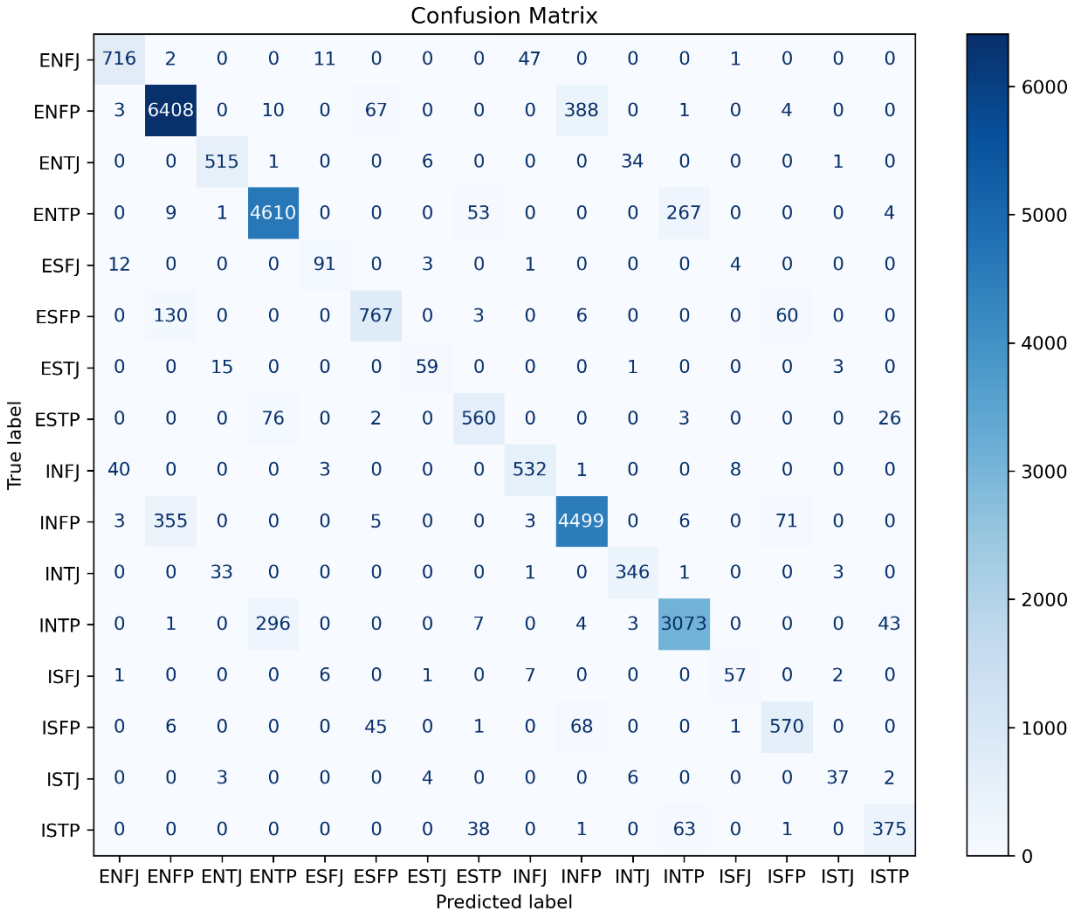

Fig 17: Confusion Matrix

5.Outlook

Although the model achieved a prediction accuracy of 90.6%, there is still room for improvement, particularly in handling minority classes like ISTJ. The model struggles to accurately predict ISTJ due to the limited representation of this type in the dataset. Collecting additional data specifically from individuals with an ISTJ MBTI type would help balance the class distribution, enabling the model to better learn the unique patterns and characteristics of this type, ultimately improving overall accuracy and fairness in predictions.

To address misclassification issues, especially for frequently confused classes, we can adopt a more targeted modeling approach. Instead of using a single model for all classes, we can train specialized models for confusing class pairs, such as ENFP-INFP or INTP-INTJ, to better capture their unique differences. Additionally, a hierarchical model structure can be implemented: a global classifier would first group samples into broader categories (e.g., intuitive vs. sensing types), followed by subgroup classifiers to distinguish between similar classes within each category. This approach enables a more nuanced understanding of the data, improving prediction accuracy for challenging classes while maintaining overall performance.

6.References

[1] https://www.kaggle.com/code/joshuakab/predict-personality-type-acc-95
[2] https://www.kaggle.com/code/samanyuk/personality-prediction-90-accuracy
Data source: **https://www.kaggle.com/datasets/stealthtechnologies/predict-people-personality-types/.**