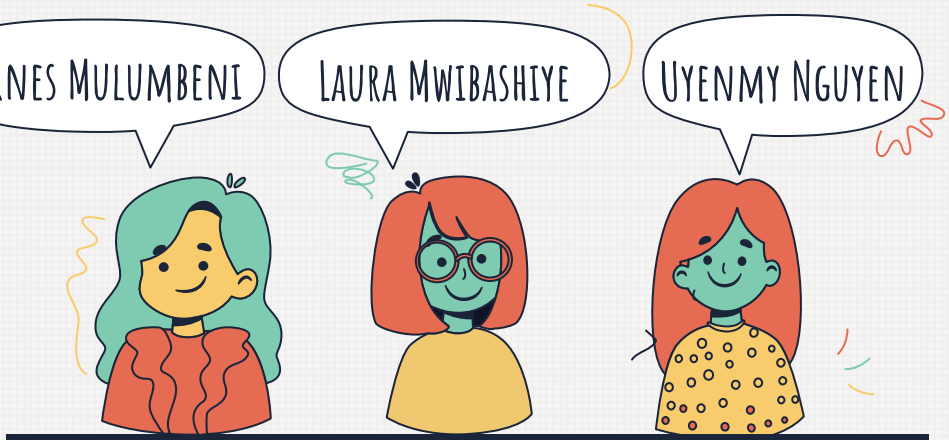
Leveraging Data Science to Predict Personality Types: An MBTI Classification Project



DATA 3421: FINAL PROJECT
MONDAY, APRIL 21ST 2025

Table of contents

Background & Motivations 01

02 Research Questions & Methodology

Exploratory Data Analysis & Data Preprocessing 03

04 ML Models & Comparison

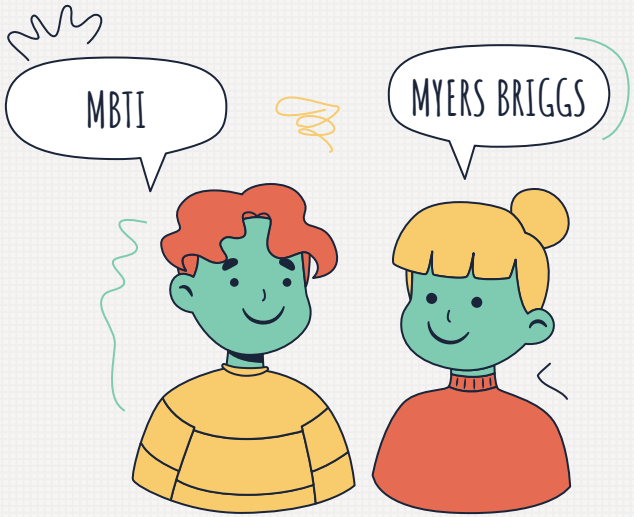
All Model Comparison & Best Model 05

06 Challenges & Future Work

01

Background & Motivations

What is our project about?



Background & Motivations

Background:

The **Myers-Briggs Type Indicator (MBTI)** is a widely used framework for classifying individuals into one of 16 personality types based on four key dimensions: Extraversion vs. Introversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving.

Motivations:

- **Advancing Machine Learning Techniques**: The project offers an opportunity to experiment with different machine learning models, data preprocessing, feature selection, and hyperparameter tuning to improve prediction accuracy.
- **Data-Driven Insights**: By automating personality prediction, we can uncover valuable insights into which features (demographic or behavioral) are most predictive of personality types, enhancing our understanding of human behavior.

Pop Quiz! What is your MBTI?

E — Extroverts
are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.

S — Sensors
are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.

T — Thinkers
tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.

J — Judgers
tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.

I — Introverts
often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.

N — Intuitives
prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.

F — Feelers
tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

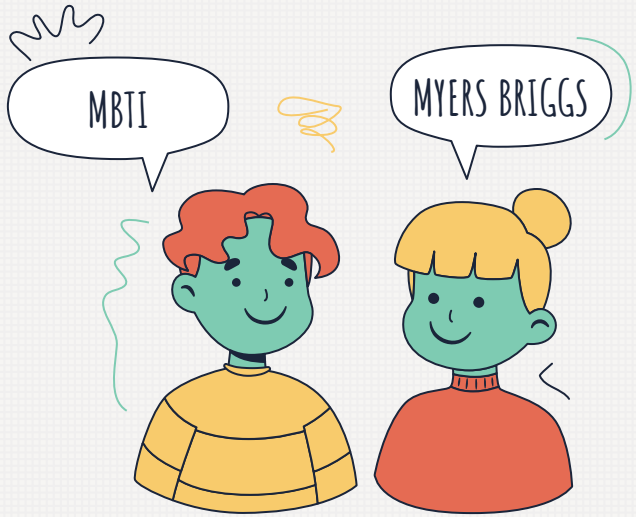
P — Perceivers
prefer to keep their options open, like to be able to act spontaneously, and like to be flexible with making plans.

THE 16 MTBI PERSONALITY TYPES

ISTJ — THE LOGISTICIAN: Practical and fact-minded individuals, whose reliability cannot be doubted	INTJ — THE ARCHITECT: Imaginative and strategic thinkers, with a plan for everything
ISFJ — THE DEFENDER: Very dedicated and warm protectors, always ready to defend their loved ones	INTP — THE LOGICIAN: Innovative inventors with an unquenchable thirst for knowledge
ESTJ — THE EXECUTIVE: Excellenct administrators, unsurpassed at managing people	ENTJ — THE COMMANDER: Bold, imaginative and strong-willed leaders, always finding a way - or making one
ESFJ — THE CONSUL: Extraordinarily caring, social and popular people, always eager to help	ENTP — THE DEBATER: Smart and curious thinkers who cannot resist an intellectual challenge
ISTP — THE VIRTUOSO: Bold and practical experimenters, masters of all kinds of tools	INFJ — THE ADVOCATE: Quiet and mystical, yet very inspiring and tireless idealists
ISFP — THE ADVENTURER: Flexible and charming artists, always ready to explore and experience something new	INFP — THE MEDIATOR: Poetic, kind and altruistic people, always eager to help a good cause
ESTP — THE ENTREPRENEUR: Smart, energetic and very perceptive people, who truly enjoy living on the edge	ENFJ — THE PROTAGONIST: Charismatic and inspiring leaders, able to mesmerize their listeners
ESFP — THE ENTERTAINER: Spontaneous, energetic and enthusiastic entertainers - are never boring	ENFP — THE CAMPAIGNER: Enthusiastic, creative and sociable free spirits, who can always find a reason to smile

02

Research Questions & Methodology



What is our project about?

Research Questions and Methodology

1. Which features of the dataset contribute most to predicting a person's MBTI personality type?

- **Methodology**: Use feature importance from Random Forest. Correlation analysis for numerical features.

2. How do preprocessing steps like handling outliers and feature selection affect model performance?

- **Methodology**: Compare model performance before and after outlier removal and feature selection using evaluation metrics.

3. Which machine learning model performs best for predicting MBTI personality types?

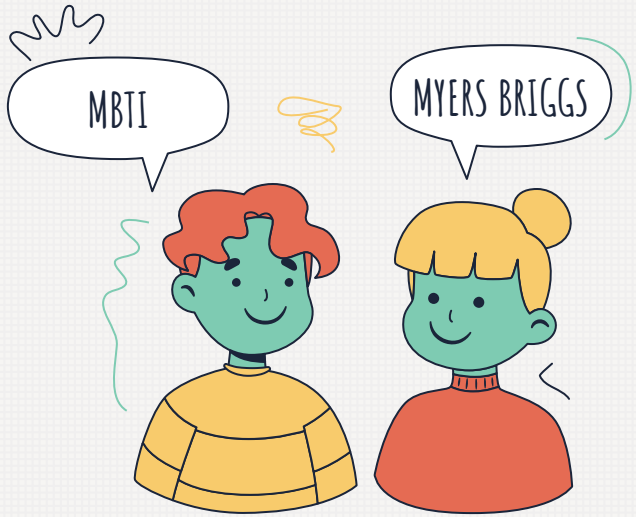
- **Methodology**: Train multiple models and compare their performance metrics.

4. Does hyperparameter tuning improve the performance of machine learning models?

- **Methodology**: Perform hyperparameter tuning using GridSearchCV, and compare model performance.

03

Exploratory Data Analysis & Data Preprocessing



What is our project about?

Data Exploration: first 15 rows & Missing Values

	Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest	Personality
0	21.0	Female	1	5.89208	2.144395	7.32363	5.462224	Arts	ENTP
1	24.0	Female	1	2.48366	3.206188	8.06876	3.765012	Unknown	INTP
2	26.0	Female	1	7.02910	6.469302	4.16472	5.454442	Others	ESFP
3	30.0	Male	0	5.46525	4.179244	2.82487	5.080477	Sports	ENFJ
4	31.0	Female	0	3.59804	6.189259	5.31347	3.677984	Others	ISFP
5	33.0	Female	0	1.06869	7.143507	3.84411	6.347241	Sports	ISFJ
6	32.0	Female	0	6.29802	6.223903	7.90633	6.705588	Arts	ESTJ
7	27.0	Male	1	3.98957	4.406797	5.09055	5.556500	Technology	INFP
8	30.0	Male	0	1.55058	6.652428	0.57707	6.919573	Unknown	ISFJ
9	26.0	Female	1	7.02255	6.929234	9.49484	6.052261	Arts	ESTP
10	32.0	Male	0	3.98624	6.287163	1.83208	5.447141	Arts	ISFP
11	32.0	Male	0	4.53003	4.627212	0.61009	5.558510	Arts	ENFP
12	27.0	Female	0	9.29553	6.634298	7.19146	5.219354	Sports	ESTJ
13	28.0	Male	0	1.01677	5.156652	8.26066	6.966520	Arts	INTJ
14	31.0	Male	0	1.14596	6.498268	2.98133	6.820404	Others	ISFJ

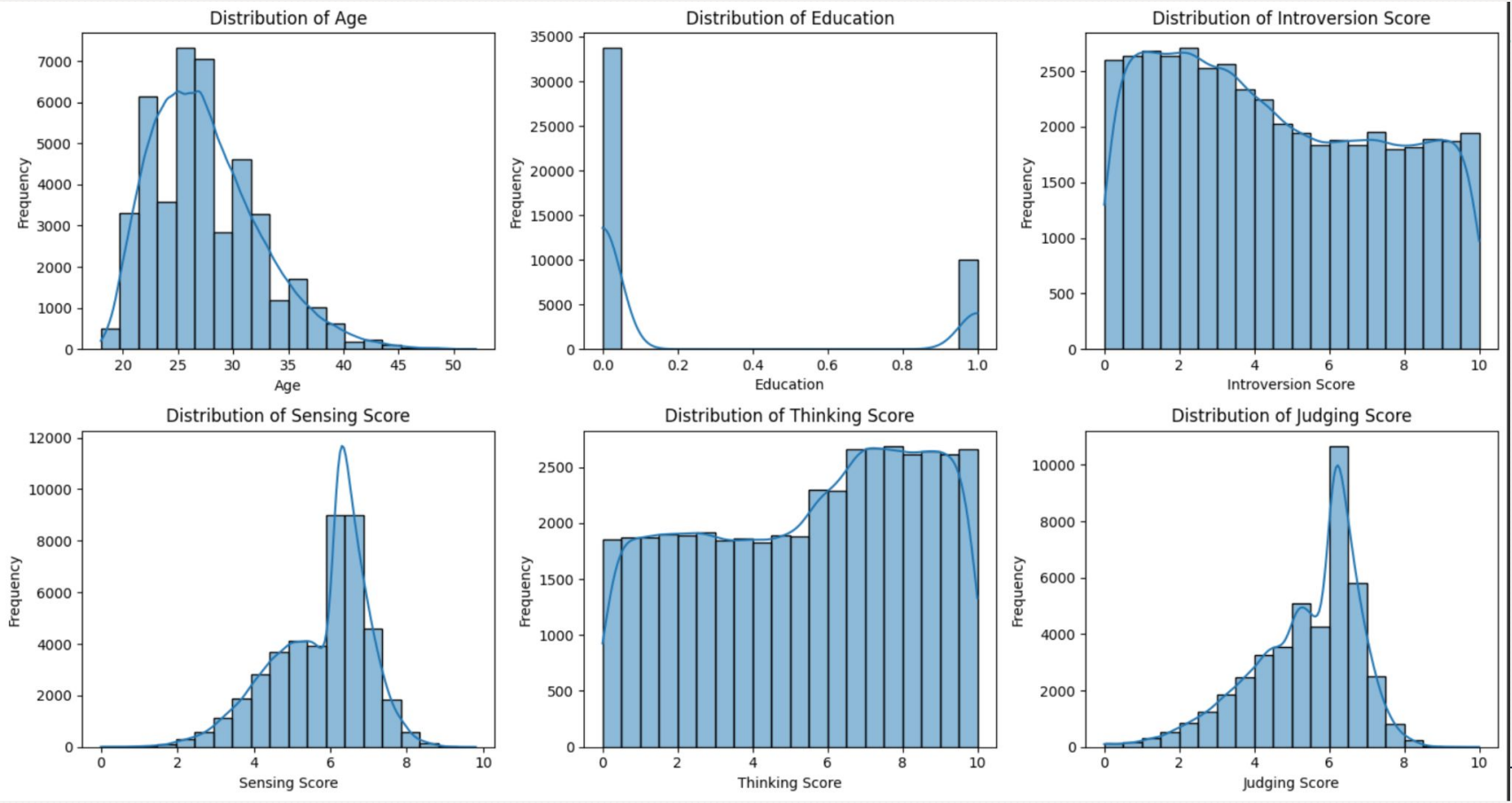
	0
Age	0
Gender	0
Education	0
Introversion Score	0
Sensing Score	0
Thinking Score	0
Judging Score	0
Interest	0
Personality	0

Data Exploration: Info & Summary of Statistics

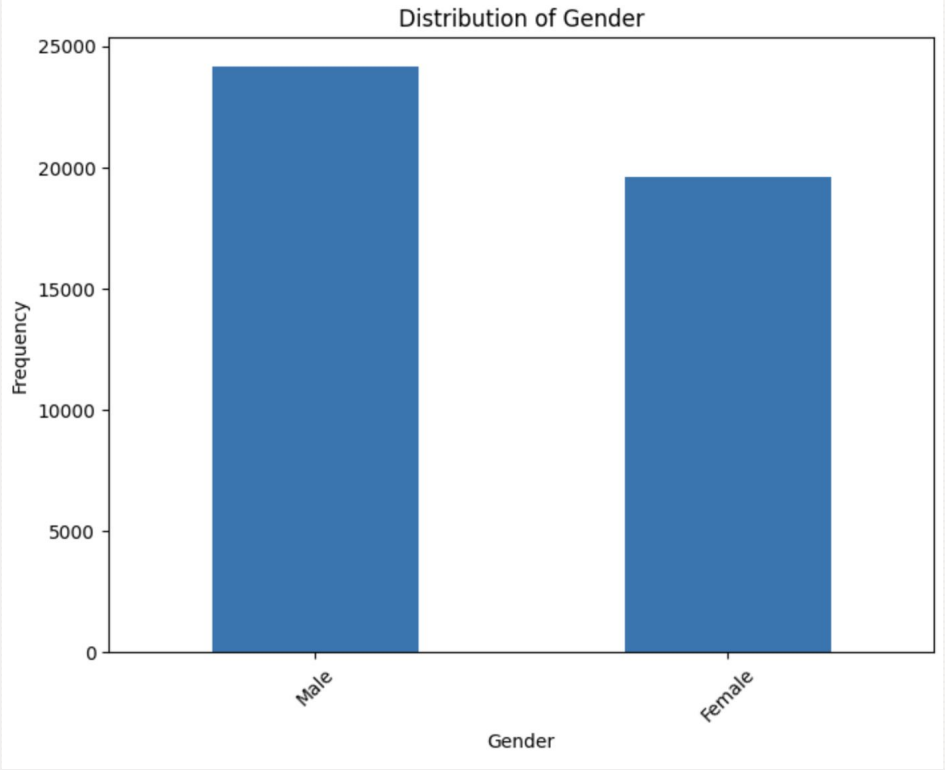
```
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 43744 non-null  float64
 1   Gender              43744 non-null  object
 2   Education           43744 non-null  int64
 3   Introversion Score  43744 non-null  float64
 4   Sensing Score       43744 non-null  float64
 5   Thinking Score      43744 non-null  float64
 6   Judging Score       43744 non-null  float64
 7   Interest            43744 non-null  object
 8   Personality         43744 non-null  object
dtypes: float64(5), int64(1), object(3)
```

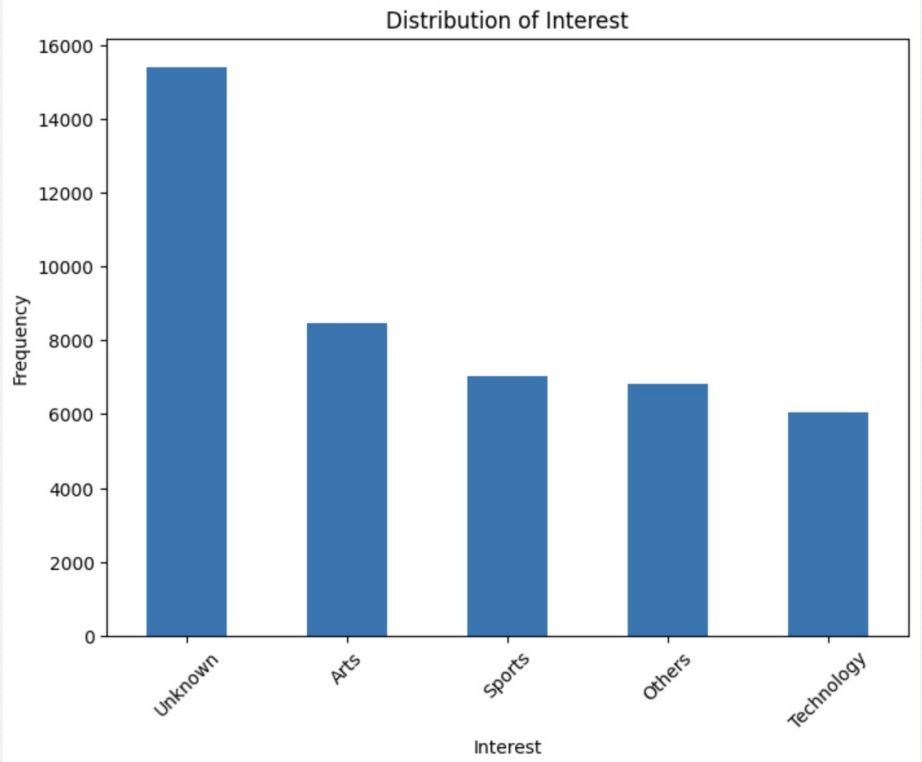
	Age	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score
count	43744.000000	43744.000000	43744.000000	43744.000000	43744.000000	43744.000000
mean	27.437203	0.229014	4.588349	5.780074	5.419131	5.391041
std	4.893805	0.420203	2.902628	1.241648	2.900785	1.442413
min	18.000000	0.000000	0.000150	0.000000	0.000320	0.000000
25%	24.000000	0.000000	2.067020	4.953340	2.895750	4.511842
50%	27.000000	0.000000	4.261680	6.162928	5.769870	5.771635
75%	30.000000	0.000000	7.085002	6.622978	7.923503	6.409583
max	52.000000	1.000000	9.999920	9.803837	9.999770	10.000000

Data Exploration: Distribution of Numerical features

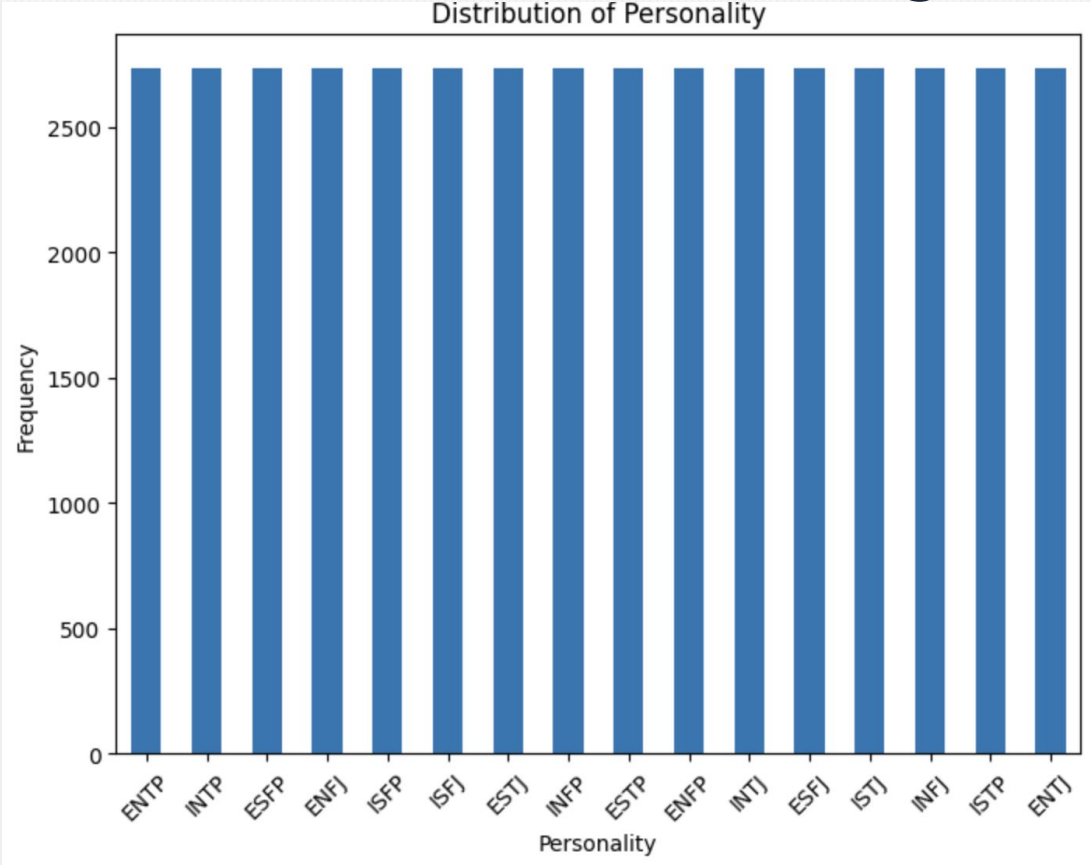


Data Exploration: Distribution of categorical features





Data Exploration: Distribution of Target



Data Exploration: Baseline Model

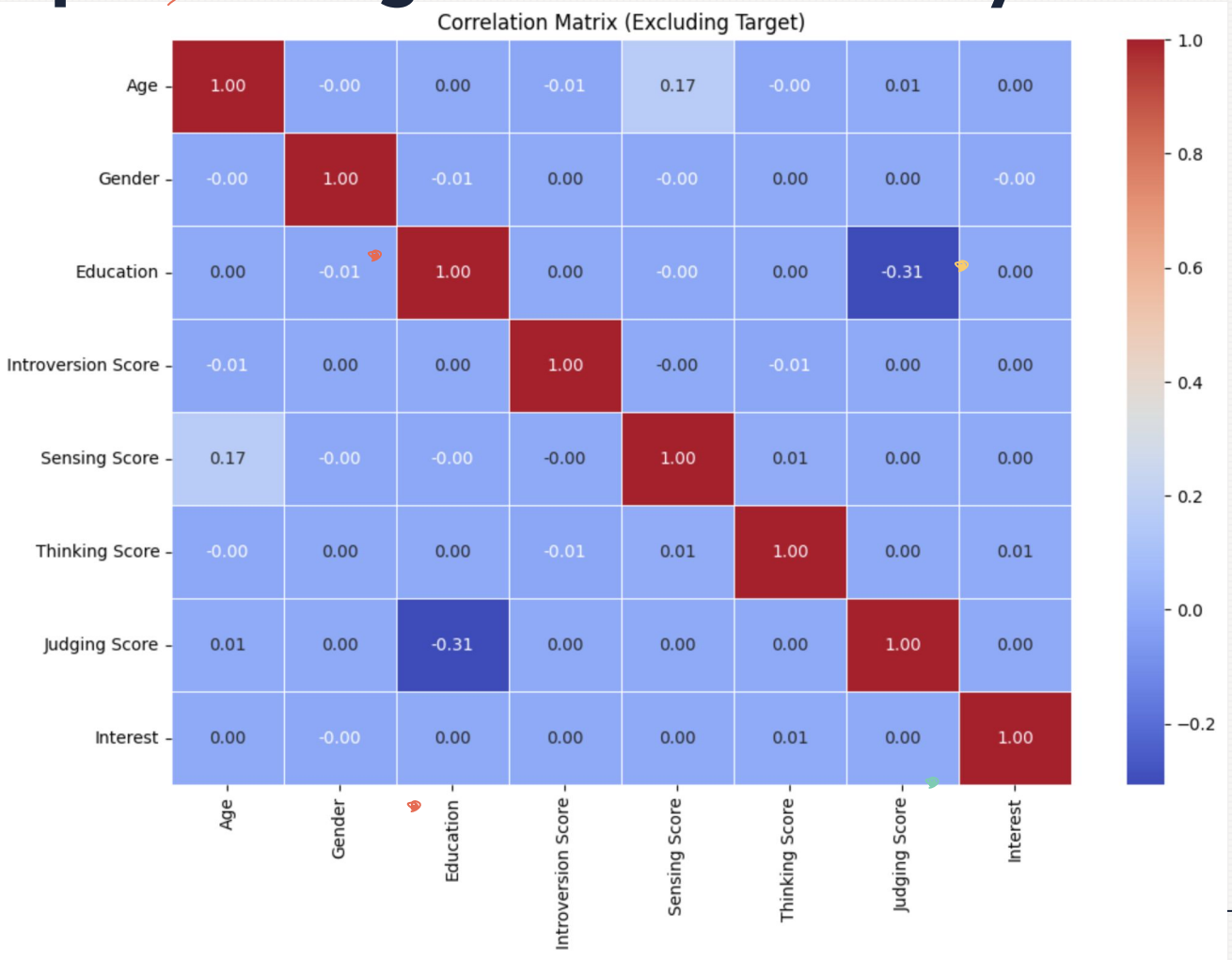
Accuracy of the Dummy Classifier: 0.06
Classification Report:

	precision	recall	f1-score	support
ENFJ	0.05	0.06	0.05	531
ENFP	0.08	0.08	0.08	517
ENTJ	0.06	0.06	0.06	559
ENTP	0.07	0.07	0.07	571
ESFJ	0.05	0.06	0.06	532
ESFP	0.07	0.07	0.07	547
ESTJ	0.04	0.04	0.04	588
ESTP	0.07	0.07	0.07	555
INFJ	0.07	0.07	0.07	540
INFP	0.05	0.05	0.05	558
INTJ	0.05	0.05	0.05	550
INTP	0.07	0.07	0.07	551
ISFJ	0.06	0.06	0.06	574
ISFP	0.07	0.07	0.07	513
ISTJ	0.06	0.06	0.06	509
ISTP	0.07	0.06	0.07	554
accuracy			0.06	8749
macro avg	0.06	0.06	0.06	8749
weighted avg	0.06	0.06	0.06	8749

Data Preprocessing: Label Encoding

	Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest	Personality
0	21.0	1	1	5.89208	2.144395	7.32363	5.462224	0	3
1	24.0	1	1	2.48366	3.206188	8.06876	3.765012	4	11
2	26.0	1	1	7.02910	6.469302	4.16472	5.454442	1	5
3	30.0	0	0	5.46525	4.179244	2.82487	5.080477	2	0
4	31.0	1	0	3.59804	6.189259	5.31347	3.677984	1	13
5	33.0	1	0	1.06869	7.143507	3.84411	6.347241	2	12
6	32.0	1	0	6.29802	6.223903	7.90633	6.705588	0	6
7	27.0	0	1	3.98957	4.406797	5.09055	5.556500	3	9
8	30.0	0	0	1.55058	6.652428	0.57707	6.919573	4	12
9	26.0	1	1	7.02255	6.929234	9.49484	6.052261	0	7
10	32.0	0	0	3.98624	6.287163	1.83208	5.447141	0	13
11	32.0	0	0	4.53003	4.627212	0.61009	5.558510	0	1
12	27.0	1	0	9.29553	6.634298	7.19146	5.219354	2	6
13	28.0	0	0	1.01677	5.156652	8.26066	6.966520	0	10
14	31.0	0	0	1.14596	6.498268	2.98133	6.820404	1	12

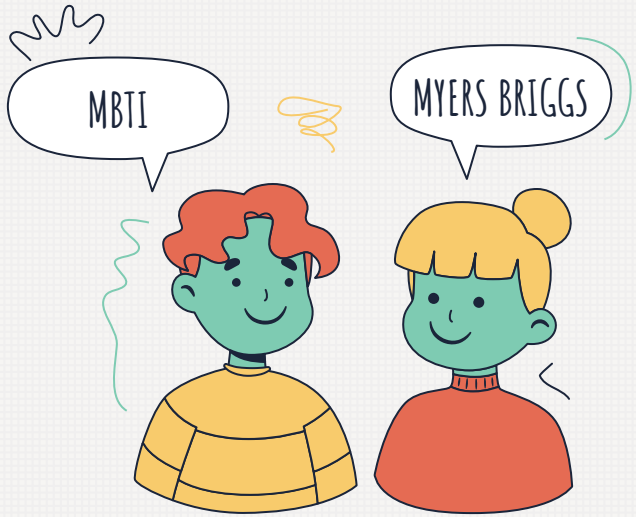
Data Preprocessing: Multicollinearity



04

ML Models & Comparison

What is our project about?



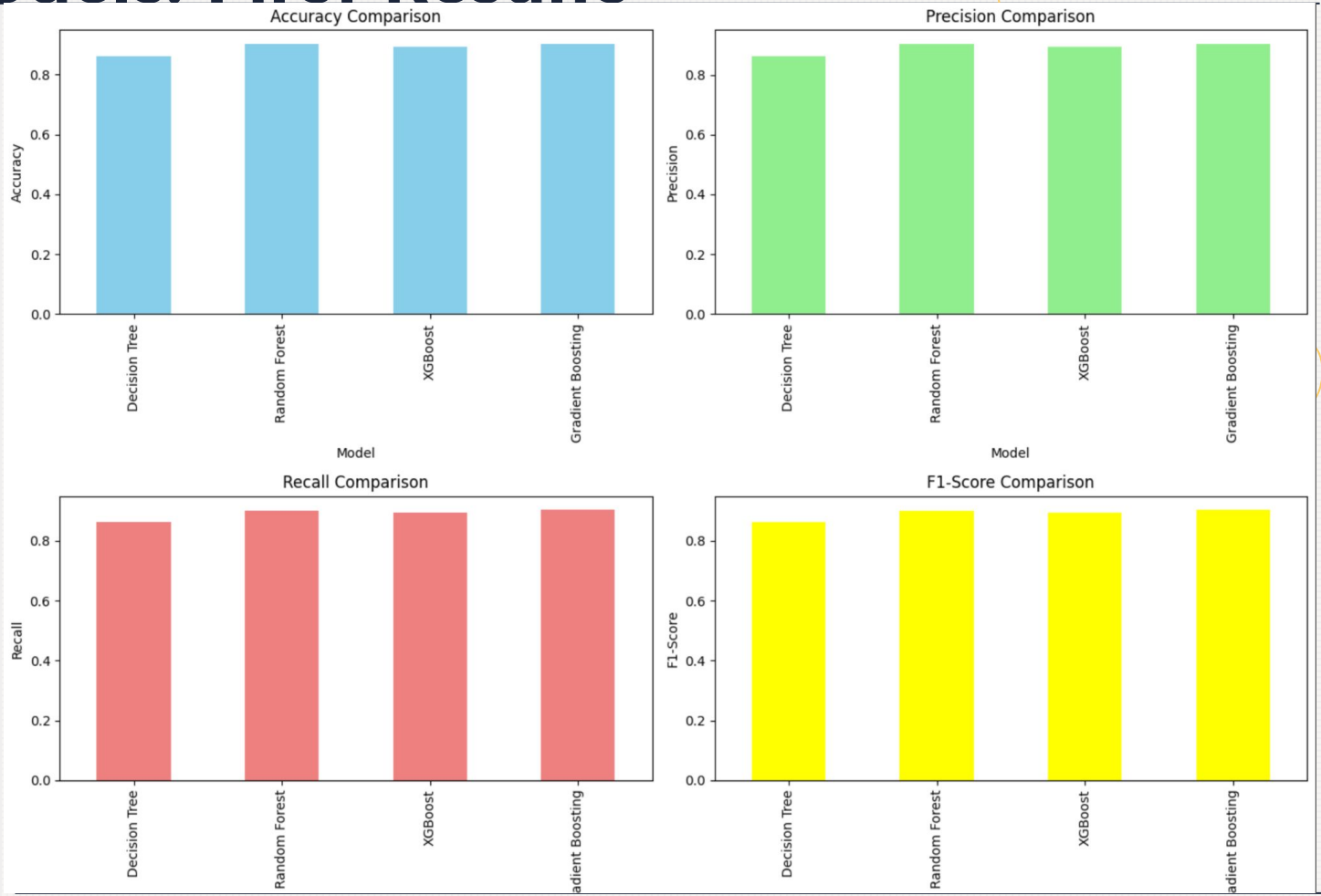
ML Models: Models Selection

```
Logistic Regression - Mean Accuracy: 0.6538 ± 0.0038
KNN - Mean Accuracy: 0.7016 ± 0.0061
Decision Tree - Mean Accuracy: 0.8668 ± 0.0035
Random Forest - Mean Accuracy: 0.8988 ± 0.0046
SVM - Mean Accuracy: 0.7420 ± 0.0030
XGBoost - Mean Accuracy: 0.8926 ± 0.0039
Gradient Boosting - Mean Accuracy: 0.8991 ± 0.0031
```

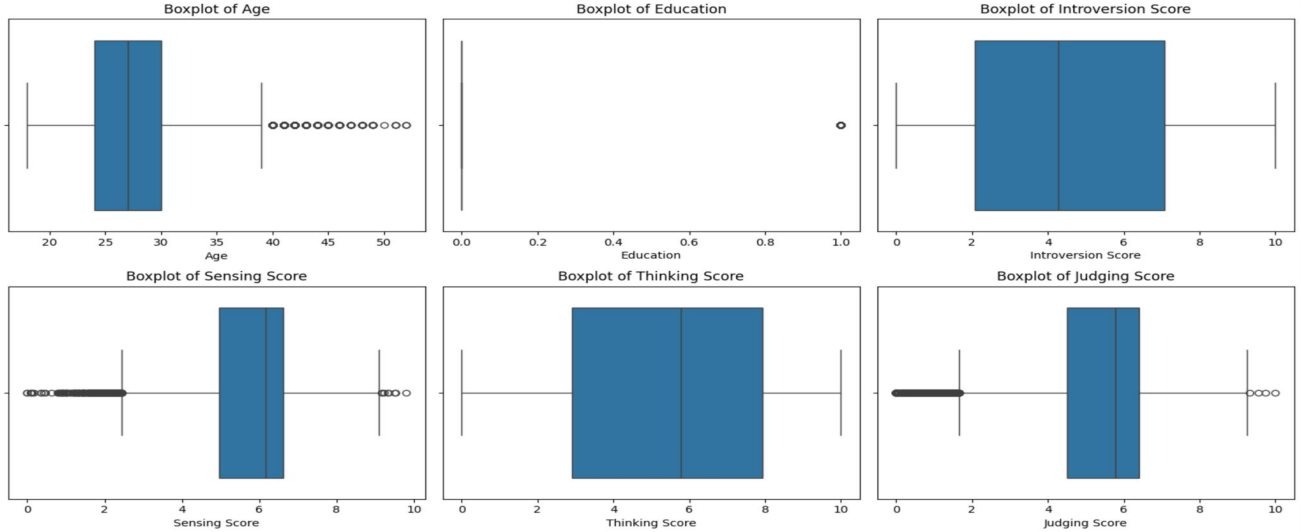
ML Models: First Results

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.862041	0.862336	0.862041	0.862038
1	Random Forest	0.901817	0.902844	0.901817	0.901835
2	XGBoost	0.893588	0.894362	0.893588	0.893595
3	Gradient Boosting	0.903075	0.904046	0.903075	0.903084

ML Models: First Results



Data Preprocessing: Handling Outliers



	Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest
0	3.091042	0.693147	0.693147	1.930373	1.145621	2.119098	1.865974	0.000000
1	3.218876	0.693147	0.693147	1.248083	1.436557	2.204836	1.561300	1.609438
2	3.295837	0.693147	0.693147	2.083072	2.010802	1.641851	1.864769	0.693147
3	3.433987	0.000000	0.000000	1.866442	1.644659	1.341524	1.805083	1.098612
4	3.465736	0.693147	0.000000	1.525630	1.972588	1.842685	1.542867	0.693147

Outlier Counts per Feature:

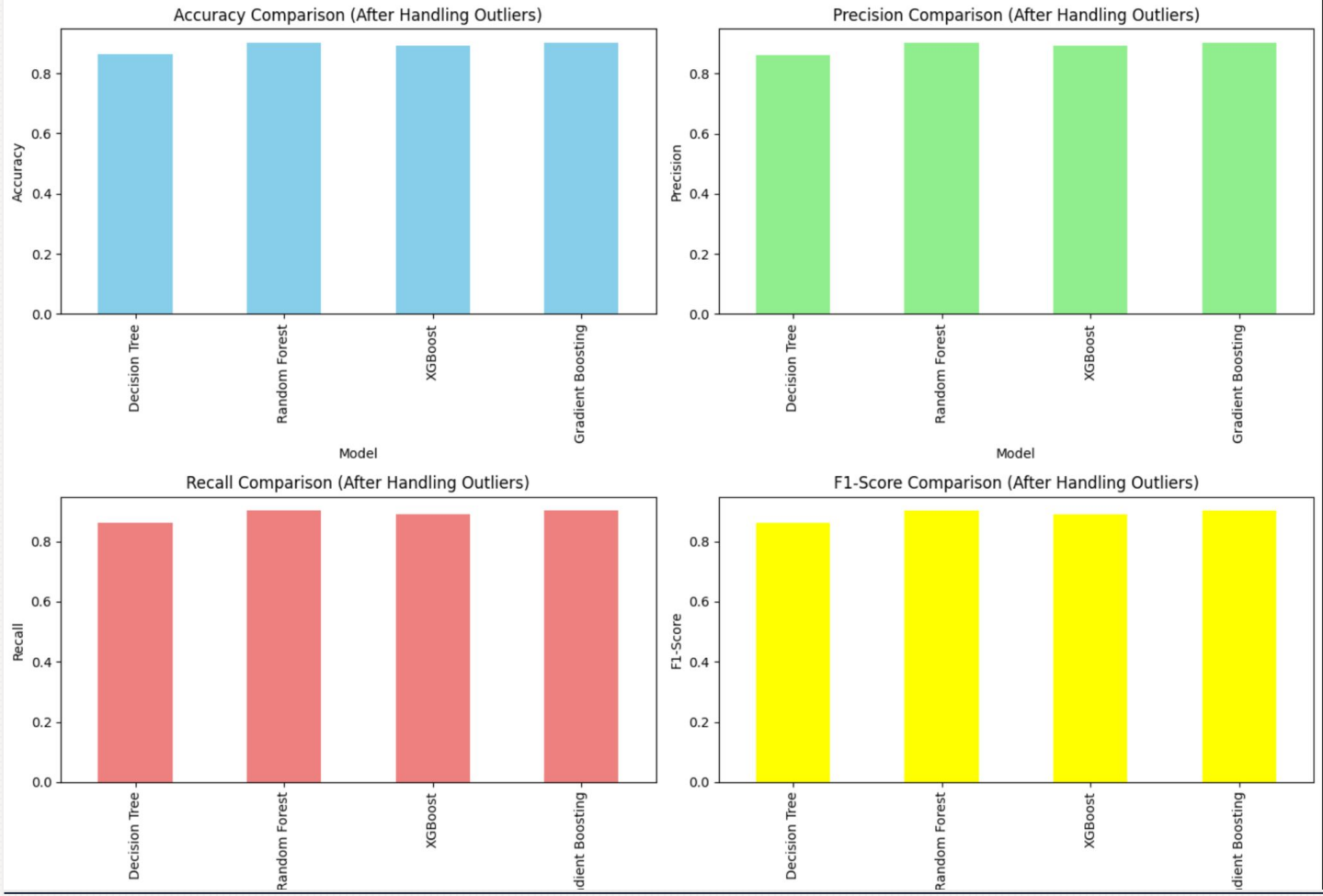
	0
Age	866
Gender	0
Education	10018
Introversion Score	0
Sensing Score	472
Thinking Score	0
Judging Score	735
Interest	0

dtype: int64

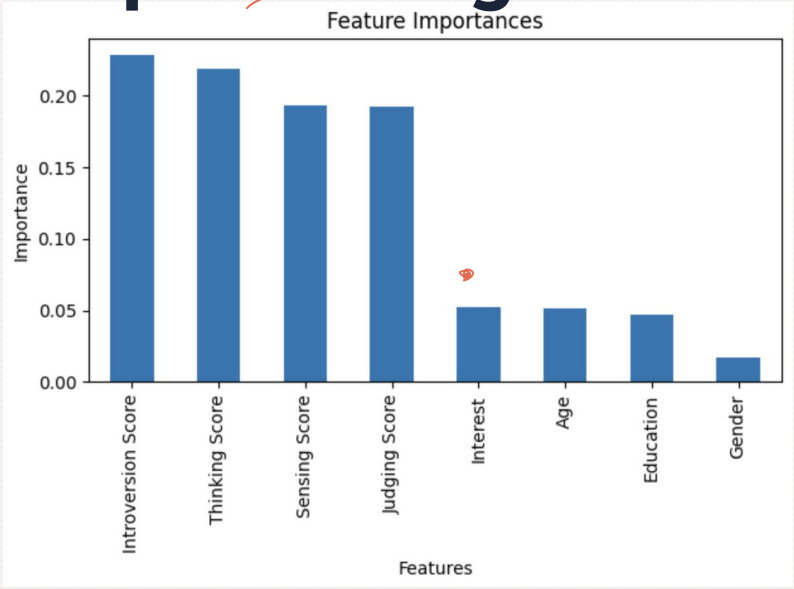
ML Models: Second Results

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.862270	0.862550	0.862270	0.862262
1	Random Forest	0.901817	0.902882	0.901817	0.901828
2	XGBoost	0.891416	0.892287	0.891416	0.891450
3	Gradient Boosting	0.902732	0.903678	0.902732	0.902744

ML Models: Second Results



Data Preprocessing: Features Selection



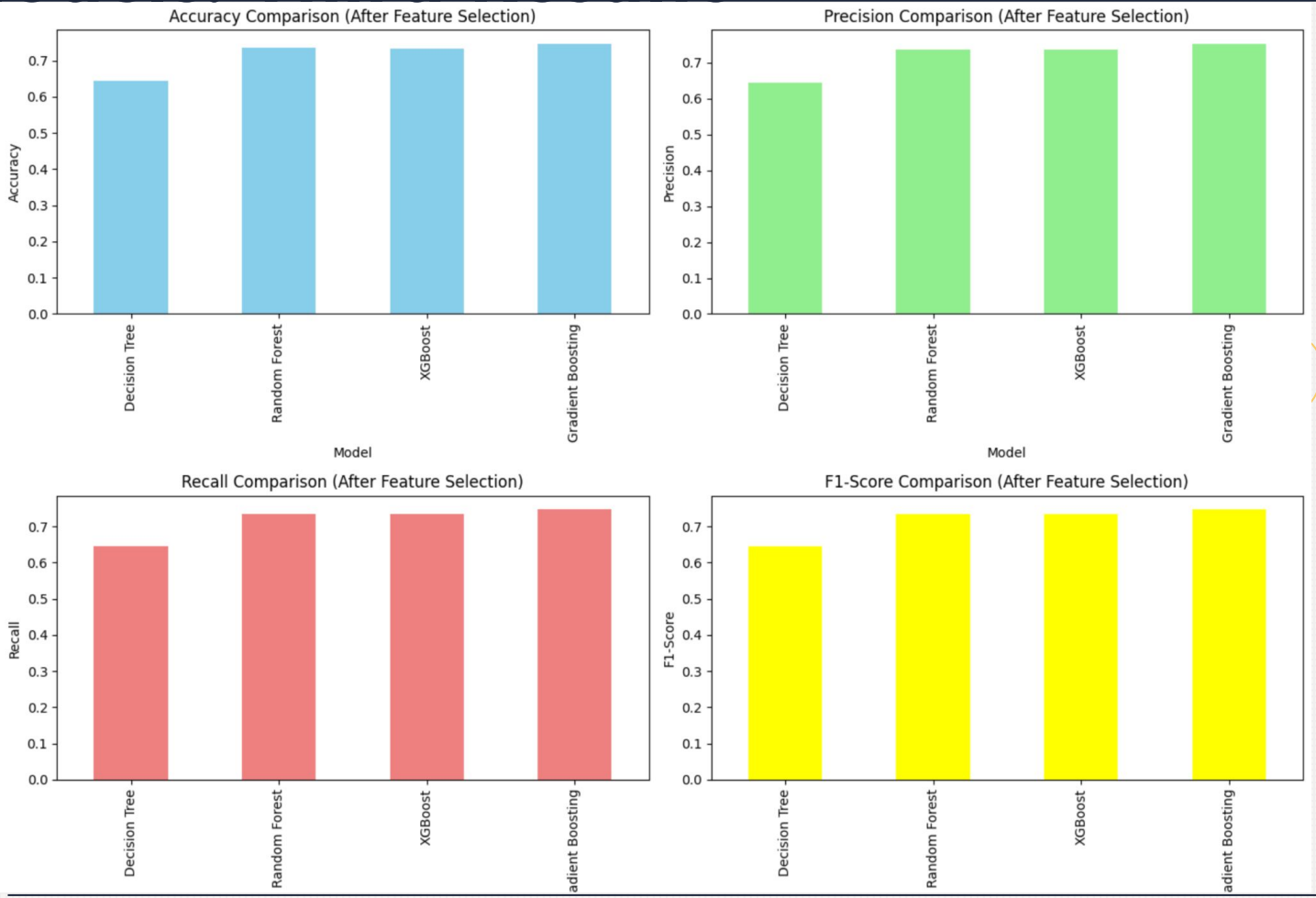
	Feature	Importance
3	Introversion Score	0.228210
5	Thinking Score	0.218731
4	Sensing Score	0.192891
6	Judging Score	0.192414
7	Interest	0.052081
0	Age	0.051050
2	Education	0.047522
1	Gender	0.017101

	Introversion Score	Sensing Score	Thinking Score	Judging Score	Personality
0	5.89208	2.144395	7.32363	5.462224	3
1	2.48366	3.206188	8.06876	3.765012	11
2	7.02910	6.469302	4.16472	5.454442	5
3	5.46525	4.179244	2.82487	5.080477	0
4	3.59804	6.189259	5.31347	3.677984	13

ML Models: Third Results

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.644874	0.644999	0.644874	0.644429
1	Random Forest	0.735170	0.737887	0.735170	0.734808
2	XGBoost	0.733570	0.737106	0.733570	0.733524
3	Gradient Boosting	0.746714	0.753667	0.746714	0.746909

ML Models: Third Results



ML Models: Tuning

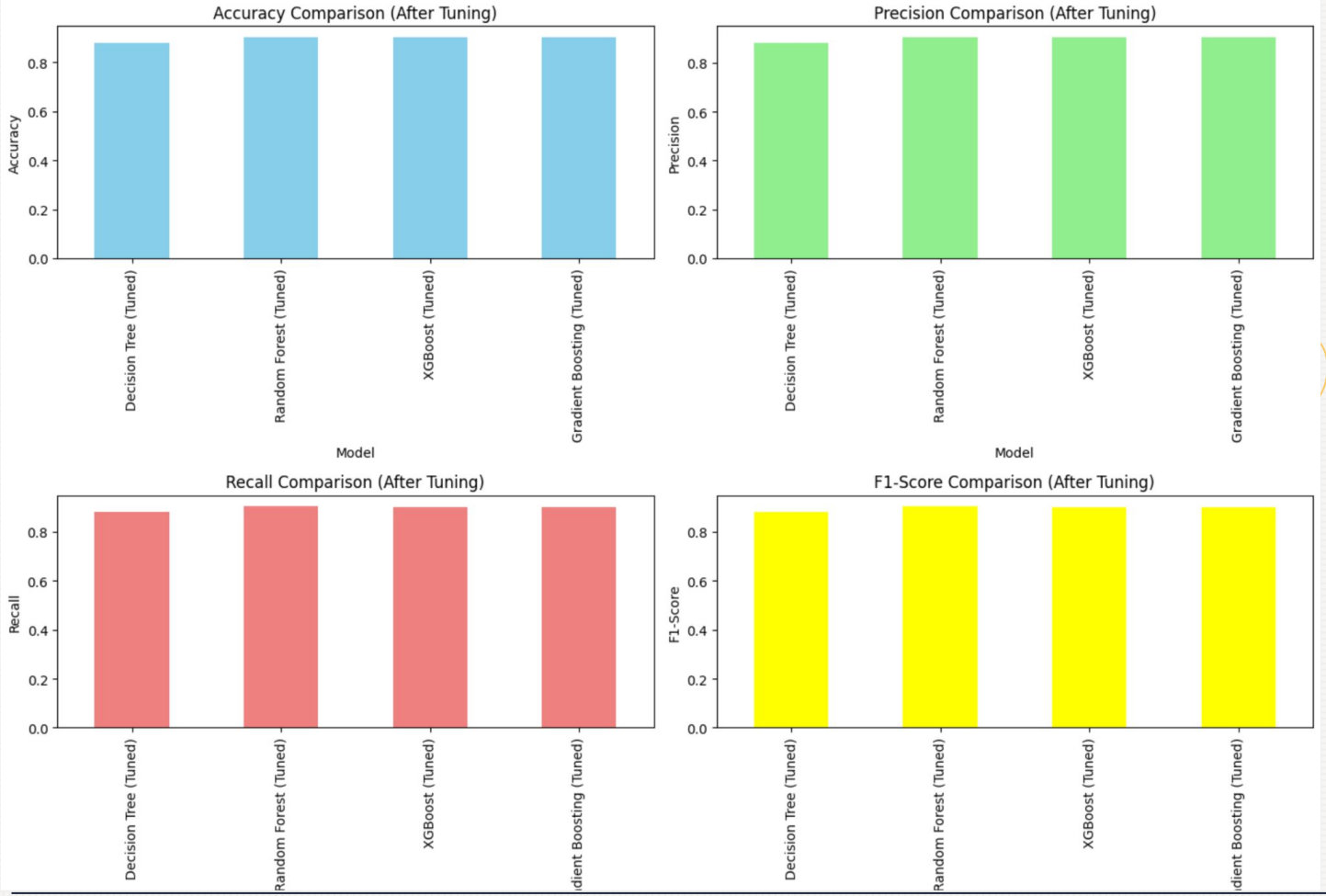
```
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier

param_grid_dt = {
    'max_depth': [3, 5, 10, 20],  # Max depth of the tree
    'min_samples_split': [2, 5, 10],  # Min samples required to split
    'min_samples_leaf': [1, 2, 4]  # Min samples required at leaf node
}
```

ML Models: Fourth Results (Tuning)

	Model	Best Hyperparameters	Accuracy	Precision	Recall	F1-Score
0	Decision Tree (Tuned)	{'max_depth': 10, 'min_samples_leaf': 1, 'min_...	0.880558	0.881909	0.880558	0.880617
1	Random Forest (Tuned)	{'max_depth': None, 'n_estimators': 300}	0.903189	0.904315	0.903189	0.903218
2	XGBoost (Tuned)	{'learning_rate': 0.1, 'n_estimators': 100}	0.902160	0.903233	0.902160	0.902170
3	Gradient Boosting (Tuned)	{'learning_rate': 0.1, 'n_estimators': 200}	0.901817	0.902559	0.901817	0.901822

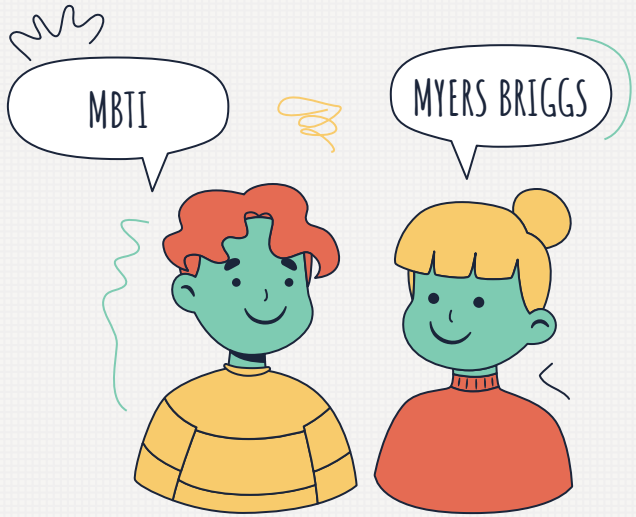
ML Models: Fourth Results



05

All Models Comparison & Best Model

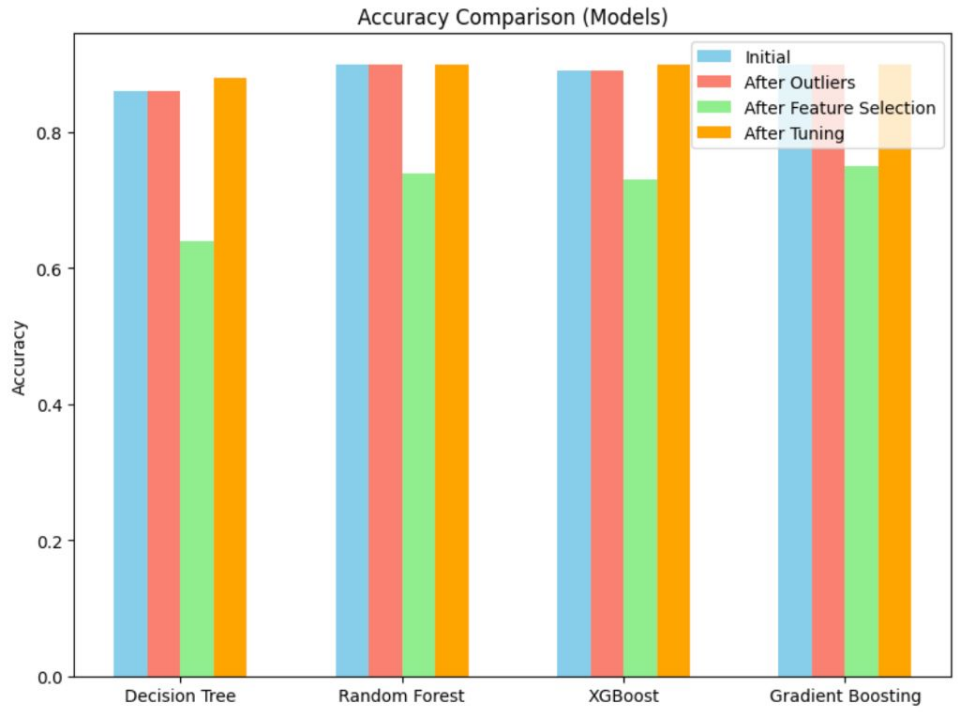
What is our project about?

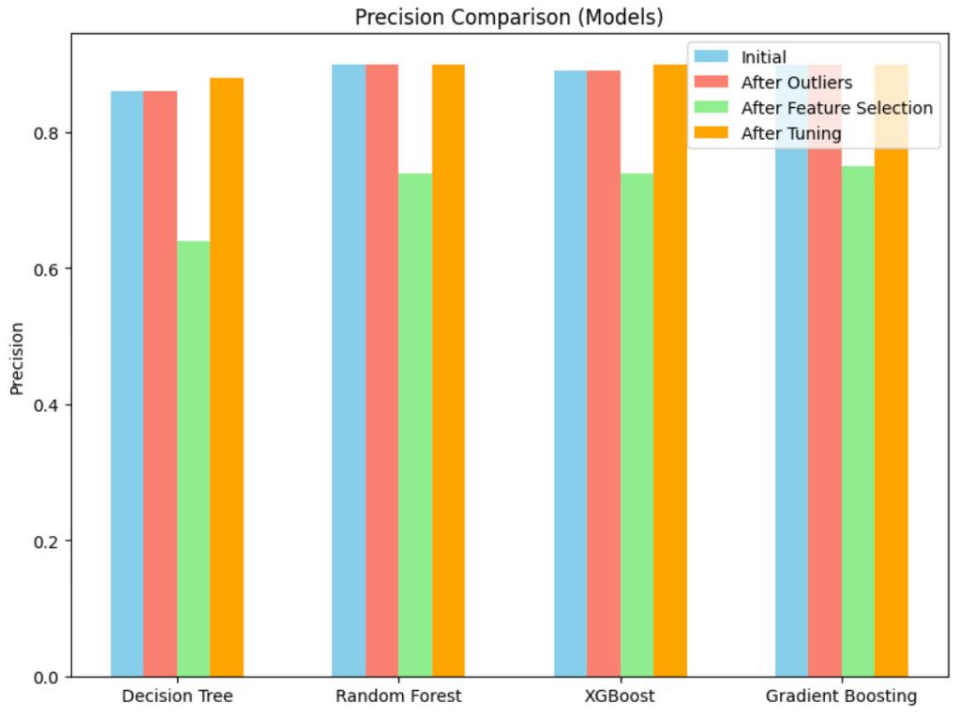


All Models Comparison

	Model	Accuracy	Precision	Recall	F1-Score	Best Hyperparameters
0	Decision Tree	0.862041	0.862336	0.862041	0.862038	NaN
1	Random Forest	0.901817	0.902844	0.901817	0.901835	NaN
2	XGBoost	0.893588	0.894362	0.893588	0.893595	NaN
3	Gradient Boosting	0.903075	0.904046	0.903075	0.903084	NaN
4	Decision Tree	0.862270	0.862550	0.862270	0.862262	NaN
5	Random Forest	0.901817	0.902882	0.901817	0.901828	NaN
6	XGBoost	0.891416	0.892287	0.891416	0.891450	NaN
7	Gradient Boosting	0.902732	0.903678	0.902732	0.902744	NaN
8	Decision Tree	0.644874	0.644999	0.644874	0.644429	NaN
9	Random Forest	0.735170	0.737887	0.735170	0.734808	NaN
10	XGBoost	0.733570	0.737106	0.733570	0.733524	NaN
11	Gradient Boosting	0.746714	0.753667	0.746714	0.746909	NaN
12	Decision Tree (Tuned)	0.880558	0.881909	0.880558	0.880617	{'max_depth': 10, 'min_samples_leaf': 1, 'min_...
13	Random Forest (Tuned)	0.903189	0.904315	0.903189	0.903218	{'max_depth': None, 'n_estimators': 300}
14	XGBoost (Tuned)	0.902160	0.903233	0.902160	0.902170	{'learning_rate': 0.1, 'n_estimators': 100}
15	Gradient Boosting (Tuned)	0.901817	0.902559	0.901817	0.901822	{'learning_rate': 0.1, 'n_estimators': 200}

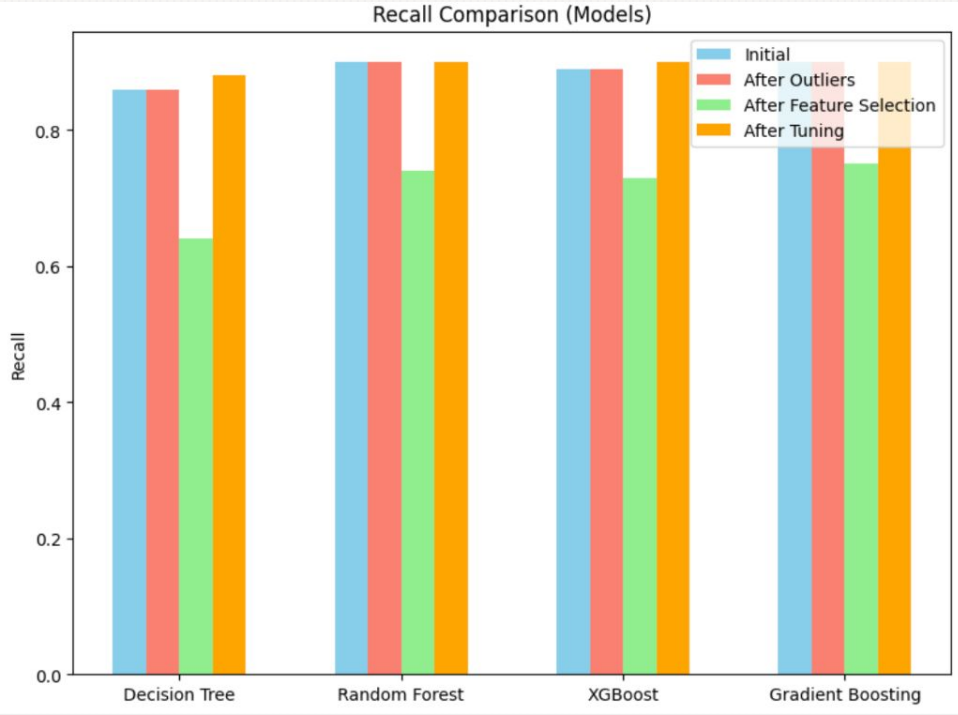
All Models Comparison

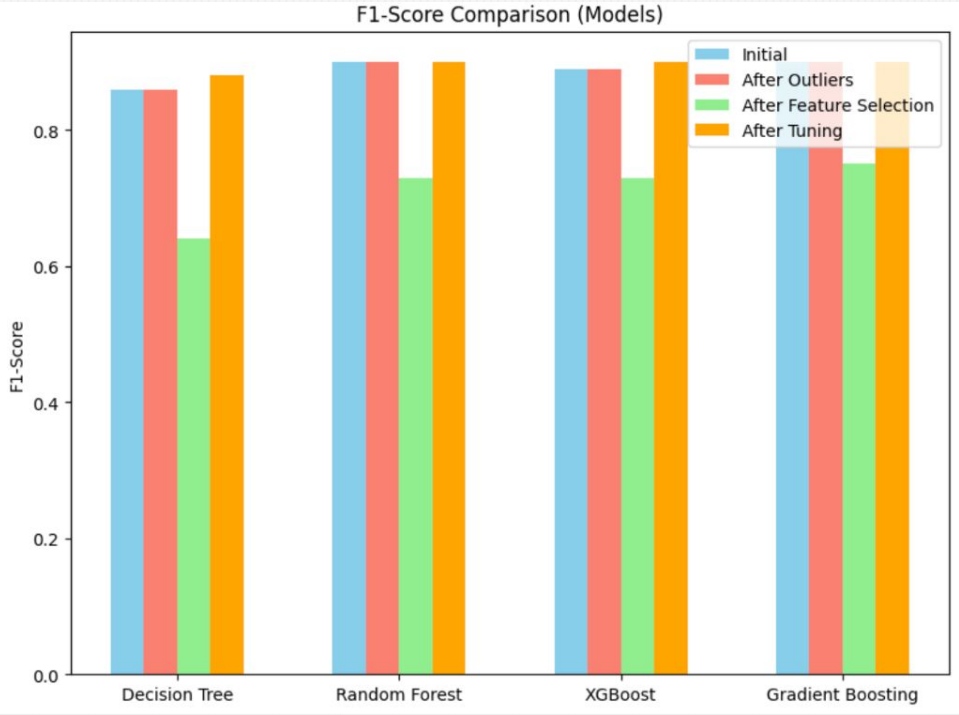




Best Model: Random Forest 0.903189 (After Tuning)

All Models Comparison

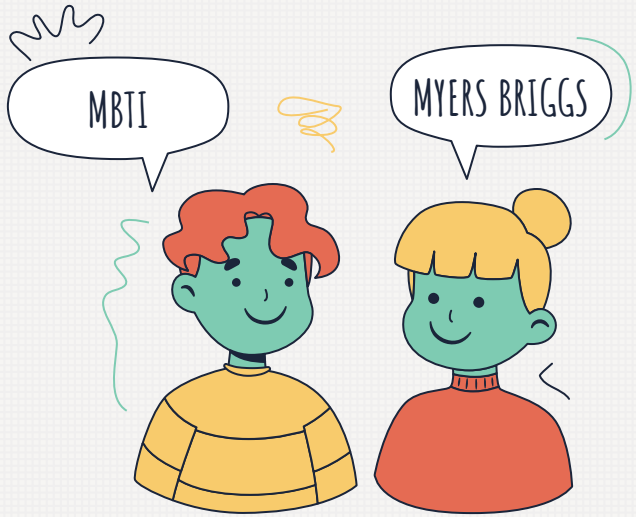




06

Challenges & Future Work

What is our project about?



Challenges

Long Running Time During Hyperparameter Tuning:

- The model training and hyperparameter tuning took too long when trying to optimize multiple hyperparameters. To overcome this, the number of hyperparameters was reduced to make the optimization more efficient and manageable.

Feature Selection Impact:

- The feature selection process negatively impacted model performance. Although the intent was to reduce dimensionality, the models performed worse.

Future Work

Utilizing High-Performance Computing (HPC):

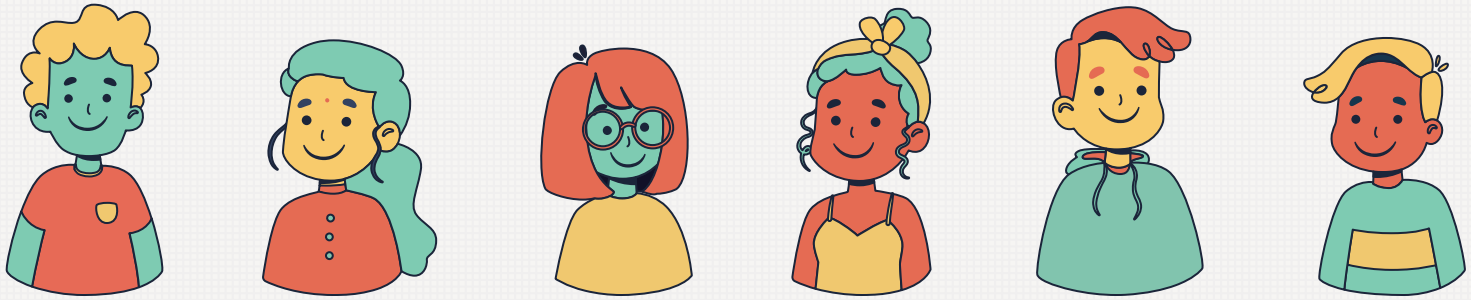
- To handle long running times more effectively, in the future we could leverage High-Performance Computing (HPC) resources. This would allow for faster model training and hyperparameter tuning.

Ensemble Methods:

- In future work, experimenting with ensemble methods like stacking or blending could further improve the performance of the models by combining the strengths of different algorithms.

Thank you!

- **Do you have any questions?**



Resources

Predict people personality types. (2025, January 7). Kaggle.

https://www.kaggle.com/datasets/stealthtechnologies/predict-people-personality-types