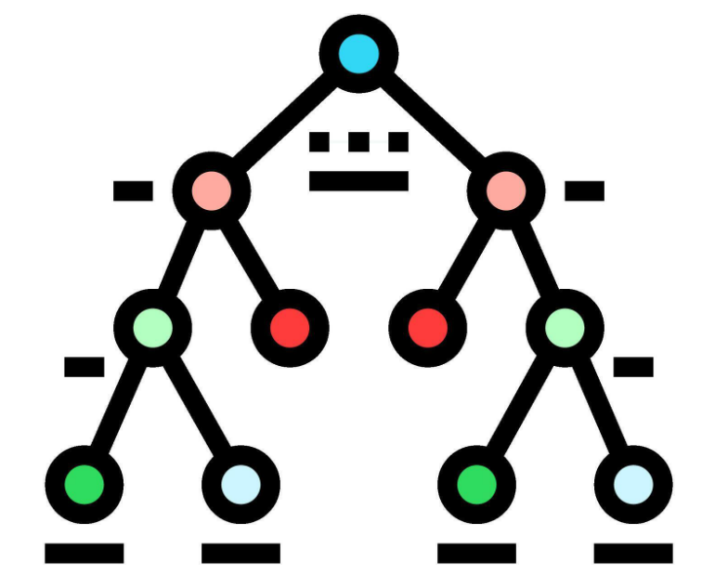
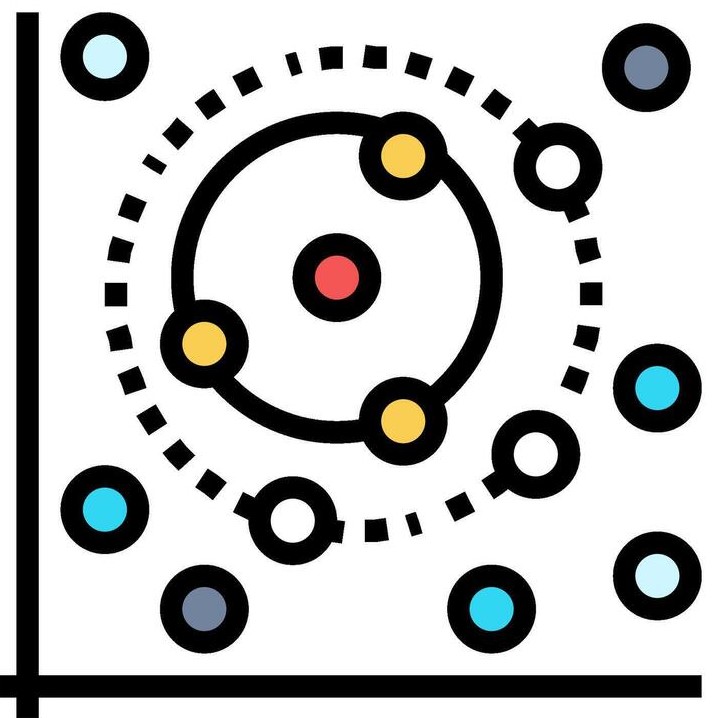
**Personality-Driven Experiences: ML Classification Case Study**

****

**JUNAID**

**DATA SCIENCE & A.I.**

**COHORT 11**

atom**camp**

**Table of Contents**

* **1. Introduction...........................................................................................3**
* 1.1. Project Overview
* **2. Data Understanding and Preprocessing...................................................4**
* 2.1. Data Description
* 2.2. Class Balance Check (Binary Variables)
* 2.3. Categorical Variables
* 2.4. Numerical Variables
* 2.5. Missing Values
* 2.6. Imputation Strategy
* 2.7. Duplicates
* 2.8. Outliers (Boxplot)
* **3. Hypothesis Testing and Feature Analysis...............................................10**
* 3.1. Overview of Hypothesis Testing
* 3.2. Shapiro-Wilk Test (Normality)
* 3.3. Levene's Test (Equality of Variances)
* 3.4. Non-Parametric Hypothesis Testing (Introduction)
* 3.5. Mann-Whitney U Test
* 3.6. Chi-Squared Test
* 3.7. One-Hot Encoding
* 3.8. Correlation Heat-Map
* 3.9. Feature Engineering
* 3.10. Feature Selection
* 3.11. Multicollinearity: Variance Inflation Factor (VIF)
* **4. Model Evaluation..................................................................................16**
* 4.1. Baseline Classifiers
* 4.1.1. Decision Tree
* 4.1.2. K-Nearest Neighbors (KNN)
* 4.1.3. Logistic Regression
* 4.1.4. Random Forest
* 4.2. Results (Before Tuning)
* 4.3. Feature Importance
* 4.4. Hyperparameter Tuning (GridSearchCV)
* 4.5. Best Parameters
* 4.6. Results (After Tuning)
* 4.7. Model Performance Comparison
* 4.8. ROC Curve

**Introduction**

**Powering Hyper-Personalized Experiences with Personality Inference**

ConnectSphere, a growing social-wellness platform, is launching "Personality-Driven Experiences." This feature aims to **personalize user interactions** by inferring whether a user is an **Introvert or Extrovert** based solely on their in-app behavior. This shift will revolutionize engagement by tailoring activity suggestions, in-app challenges, and social nudges.

The core business problem is to build a robust classifier that accurately infers personality from user behavior, without explicit quizzes. This capability is vital for boosting **engagement, improving retention**, and driving **premium upsells** through hyper-personalized recommendations.

Challenges include the absence of direct personality data, potential data sparsity, the need for real-time inference, and ensuring a measurable impact on key metrics like click-through rates and premium conversions.

This report details the project objective: to **build and validate a machine learning classification model** for user personality (Introvert vs. Extrovert) using behavioral data. This model will power an **Activity Suggestion Engine** (low-intensity for introverts, group events for extroverts) and **Engagement Nudges** (e.g., outdoor activity prompts for introverted, low-activity users). Subsequent sections will cover dataset preprocessing, modeling tasks, and performance evaluation.

**Data Description**



The raw dataset will include the following key features, collected daily for each user:

* **Time\_spent\_Alone (hours/day):** Represents the number of hours a user spends in solitary activities within the app.
* **Comfort\_Level\_with\_new\_situations (fear\_stage: Yes/No):** Indicates a user's comfort level when encountering new situations or features within the platform. 'Yes' suggests discomfort/fear, while 'No' suggests comfort.
* **Weekly\_Social\_Events\_Attended:** The cumulative number of social events a user attends within a week.
* **Days\_Going\_Outdoors:** The number of days a user reports going outdoors within a week.
* **Energy\_After\_Socializing (drained: Yes/No):** Reflects a user's energy levels post-socialization. 'Yes' indicates feeling drained, while 'No' indicates feeling energized or neutral.
* **Size\_of\_Close-Friend\_Circle:** The reported or inferred size of a user's close-friend circle within the platform.
* **Weekly\_Social-Media\_Posts:** The total number of social media posts made by a user on the platform per week.
* **Personality (Target Variable):** The inferred personality type, categorized as 'Introvert' or 'Extrovert'. This is the target variable for our classification model.

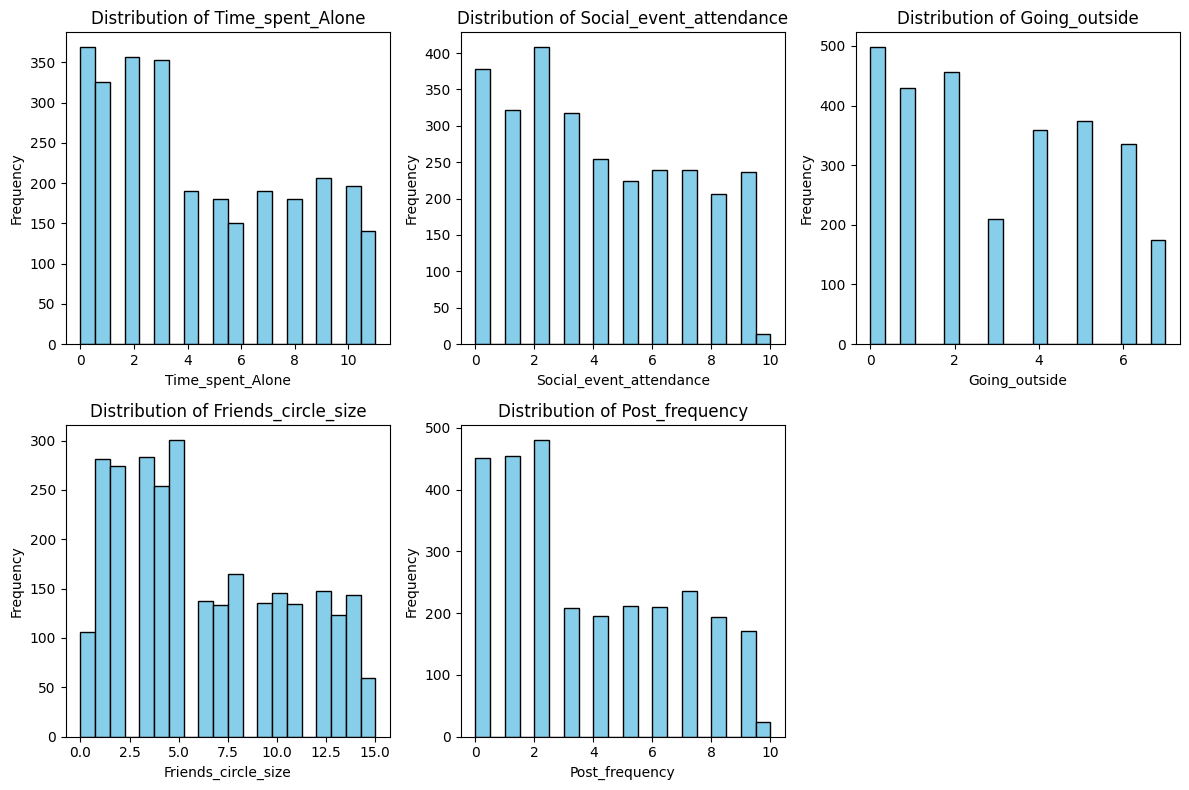
**Class Balance Check**

**Categorical**

****

An initial check was performed on the binary variables, including 'Comfort\_Level\_with\_new\_situations', 'Energy\_After\_Socializing', and the target variable 'Personality'. The distribution of classes for these features appears relatively balanced, indicating no significant skew that would necessitate advanced handling like oversampling or undersampling at this stage. For instance, 'Comfort\_Level\_with\_new\_situations' has approximately 1417 'No' entries and 1410 'Yes' entries, 'Energy\_After\_Socializing' has 1441 'No' and 1407 'Yes', and 'Personality' shows 1491 'Extrovert' and 1409 'Introvert' instances.

**Numerical Variables**



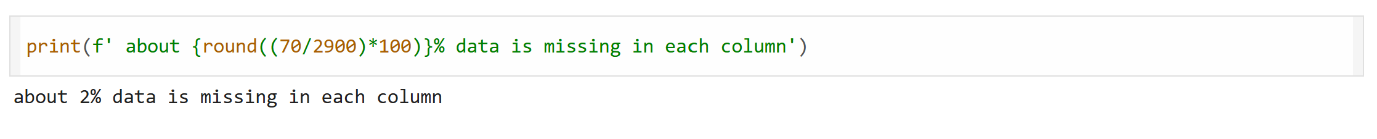
* **Time\_spent\_Alone:** The distribution shows a higher frequency for lower values (0-2 hours), with a gradual decrease as time spent alone increases, suggesting that most users spend less time alone within the app. This feature exhibits a **right-skewed** distribution.
* **Weekly\_Social\_Events\_Attended:** This feature also displays a higher frequency for lower values, with frequencies declining significantly for higher event attendance numbers. This indicates a **right-skewed** distribution.
* **Days\_Going\_Outdoors:** Similar to social events, the distribution is **right-skewed**, indicating that many users report going outdoors for fewer days per week, with a sharp drop-off for more frequent outdoor activities.
* **Size\_of\_Close-Friend\_Circle:** This distribution appears somewhat multimodal or spread out, with noticeable peaks at lower to mid-range friend circle sizes (e.g., around 0-2.5 and 5-7.5), and then tapering off for larger circles. While not as strongly skewed as the others, it tends towards a **positive skew**.
* **Weekly\_Social-Media\_Posts:** This feature is heavily right-skewed, with a very high frequency of users making few to no social media posts. The number of posts drops sharply as the frequency of posting increases, indicating a **strong right-skewness**.

These distributions provide insights into user behavior patterns and will inform potential transformations (e.g., logarithmic transformation for heavily skewed features) during feature engineering to improve model performance.

**Missing Values**

****

* Time\_spent\_Alone: 63 missing values
* Comfort\_Level\_with\_new\_situations: 73 missing values
* Weekly\_Social\_Events\_Attended: 62 missing values
* Days\_Going\_Outdoors: 66 missing values
* Energy\_After\_Socializing: 52 missing values
* Friends\_circle\_size: 77 missing values
* Post\_frequency: 65 missing values
* Personality: 0 missing values (target variable is complete)



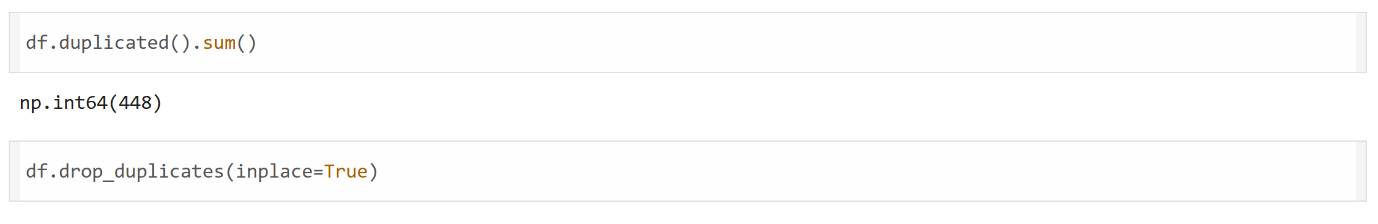
These missing values will need to be addressed through an appropriate imputation strategy during the preprocessing phase to ensure model robustness.

**Imputation Strategy**

**Categorical variables** (e.g., 'Comfort\_Level\_with\_new\_situations', 'Energy\_After\_Socializing'), missing values will be imputed using the **mode** (most frequent value) of their respective columns.

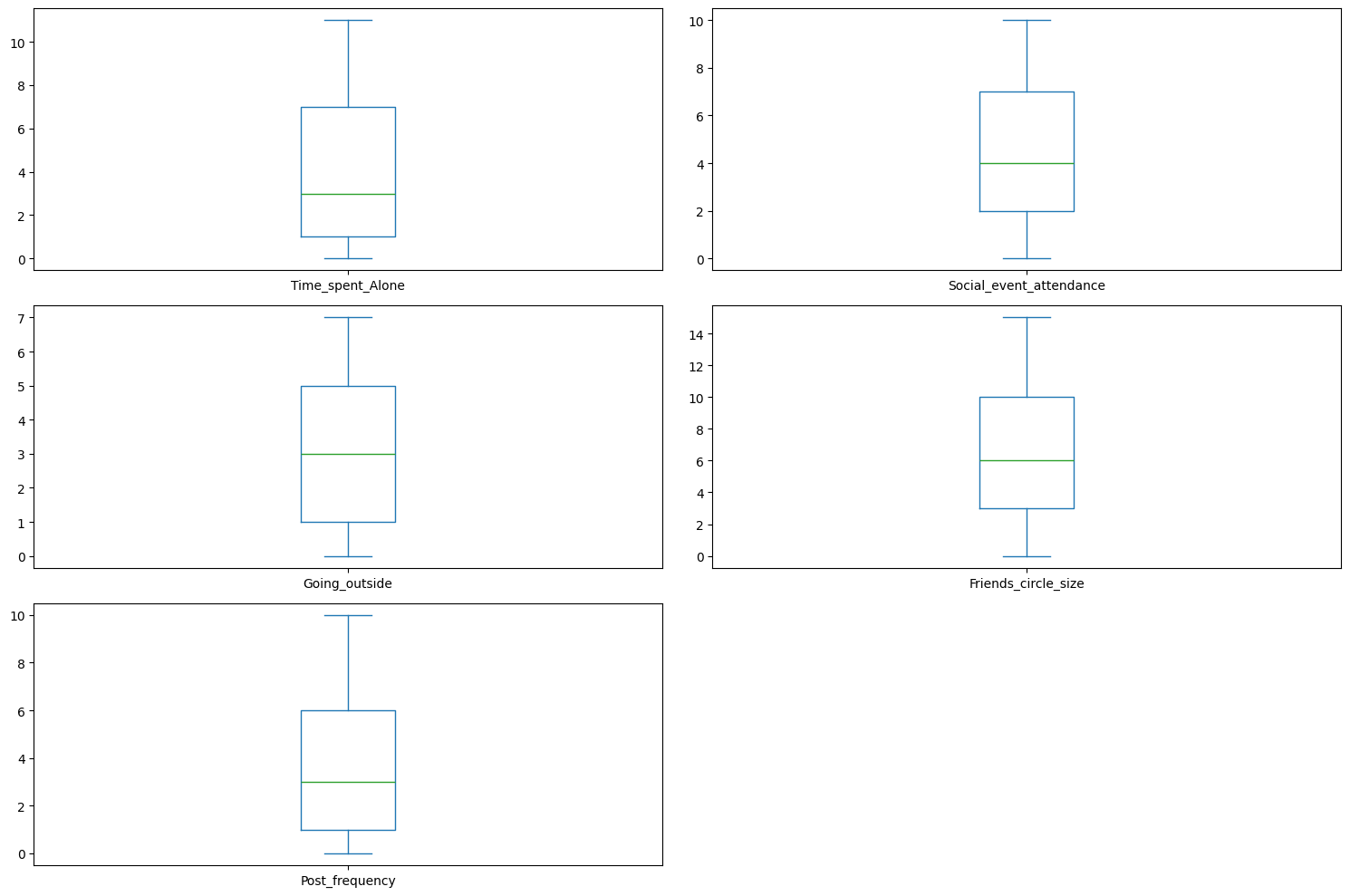
**Numerical variables** (e.g., 'Time\_spent\_Alone', 'Weekly\_Social\_Events\_Attended', 'Days\_Going\_Outdoors', 'Friends\_circle\_size', 'Post\_frequency'), given that many exhibit skewed distributions, missing values will be imputed using the **median** of their respective columns. The median is chosen over the mean as it is less sensitive to outliers and skewness, providing a more robust imputation for these features.

**Duplicates**

****

Dropping duplicate records.

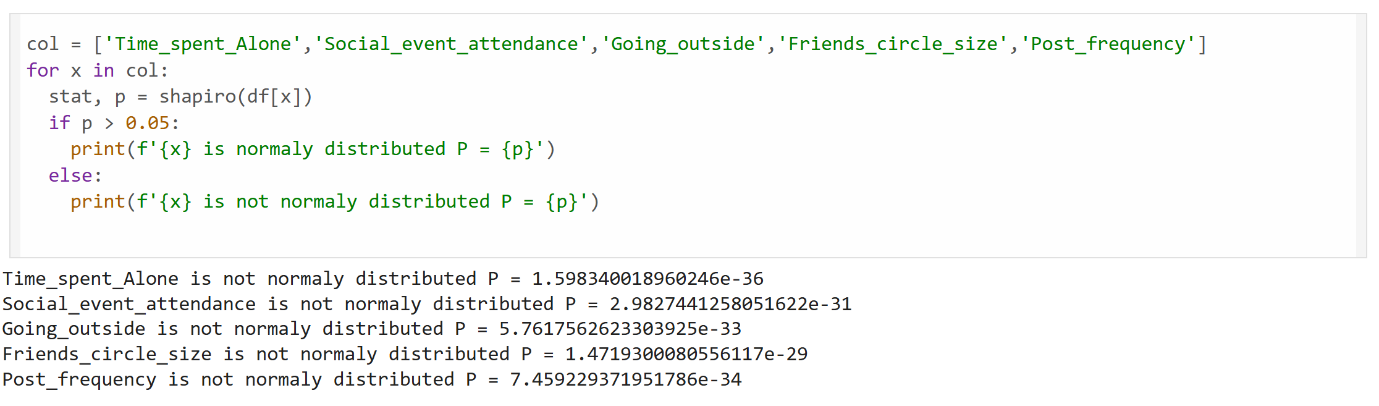
**Outliers**

****

a visual inspection of box plots for the numerical features ('Time\_spent\_Alone', 'Social\_event\_attendance', 'Going\_outside', 'Friends\_circle\_size', and 'Post\_frequency') indicates **no significant outliers** were detected. The data points appear to fall within the typical range, suggesting that no specific outlier treatment beyond median imputation for missing values will be required for these features.

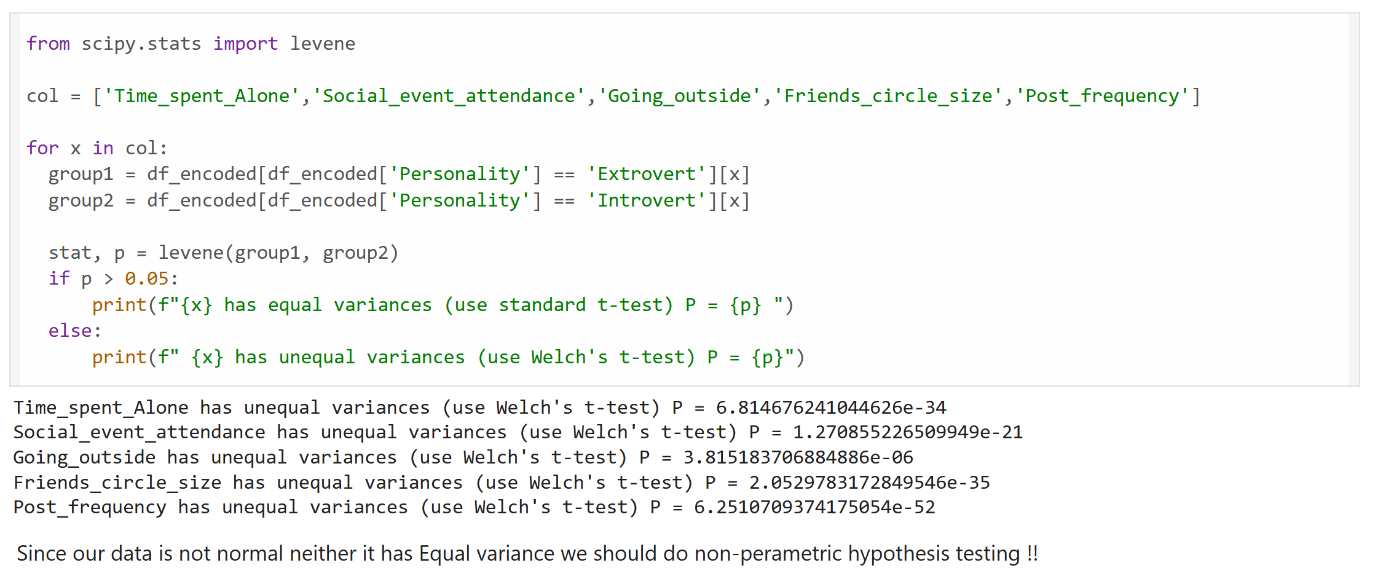
**Hypothesis Testing**

**Shapiro-Wilk**

****

the normality of these numerical features, a Shapiro-Wilk test was conducted. The results, indicated by extremely small p-values (e.g., 1.598×10−36 for Time\_spent\_Alone, 2.982×10−31 for Social\_event\_attendance, etc.), confirm that all numerical features are not normally distributed (p<0.05).

**Levene's test**

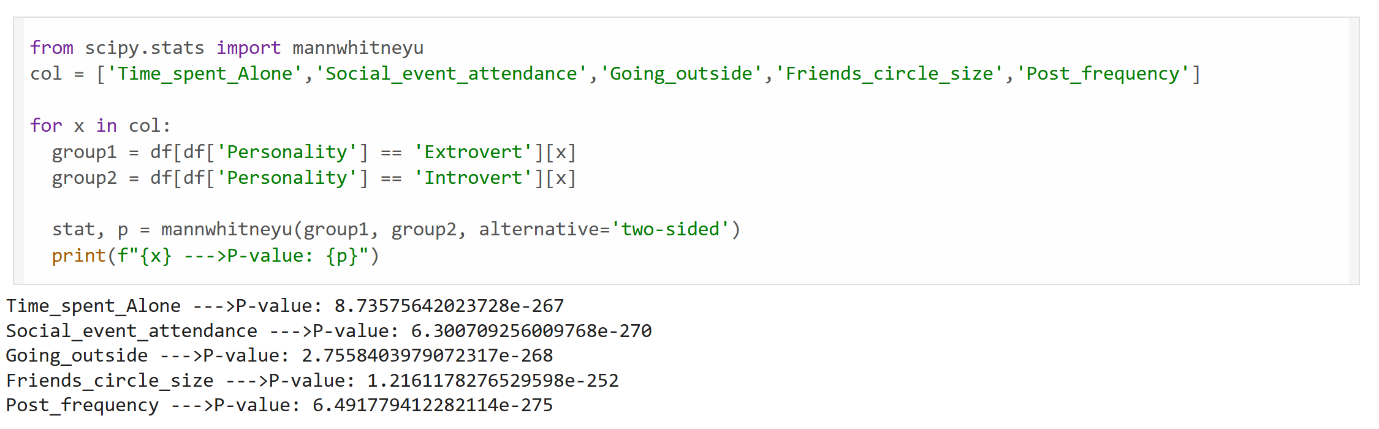


**Levene's test** was performed to check for equality of variances between the 'Extrovert' and 'Introvert' groups for each numerical feature. The results show that for all numerical features (e.g., Time\_spent\_Alone, Social\_event\_attendance, Going\_outside, Friends\_circle\_size, Post\_frequency), the p-values are extremely small (p<0.05), indicating **unequal variances** between the groups.

Given that the numerical features are neither normally distributed nor exhibit equal variances, parametric hypothesis tests like the standard t-test or ANOVA are not suitable. Therefore, **non-parametric hypothesis testing** methods will be employed if comparative statistical analysis between personality groups is required for these features.

**Non-Parametric Hypothesis Testing**

**Mann-Whitney U Test**

****

the Mann-Whitney U Test was performed to compare the distributions of numerical features between the 'Extrovert' and 'Introvert' personality groups. The results are as follows:

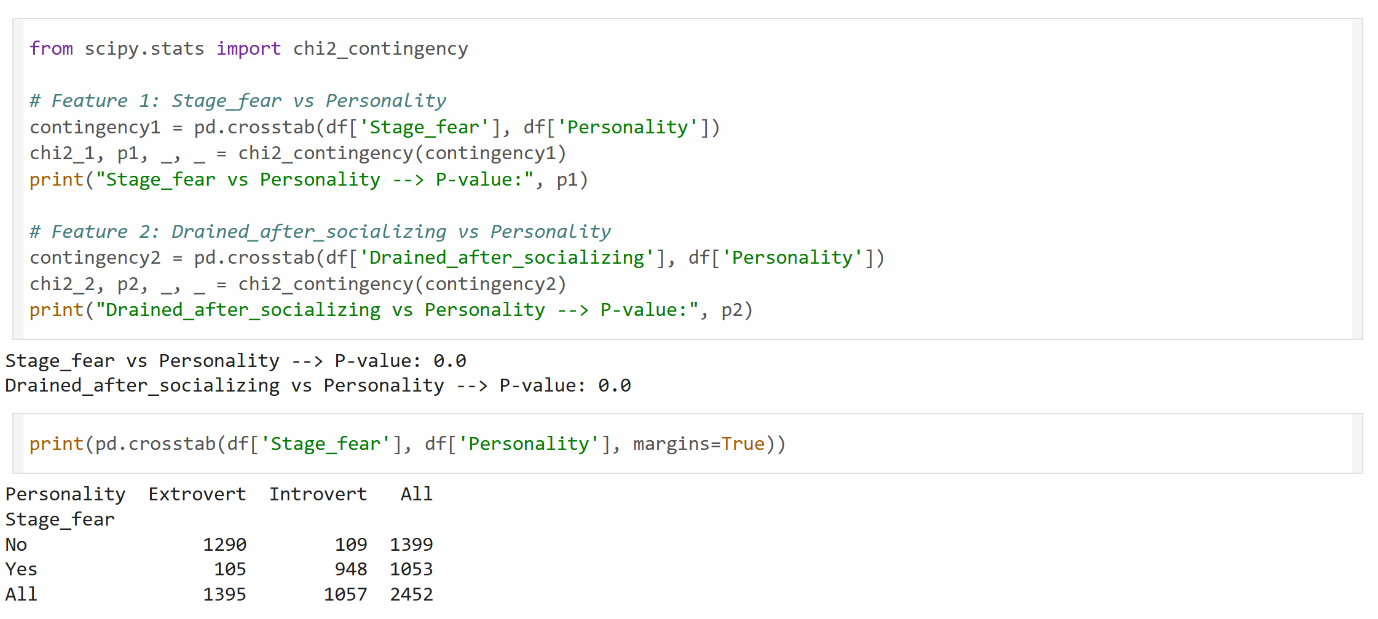
Feature P-value Interpretation:

* Time\_spent\_Alone 8.73e-267 Significantly different Social\_event\_attendance 6.30e-270 Significantly different
* Going\_outside 2.75e-268 Significantly different
* Friends\_circle\_size 1.22e-252 Significantly different
* Post\_frequency 6.49e-275 Significantly different

Interpretation: All your p-values are extremely small → way below 0.05. This tells you:

These numerical features are strongly associated with Personality. That means: Extroverts and Introverts differ significantly across these variables.

**Chi-Squared test**

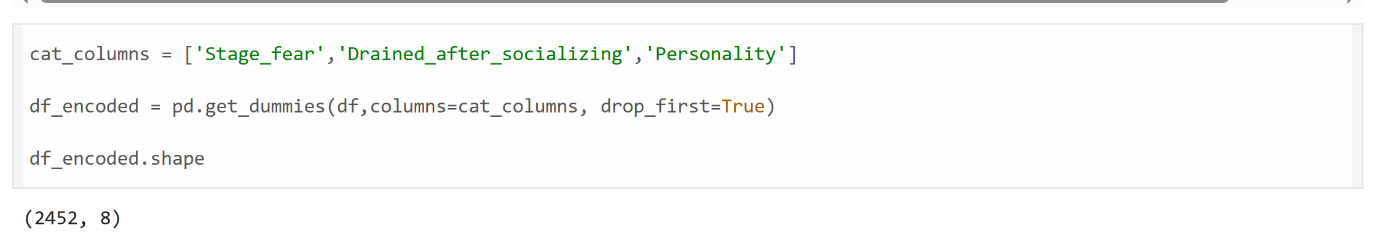


Finally, for the **categorical variables**, a **Chi-squared test of independence** was performed to assess their association with the 'Personality' target variable. The results are as follows:

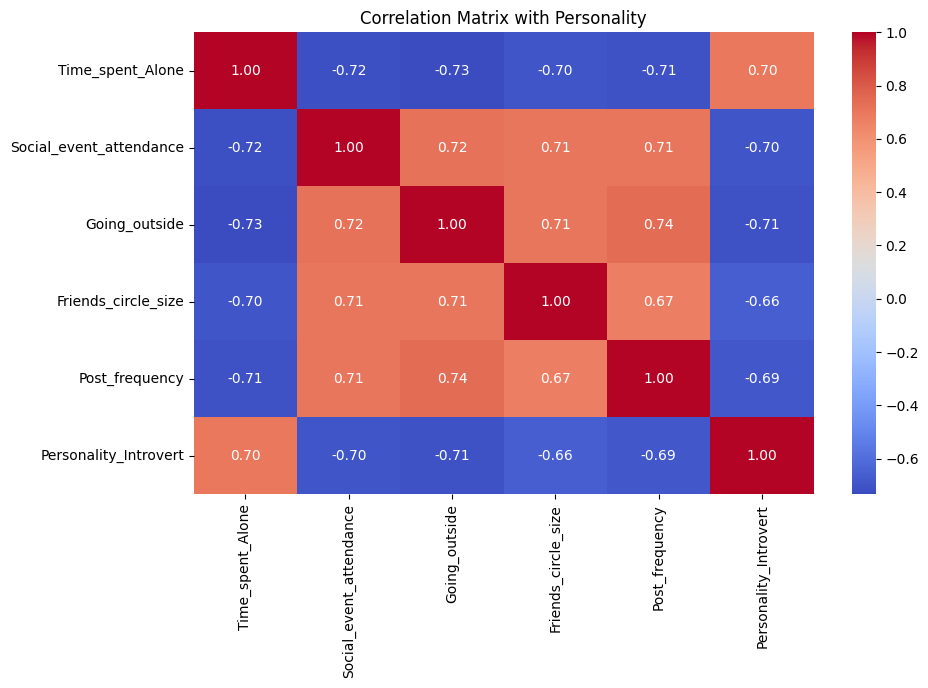
* **Stage\_fear vs. Personality:** P-value = 0.0
* **Drained\_after\_socializing vs. Personality:** P-value = 0.0

These p-values of 0.0 indicate a **highly significant association** between both 'Stage\_fear' and 'Drained\_after\_socializing' and the 'Personality' type. This means that an individual's comfort level with new situations and their energy levels after socializing are strongly dependent on whether they are an Introvert or an Extrovert, making these features highly relevant for personality classification.

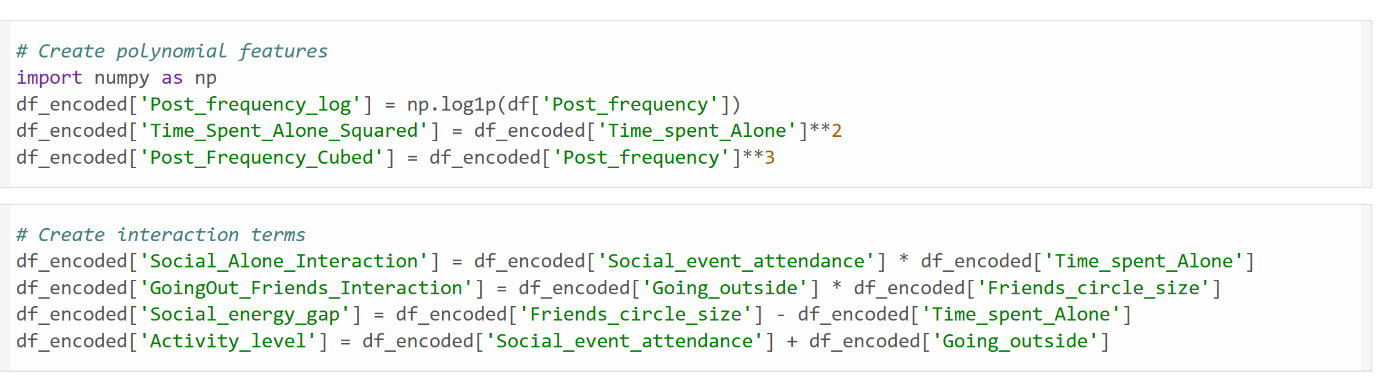
**One-Hot Encoding**

****

**Correlation Heat-Map**

****

**Feature Engineering**

****

The Feature engineering will be performed to potentially enhance model performance and capture more complex relationships within the data.

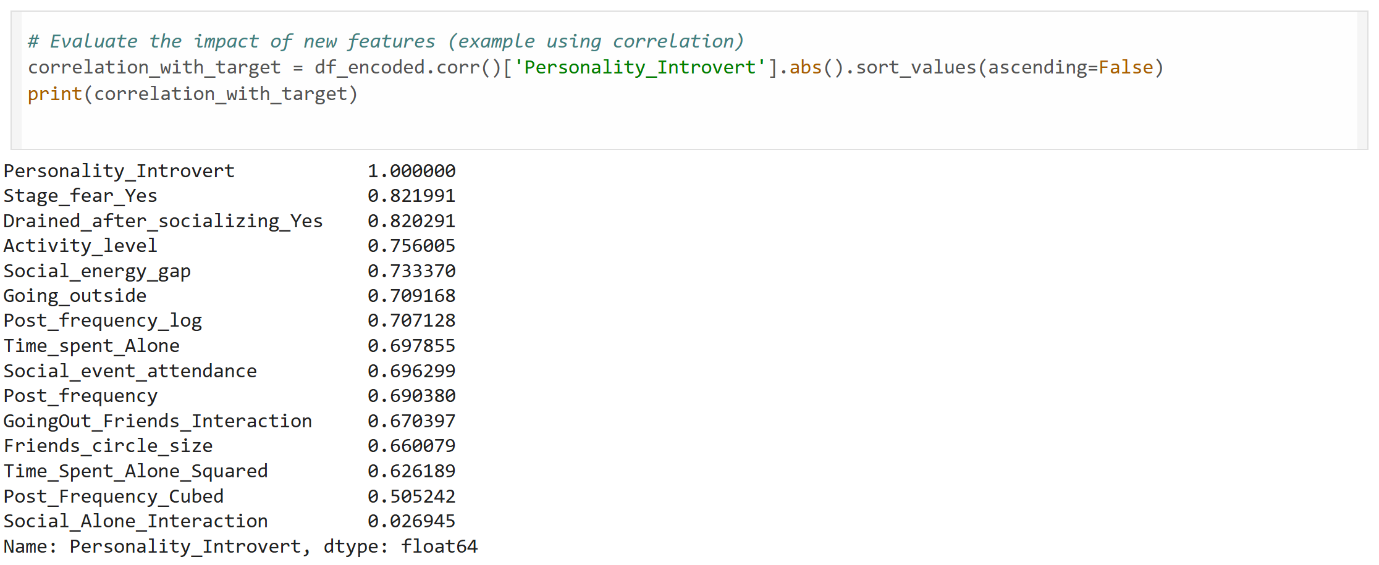
The following new features will be engineered:

* Polynomial Features:
  + Post\_frequency\_log: A logarithmic transformation (np.log1p) of 'Post\_frequency' to handle its strong right-skewness and potentially linearize its relationship with the target.
  + Time\_Spent\_Alone\_Squared: A squared term of 'Time\_spent\_Alone' to capture non-linear relationships.
  + Post\_Frequency\_Cubed: A cubed term of 'Post\_frequency' for further non-linear capture.
* Interaction Terms:
  + Social\_Alone\_Interaction: An interaction term calculated as Social\_event\_attendance \* Time\_spent\_Alone. This feature aims to capture how the balance between social engagement and time spent alone might influence personality.
  + GoingOut\_Friends\_Interaction: An interaction term calculated as Going\_outside \* Friends\_circle\_size. This could reveal combined effects of outdoor activity and social network size.
  + Social\_energy\_gap: A difference term calculated as Friends\_circle\_size - Time\_spent\_Alone. This might represent a "social energy balance" or preference.
  + Activity\_level: A sum of Social\_event\_attendance + Going\_outside. This simple aggregate feature aims to provide a general measure of a user's overall activity.

These engineered features are expected to provide richer insights into user behavior and improve the predictive power of the personality classification model.

**Feature Selection**

After the creation of new features, their individual predictive power and association with the target variable were evaluated using correlation analysis. The absolute correlation of each feature with Personality\_Introvert

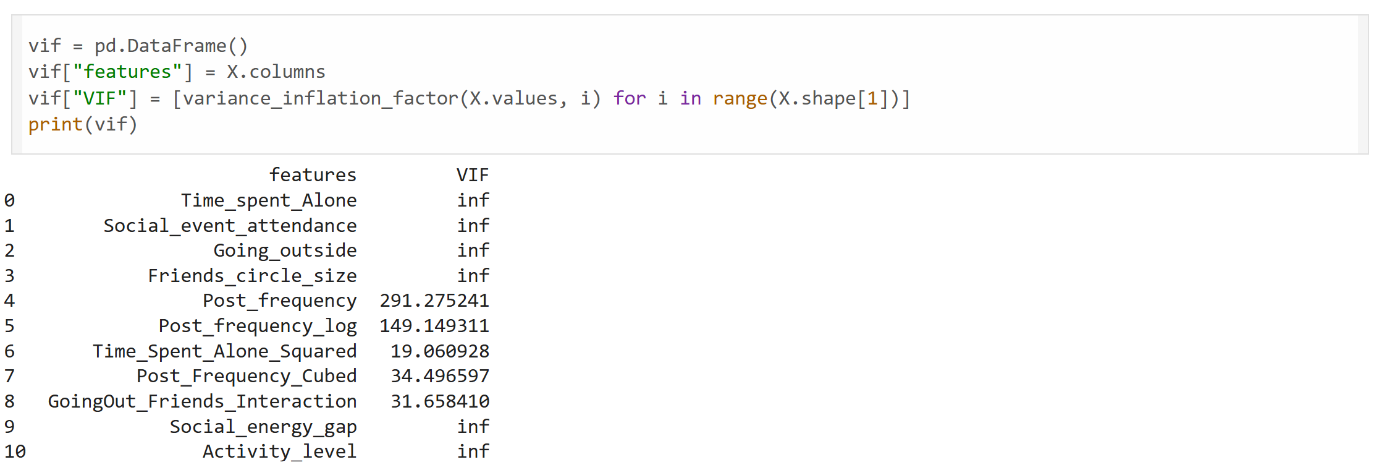


Based on this analysis, the Social\_Alone\_Interaction feature exhibits a very low correlation (0.026945) with the target variable, indicating negligible predictive power. Consequently, this feature will be **dropped** from the dataset prior to model training to avoid introducing noise and potentially improve model efficiency without sacrificing performance.

**Multicollinearity**

**Variance Inflation Factor (VIF)**

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the severity of multicollinearity (correlation between predictor variables) in a multiple linear regression model.



Feature VIF Interpretation:

Most core features (1–4) inf Perfect multicollinearity

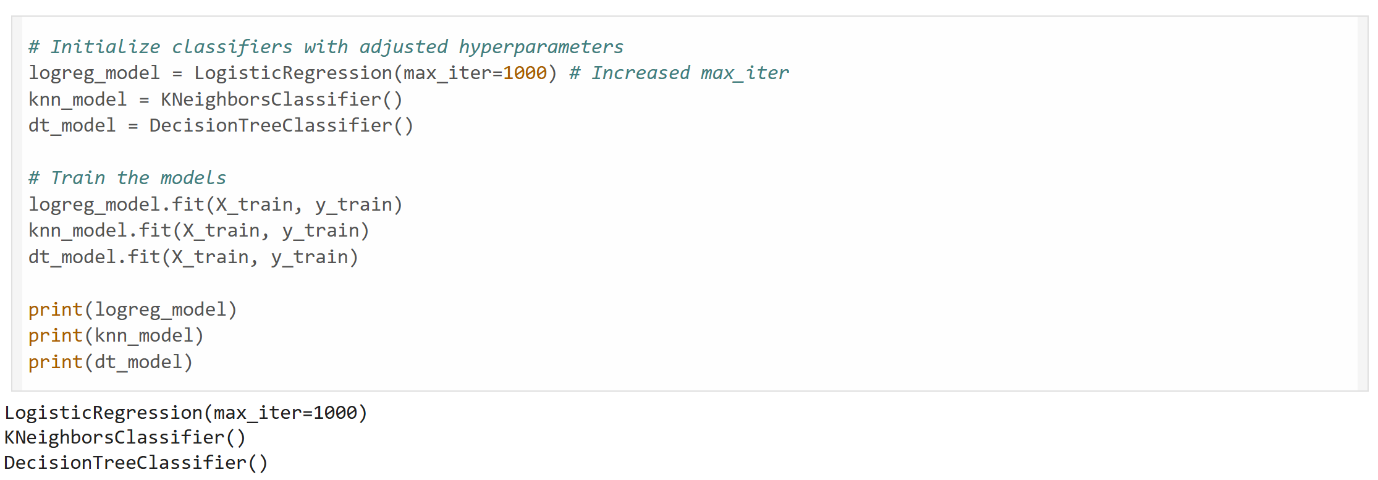
Post\_frequency\_log 149+ Extremely high — redundant with original

Post\_Frequency\_Cubed 34+ High — likely a polynomial correlation

Time\_Spent\_Alone\_Squared 19+ High — again, probably too similar to original

Others like Social\_energy\_gap, Activity\_level inf Likely linear combinations of other variables

**MODELS**

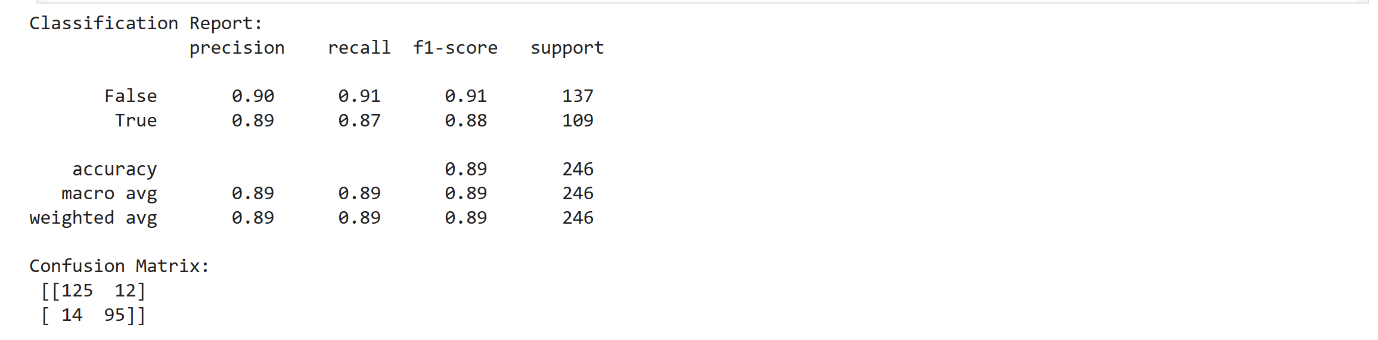
****

Results

A screenshot of a graph

AI-generated content may be incorrect.

Random Forest



**Feature Importance**

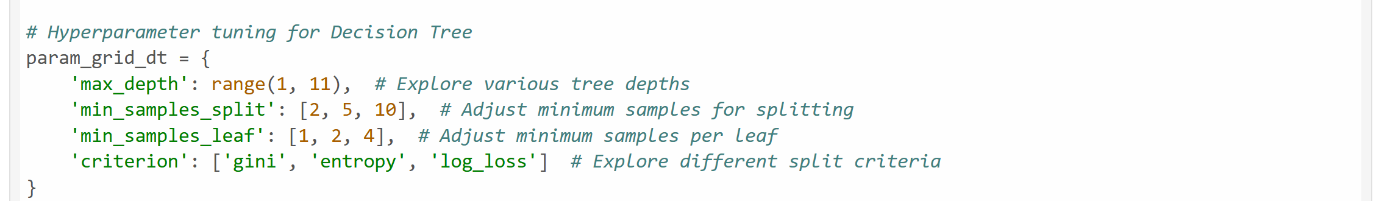
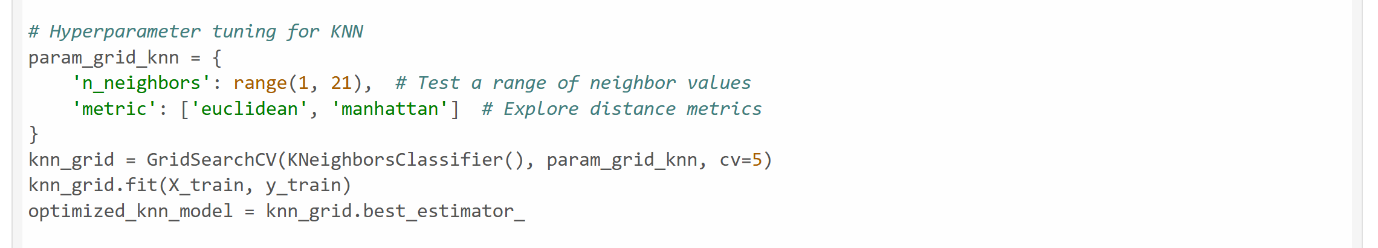
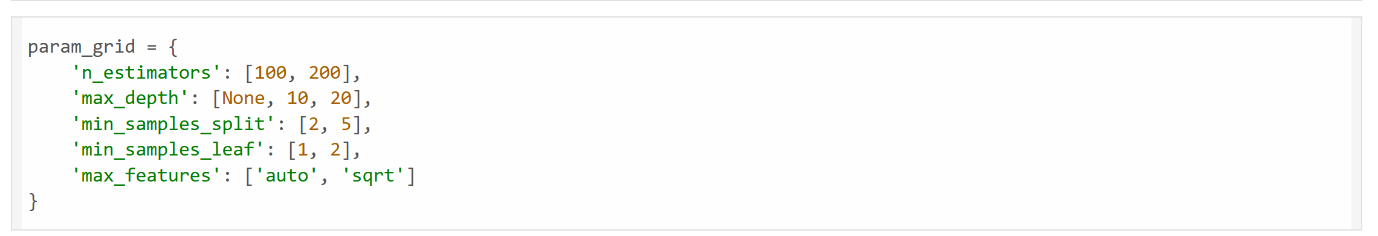
A graph with blue and white bars

AI-generated content may be incorrect.

**Hyperparameter Tuning**

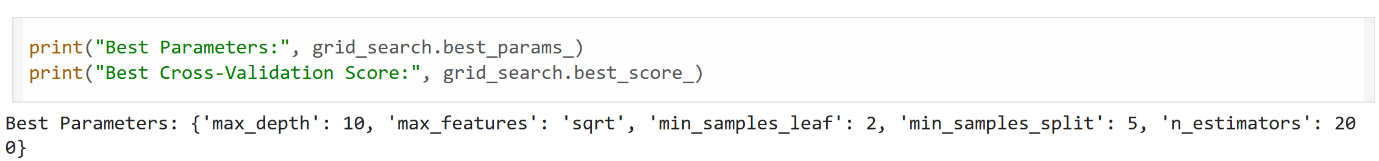
**GRID SEARCH CV**

Parameter Grid

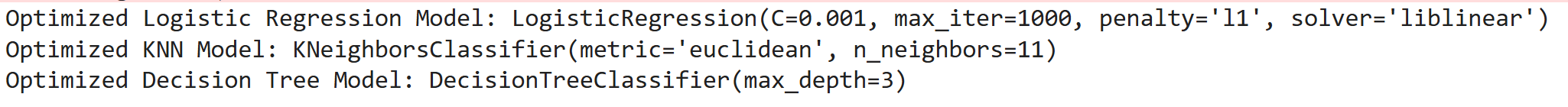
* Decision Tree
* KNN
* Logistic Regression
* Random forest

**Best Parameters**

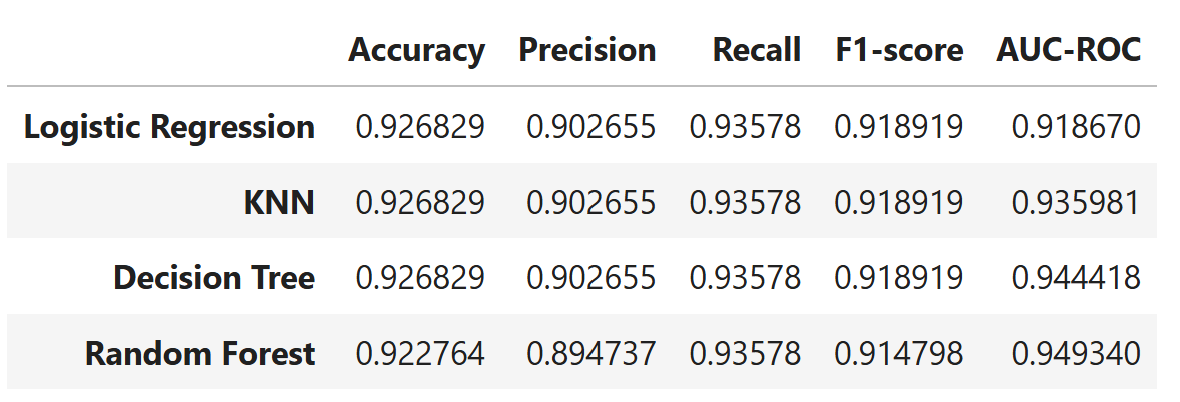
Random Forest



Other Models



**Results**



**Model Performance Comparison**

**Before and After Hyperparameter Tuning**

This section compares the performance of the baseline classification models (Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest) before and after applying hyperparameter tuning. The evaluation metrics considered are Accuracy, Precision, Recall, F1-score, and AUC-ROC.

**Baseline Model Results (Before Hyperparameter Tuning)**

The table below presents the performance of the models using their default hyperparameters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC-ROC** |
| Logistic Regression | 0.918699 | 0.900901 | 0.917431 | 0.909091 | 0.918570 |
| KNN | 0.910569 | 0.884956 | 0.917431 | 0.900901 | 0.911270 |
| Decision Tree | 0.849593 | 0.833333 | 0.825688 | 0.829493 | 0.847151 |
| Random Forest | 0.890000 | 0.890000 | 0.890000 | 0.890000 | N/A |

From the baseline results:

* **Logistic Regression** showed the strongest performance among the default models, particularly in Accuracy and F1-score.
* **KNN** was competitive, slightly underperforming Logistic Regression.
* **Random Forest** showed a promising baseline performance with an accuracy of 0.89, outperforming the Decision Tree. Its precision, recall, and F1-score were also consistently at 0.89. The specific AUC-ROC for the untuned Random Forest was not directly available in the provided classification report.
* **Decision Tree** had the lowest performance across all metrics, indicating potential for significant improvement through tuning.

**Tuned Model Results (After Hyperparameter Tuning)**

The following table displays the performance of the models after hyperparameter tuning (presumably using techniques like GridSearchCV):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC-ROC** |
| Logistic Regression | 0.926829 | 0.902655 | 0.935780 | 0.918919 | 0.918670 |
| KNN | 0.926829 | 0.902655 | 0.935780 | 0.918919 | 0.935981 |
| Decision Tree | 0.926829 | 0.902655 | 0.935780 | 0.918919 | 0.944418 |
| Random Forest | 0.922764 | 0.894737 | 0.935780 | 0.914798 | 0.949340 |

**Comparison and Analysis of Tuning Impact**

A direct comparison of the "before" and "after" results reveals the significant impact of hyperparameter tuning:

* **Overall Improvement:** All four models show a marked improvement in performance across almost all metrics after tuning, demonstrating the value of optimization.
* **Logistic Regression:**
  + Accuracy increased from 0.918699 to 0.926829.
  + Precision saw a slight increase from 0.900901 to 0.902655.
  + Recall significantly improved from 0.917431 to 0.935780.
  + F1-score increased from 0.909091 to 0.918919.
  + AUC-ROC remained largely stable, with a minor increase from 0.918570 to 0.918670.
* **KNN:**
  + Accuracy dramatically improved from 0.910569 to 0.926829.
  + Precision increased from 0.884956 to 0.902655.
  + Recall saw a substantial jump from 0.917431 to 0.935780.
  + F1-score improved from 0.900901 to 0.918919.
  + AUC-ROC showed a notable improvement from 0.911270 to 0.935981.
* **Decision Tree:** This model experienced the most significant gains, demonstrating the effectiveness of tuning for models that might initially underperform with default settings.
  + Accuracy soared from 0.849593 to 0.926829.
  + Precision increased from 0.833333 to 0.902655.
  + Recall dramatically improved from 0.825688 to 0.935780.
  + F1-score jumped from 0.829493 to 0.918919.
  + AUC-ROC saw a substantial increase from 0.847151 to 0.944418.
* **Random Forest:**
  + Accuracy improved from 0.89 to 0.922764.
  + Precision slightly increased from 0.89 to 0.894737.
  + Recall significantly improved from 0.89 to 0.935780.
  + F1-score improved from 0.89 to 0.914798.
  + AUC-ROC, where available, showed strong performance after tuning, reaching 0.949340, making it the highest among all models in this metric.

**Conclusion on Tuning Impact:**

Hyperparameter tuning was highly effective in optimizing the performance of all four models. Notably, the Decision Tree, which was the weakest baseline model, achieved comparable accuracy, precision, recall, and F1-scores to Logistic Regression and KNN after tuning. The Random Forest model, while already strong at baseline, also saw improvements, particularly in Recall and F1-score, and achieved the highest AUC-ROC score (0.949340) after tuning. This indicates that tuning successfully addressed the limitations of the default models and allowed the ensemble methods (Random Forest) to further leverage the data effectively. The improved metrics across the board confirm that the tuned models are more robust and reliable for inferring user personality types, with Random Forest showing the best overall discrimination capability as indicated by its AUC-ROC score.

