Assignment ML Pipeline

Assignment submitted by:

Muhammad Haroon Jan

# Dataset:

I used dataset from kaggle that is **mdsultanulislamovi/student-stress-monitoring-datasets**. The data set insights are as follows:

* **Data Structure:** The dataset contains 843 entries and 26 columns. Most columns are of integer type, except for one object type column ('Which type of stress do you primarily experience?').
* **Missing Values:** There are no missing values in the dataset.
* **Duplicates:** There are no duplicate rows in the dataset.
* **Age Distribution:** The age distribution shows a peak around 19-20 years old, but there is also a single outlier at 100 years old.
* **Feature Importance:** The Random Forest model's feature importance plot suggests that 'Academic and extracurricular activities conflicting for you?' and 'Have you gained/lost weight?' are among the most important features in predicting the target variable.

# Model comparison table:

|  |  |  |
| --- | --- | --- |
| Model | Baseline Score | Tuned Score(Cross Validation) |
| KNN | 0.4451 | 0.9375 |
| Decision tree | 1 | 0.95 |
| Random Forest | 0.9939 | 0.9625 |

Based on the performance table, the **Random Forest** model performed the best.

# Summary of Performance

* **Tuned Random Forest:** This model achieved the highest tuned score of **0.9625**, indicating that it had the best performance after hyperparameter optimization.
* **Tuned KNN:** While its baseline score was low, hyperparameter tuning significantly improved its performance to **0.9375**, showcasing the value of optimization.
* **Tuned Decision Tree:** The Decision Tree's baseline score of **1.0** suggests overfitting, as it performed perfectly on the training data. After tuning, its score dropped to **0.95**, a more realistic and generalized performance.