Assignment 3

The objectives of this assignment are as follows:

- Data Analyses: Univariate, Bivariate, and Multivariate Analyses provide foundational insights crucial for feature identification and manipulation in Feature Engineering and subsequently inform the selection and preparation of features for Ensemble Methods.
- Feature Engineering: Techniques for extracting, creating, and selecting features directly draw upon the insights gained from Data Analyses. Hands-on exploration of feature impacts reinforces the understanding of how feature modifications influence model performance.
- Ensemble Methods: Leveraging well-engineered features and models from Feature Engineering, Ensemble Methods combine multiple models to enhance predictive accuracy and robustness. They capitalize on the refined features and diverse models created through prior stages.
- Hyperparameter Tuning: The optimization of hyperparameters further refines model performance. Techniques like grid search and cross-validation ensure models are tuned for optimal accuracy and robustness.

The cohesive integration of Data Analyses, Feature Engineering, Ensemble Methods, and Hyperparameter Tuning creates a cyclical process wherein insights from data analysis inform the creation of well-engineered features, subsequently enhancing ensemble methods' performance for more accurate predictions in machine learning tasks.

Problem Statement

Smoking remains a significant contributor to various health complications and is responsible for preventable illnesses and millions of deaths globally, with projections indicating a surge to 10 million smoking-related deaths by 2030. Despite concerted efforts, the complexity of factors influencing smoking cessation has hindered high success rates in quitting.

To address this challenge, the problem at hand is to utilize machine learning to predict an individual's smoking status using bio-signals. The objective is to create a predictive model that considers multiple factors such as nicotine dependency, carbon monoxide levels, daily cigarette consumption, age of smoking initiation, previous quit attempts, emotional well-being, personality traits, and motivation to quit. This model aims to assist healthcare professionals and individuals in assessing the likelihood of successfully quitting smoking, thereby improving smoking cessation outcomes.

Assignment Details

1. **Data Analysis:**

    - **Univariate Analysis**: Understanding the distribution and characteristics of individual features provides crucial insights into their relevance and potential impact on model performance. This analysis aids in identifying which features might be significant for further manipulation in Feature Engineering.

- **Bivariate Analysis**: Examining relationships between pairs of features helps identify dependencies or correlations. These identified patterns guide the selection of feature combinations that might contribute significantly to model performance during Feature Engineering.
- **Multivariate Analysis**: Assessing complex interactions among multiple features offers a deeper understanding of their combined effects. This analysis assists in feature selection and dimensionality reduction techniques employed in Feature Engineering.

2. **Feature Engineering and Preprocessing:**
   - **Techniques for Feature Engineering**: Leveraging insights from Univariate, Bivariate, and Multivariate Analyses, techniques for extracting, creating, and selecting features are employed. This step involves transforming raw data into meaningful representations that positively impact model accuracy and performance.
   - **Exploration of Feature Impacts**: Hands-on exercises involving the manipulation of features based on data analysis findings provide a practical understanding of how feature modifications directly influence model outcomes.
3. **Ensemble Methods:**

   You are **required** to **implement** the following ensemble methods:
   - **Bagging**: Utilizes bootstrapping to train multiple models independently and combines their predictions through averaging or voting.
   - **Boosting:** Builds a sequence of models, each correcting errors made by the previous ones by assigning different weights to data instances.
   - **Random Forests:** Implement a Random Forest class or functions to combine multiple Decision trees using the DecisionTreeClassifier from scikit-learn.

   These ensemble techniques leverage the enhanced features and models created through Feature Engineering. The refined features and diverse models contribute to improved predictive accuracy and robustness when combined using these ensemble methods.
4. **Hyperparameter Tuning**:
   - **Grid Search**: Exhaustively searches through a specified parameter grid to identify the best combination of hyperparameters.
   - **Randomized Search**: Randomly samples hyperparameters from specified distributions to efficiently find optimal settings.
5. **Final System**:
   - After Finishing all of you analysis, trying different ensemble methods, and finetuning them, you will end up with a final Machine learning system with the best performance.

Assignment Requirements:

1. In Data Analysis Part You should do all the 3 types of analysis, we expect that you explore at least 20 of the given features.

   - **We Expect a lot of visualizations, with your insights.**
   - Emphasize the univariate, bivariate, and multivariate aspects within the data.
   - Organize exploratory and explanatory analyses in separate notebooks.
   - The exploratory notebook resembles an organized draft.

- The explanatory notebook includes only the crucial final graphs.
- In Feature Engineering we are expecting at least normalization with its different types, removing outliers, any extra effort will be appreciated(bonus).
- In Ensemble Methods, you should implement at least the 3 types of ensembles.
- In Hyperparameter Tuning **,** You should use one of them , If you want to explore Bayesian methods will be appreciated(bonus).

Extra Notes

- Each team will be assigned a variation of the data. Please make sure you use the correct dataset.
    - Generate your dataset variant by adding and using the script in generate_data.py in your notebook.
- You **MUST** split your dataset into train, valid and test as usual.
- If you don't know how to add your analysis in an organized manner add a report.
- You should deliver your notebooks as pdf, also the original notebook.
    - You should have one for exploratory analysis.
    - You should have one for explanatory analysis.
    - You should have one for the analysis.
- You are not limited to what have been taught in the class, you can explore extra frameworks for training , like catboost , xgboost also for the visualization of data analysis you can explore matplotlib , plotly , seaborn
- You should work in groups of three. Each team should have one submission Id1_id2_id3.zip.
- Delivery will be ignored if you didn't follow the naming scheme provided in 4, any one of the team ids can be used.
- **Copies will get zeros.**
- **Assignment Deadline will be 22nd December, please start early.**

Dr. Marwan Torki

Eng. Abdallah Ashraf
Eng. Youssef Nawar
Eng. Mohamed Badran
Eng. Zeyad Ezzat

Grading Scheme:

- Data Analysis 25%
- Feature Engineering 25%
- Ensemble Methods 30%
- Hyperparameter Tunning 10%
- Final System 10%
- Bonus 20% (If you have actually good work relative to the rest of the students)

Dr. Marwan Torki

Eng. Abdallah Ashraf Eng. Mohamed Badran
Eng. Youssef Nawar Eng. Zeyad Ezzat