

WORCESTER POLYTECHNIC INSTITUTE

**PROJECT 2 – CLASSIFICATION**

**PREDICTING DRINKING BEHAVIOR WITH BODY SIGNAL DATA**

Sheroz Shaikh & Jhih-ci Chao

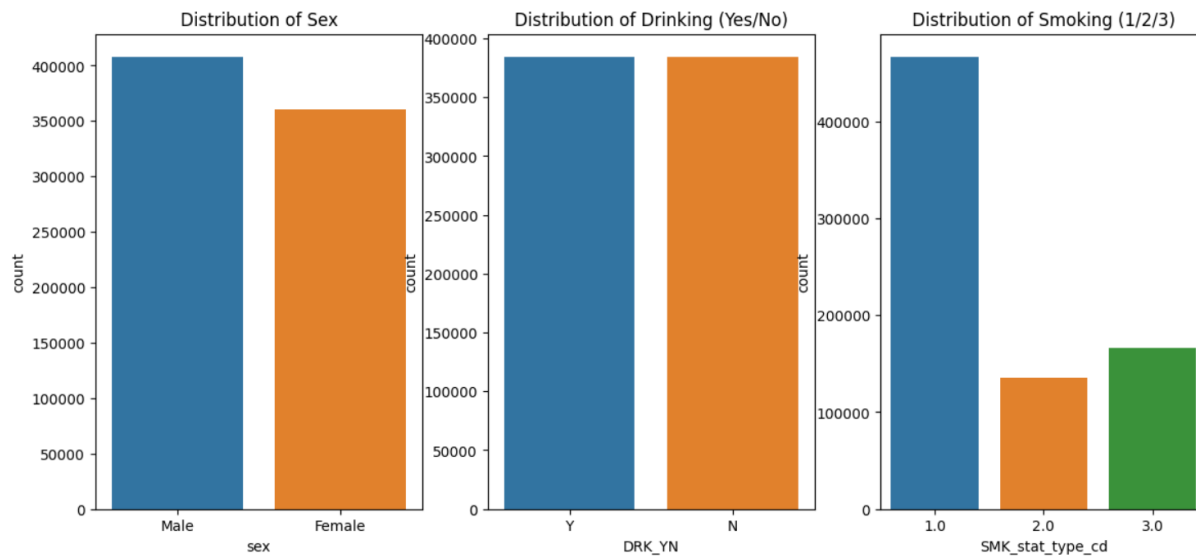
CS 548 – Knowledge Discovery & Data Mining

Prof. Roe Shraga

10/12/2023

## **Task 2 – EDA [10 points]**

### ***1. Categorical Variable Analysis***

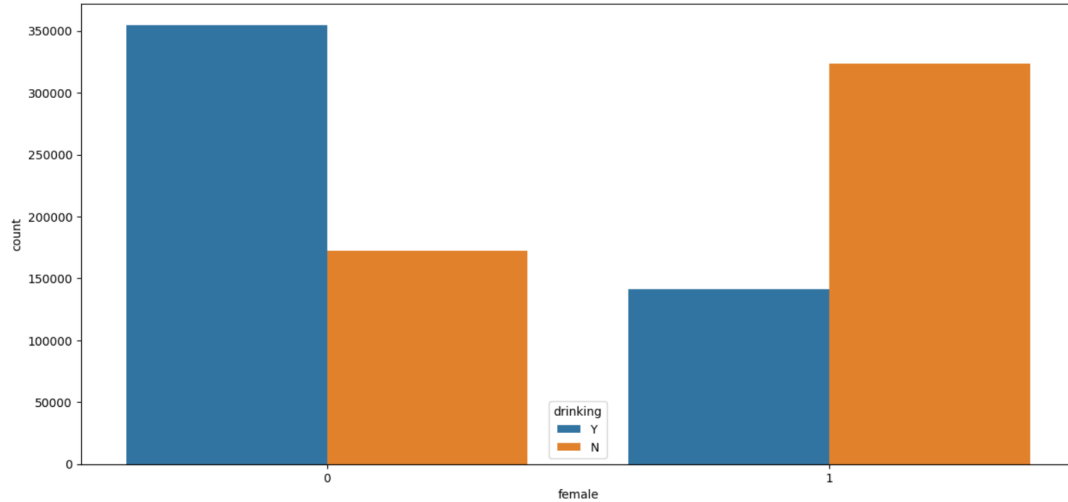


Understanding the distribution of categorical variables, like sex, drinking habits, and smoking status, provides a foundational understanding of the population being analyzed. Summarizing the counts of these categorical variables is crucial as it aids in identifying patterns, imbalances, or trends within the dataset. It facilitates a clearer interpretation of subsequent analyses by putting them into context. For instance, knowing the baseline distribution can help detect potential biases, inform sampling strategies, and influence decision-making processes. A balanced or imbalanced distribution in categories can significantly impact the outcomes of studies, making it vital to ascertain these summaries before diving into deeper analyses.

Below are the summaries:

- **Sex:** roughly balanced with slightly more male
- **Drinking:** about half the people drink and half don't.
- **Smoking:** most people never smoke, some still do, and there are less people who quit than who continue to smoke among the group that ever smoked.

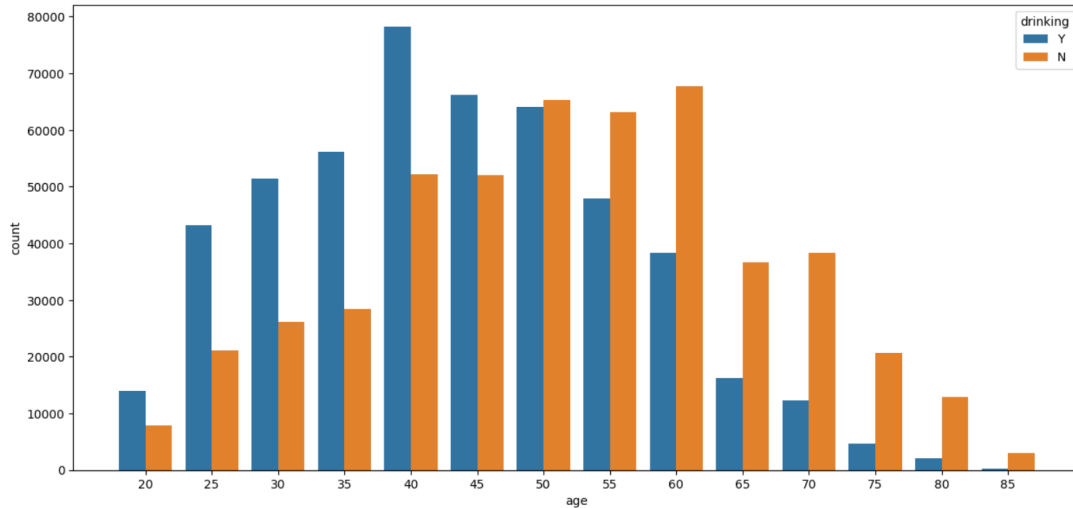
## 2. Gender Differences in Alcohol Consumption Patterns



This is a grouped bar chart representing the count of individuals based on two categorical variables: gender ("female" with categories 0 and 1) and drinking status ("drinking" with categories Y for "Yes" and N for "No"). For the category labeled "0" under "female" (indicating males, or non-females), the count of individuals who drink (Y) is noticeably lower than those who do not drink (N). On the other hand, for the category labeled "1" under "female" (indicating females), the count of individuals who drink (Y) is significantly higher than those who don't drink (N).

The data suggests a distinct difference in drinking habits between the two gender categories. While a larger number of male choose to drink, the opposite trend is observed for females, where a majority appear to not consume alcohol. This could have implications for health studies, marketing strategies, or sociological research, indicating the importance of considering gender differences when addressing alcohol consumption behaviors.

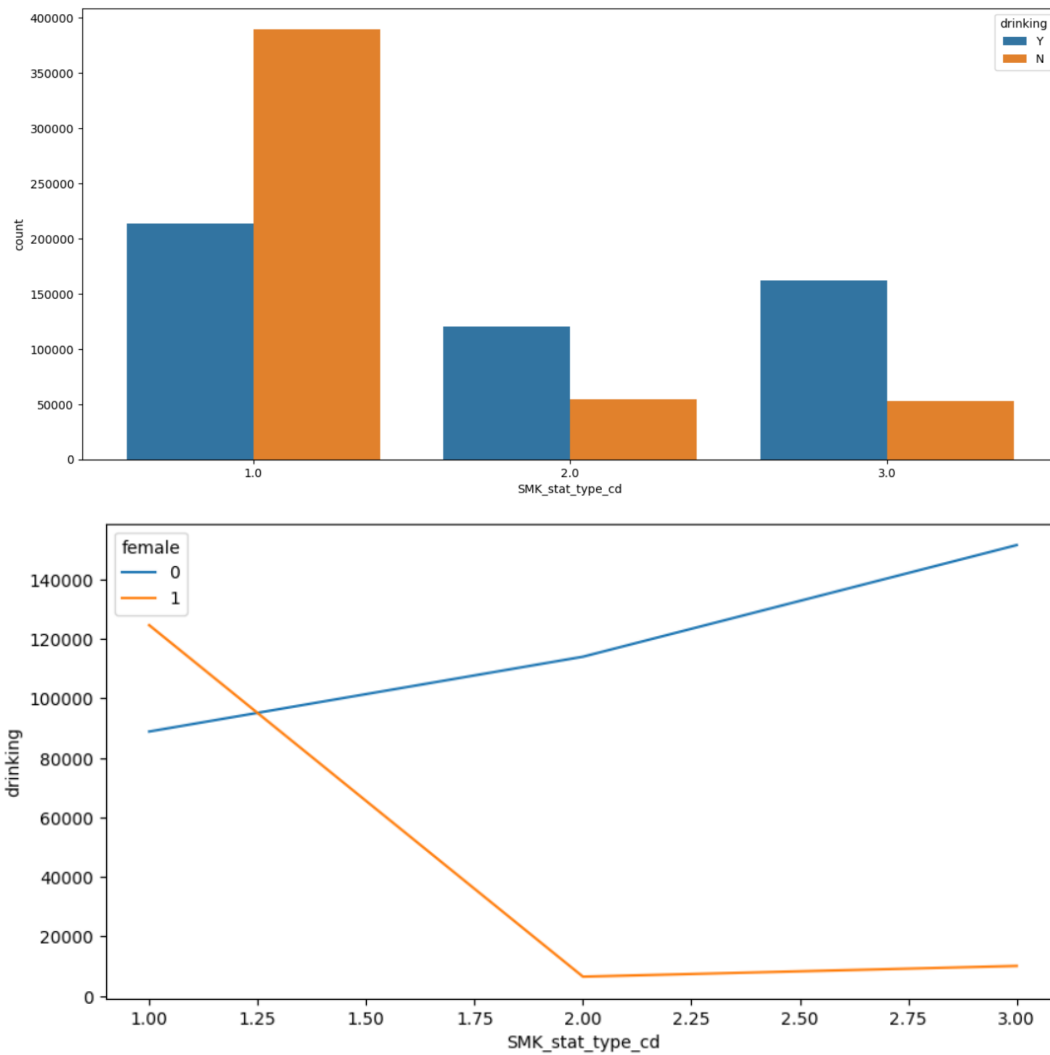
### 3. Age-wise Distribution of Alcohol Consumption



This is a grouped bar chart illustrating the count of individuals based on age and their drinking status. The x-axis represents age groups, starting from 20 and going up to 85 in increments of 5. The y-axis signifies the count of individuals. For each age group, there are two bars: one for individuals who drink (Y) and another for those who don't drink (N). From the ages of 20 to 45, there's a pronounced increase in the number of individuals who drink, peaking at the 40-45 age group. From 45 onwards, the number of individuals who drink starts to decrease, whereas the count of those who don't drink remains relatively stable throughout, with some fluctuations.

Drinking appears to be more prevalent among middle-aged individuals, especially those in the 40-45 age group. After this peak, alcohol consumption seems to gradually decrease with age. This could be due to a variety of reasons including health concerns, lifestyle changes, or societal norms associated with drinking at different life stages. The relatively consistent count of non-drinkers across age groups suggests that the decision not to drink might be influenced by factors other than age. The data might be beneficial for health campaigns targeting specific age groups or for businesses in the beverage industry strategizing their marketing efforts.

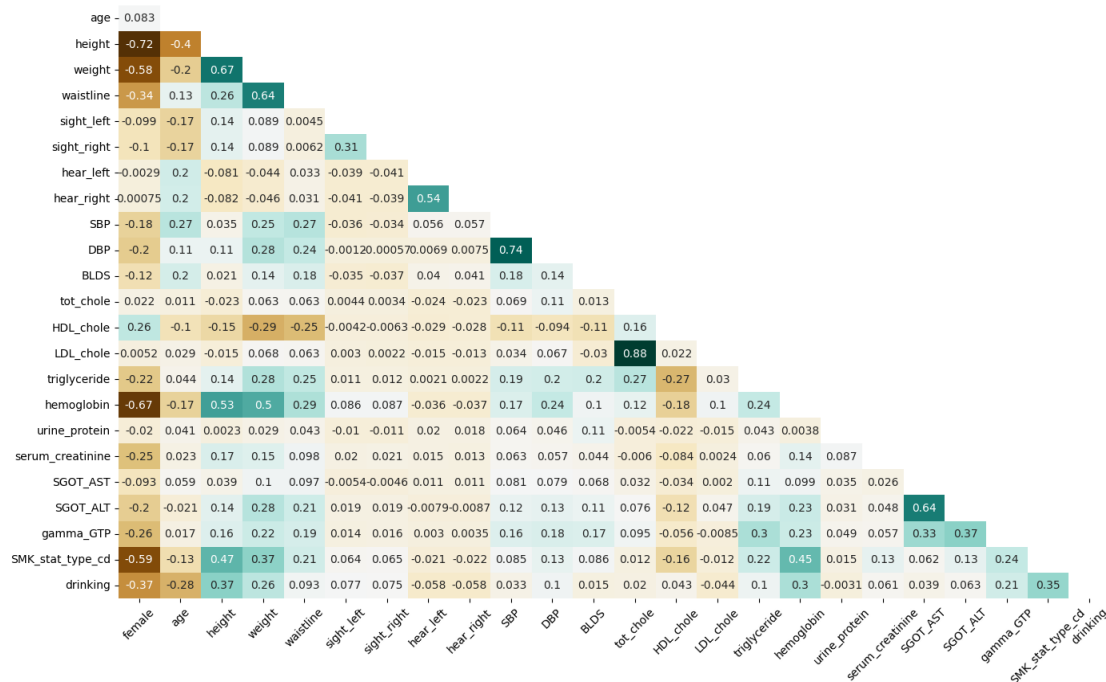
#### 4. Drinking Habits Based on Smoking Status



The first graph is a bar chart depicting the count of individuals based on their drinking habits ("Y" or "N") across the three smoking statuses: never smoked (1.0), smoked but quit (2.0), and still smoking (3.0). A majority of those who never smoked are in the "N" category, while those who ever smoked are more likely to be in the "Y" category. The second graph is a line plot showcasing the relationship between "SMK\_stat\_type\_cd" and "drinking" for two categories denoted as 'female' (0 and 1, male and female). From the data, males show a consistent increase in drinking across all three smoking statuses, while females exhibit relatively extreme drinking levels, with a steep dip for those who smoked but quit, followed by a slight rise for current smokers.

The implications are that there's a discernible relationship between smoking status and drinking habits. Those who never smoked predominantly fall into the "N" drinking category. The drinking levels of male individuals increase in line with their smoking status progression, but the trend appears to be opposite for females. This data suggests gender-based differences in the interaction between smoking and drinking habits.

## 5. Heatmap Analysis



This heatmap effectively visualizes relationships between various health metrics, enabling easier identification of patterns and correlations that might be critical for medical analysis or lifestyle recommendations. Below are some notable correlations and insights extracted from the heatmap that may influence the target attributes (smoking status and drinking):

### Smoking Status:

- Moderate positive correlation: height, weight, and hemoglobin, drinking
- Weak positive correlation: waistline, triglyceride, gamma\_GTP
- Moderate negative correlation: age, HDL\_chole
- Strong negative correlation: (being) female

### Drinking Status:

- Moderate positive correlation: height, weight, hemoglobin, SMK\_stat\_type\_cd
- Weak positive correlation: DBP, triglyceride, gamma\_GTP
- Moderate negative correlation: age, (being) female

The correlations above, especially the strong negative correlation of being female, suggest inherent physiological or societal differences based on gender. The positive correlation between drinking and metrics like weight and hemoglobin might hint at lifestyle patterns or dietary habits associated with those who consume alcohol. The negative correlation between age and drinking could indicate changing consumption habits as individuals age or potential generational differences in drinking patterns. Correlations like these can be vital in health studies, policy-making, or targeted interventions, especially when considering factors like gender and age. Interestingly, most biological metrics don't seem to be related with smoking and drinking; in fact, most metrics don't seem to be related to each other, except LDL\_chole vs. TOT\_chole (LDL and total cholesterol), DBP vs. SBP (diastolic vs. systolic blood pressure), SGOT\_ALT vs. SGOT\_AST, that show strong or moderate correlation.

### **Task 3 – Problem Definition [5 points]**

Understanding the multifaceted relationship between lifestyle choices like smoking and drinking and their impact on health is not just academic—it's a public health imperative. Our study dives into a rich dataset from Korea's National Health Insurance Service, encompassing around 991,000 records with 24 features aimed at classifying smoking and drinking behaviors. The dataset offers a compelling mix of complexity and relevance, making it an ideal playground for cutting-edge classification algorithms.

#### ***1. Problem Statement and Observations***

While lifestyle choices are personal, their consequences reverberate through public health systems. Our initial Exploratory Data Analysis (EDA) uncovers some intriguing patterns that make this study timely and significant:

- **Gender Differences:** Alcohol consumption is noticeably gender-biased, with more males opting to drink than females. This hints at the need for gender-specific health interventions.
- **Age Groups:** Alcohol consumption peaks in the 40-45 age group and tapers off thereafter, suggesting the middle-aged demographic could be a key target for preventative health measures.
- **Smoking and Drinking:** A clear interaction exists between smoking and drinking habits. Those who have never smoked predominantly do not drink, implying the need for integrated public health campaigns.
- **Health Metrics Correlations:** Strong and moderate correlations exist between specific health metrics and smoking or drinking status. For example, a strong negative correlation exists between being female and smoking, and moderate positive correlations exist between metrics like height, weight, and hemoglobin with both smoking and drinking.

#### ***2. Motivation***

Chronic diseases stemming from smoking and drinking are rampant and often fatal. Early identification of high-risk individuals can guide timely interventions, possibly averting severe health issues. Moreover, public health policies can be refined based on targeted demographics unveiled through this analysis. The dataset's volume and feature diversity allow us to tackle these issues through rigorous machine learning workflows, from data preprocessing to advanced statistical analysis.

#### ***3. Significance***

The endgame is a robust predictive model to identify individuals at high risk, offering healthcare providers a valuable tool for early intervention. Beyond healthcare, the insights gained can influence marketing strategies in the beverage industry and assist policymakers in framing gender- and age-specific health campaigns.

By dissecting this intricate web of lifestyle choices and their physiological ramifications, we aim not only for early disease detection but also to nudge people towards healthier decisions, thereby reducing healthcare costs and enhancing public well-being.

## **Task 4 – Preprocessing [15 points]**

### ***Data Cleaning***

Data cleaning is the first step in preparing the dataset for model training. It's vital to ensure the integrity and uniformity of the data, making it easier to work with. In this section, we focus on renaming columns for better readability, mapping categorical values to numerical ones, and selecting relevant columns for further analysis.

#### ***1. Column Renaming and Value Mapping (see Section 4.1 on the ipynb)***

- `sex` renamed to `female` and values mapped to 0 for Male and 1 for Female.
- `DRK_YN` renamed to `drinking`.
- *Purpose:* To make the column names and values more interpretable and consistent.

#### ***2. Column Selection (Section 4.2)***

- Specific columns chosen relevant to the task.
- *Purpose:* To narrow down the scope and focus on predicting drinking behavior.
- `Independent_variables = ['age', 'BLDS', 'female', 'gamma_GTP', 'HDL_chole', 'hemoglobin', 'SBP', 'SMK_stat_type_cd', 'tot_chole', 'triglyceride', 'weight',]`
- `dependent_variables = ['drinking']`

### ***Data Transformation & Normalization***

After cleaning, the data must be transformed into a form that is more suitable for analysis and model training. This often involves scaling features and converting categorical variables into a numerical format. In this section, we apply various scaling techniques and map categorical variables to ensure the dataset is ready for machine learning algorithms.

#### ***1. Feature Scaling - Min-Max Scaler (Section 4.3)***

- Applied to `independent_variables`.
- *Purpose:* MinMax scaling brings all values in the dataset within the range [0, 1], which is often desirable for many ML algorithms sensitive to feature scale.
- *Output:* `processed_data_4`

#### ***2. Feature Scaling - Standard Scaler (Section 4.4)***

- Applied to `independent_variables`.
- *Purpose:* Z-score normalization ensures that each feature has mean=0 and variance=1, useful for algorithms that assume features are centered around zero.
- *Output:* `processed_data_5`



3. Categorical Mapping (Section 4.5)

- Mapped categorical features like SBP and BLDS .
- **Purpose:** many algorithms require numerical input.
- **Output:** changes are reflected in the original df1

4. Feature Scaling with Categorical Mapping - Min-Max Scalar (Section 4.6)

- Same as scaling 1, only applied after categorical mapping.
- **Output:** processed\_data\_2

5. Feature Scaling with Categorical Mapping - Standard Scalar (Section 4.7)

- Same as scaling 2, only applied after categorical mapping.
- **Output:** processed\_data\_3

## **Task 5 – Model Selection, Training, and Optimization [30 points]**

### ***Model Selection and Training***

This phase involves selecting and training machine learning classifiers suitable for the task at hand. In the current scope, classifiers ranging from logistic regression to ensemble methods and neural networks are chosen. These classifiers are trained on various versions of the dataset, including raw and preprocessed data, to observe their initial performance.

Models used include:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Extra Trees
5. Support Vector Machine (SVM)
6. Gaussian Naive Bayes
7. Multinomial Naive Bayes
8. K-Nearest Neighbors (K-NN)
9. AdaBoost
10. Gradient Boosting
11. Hist Gradient Boosting
12. Bagging
13. Neural Network (MLPClassifier)

Each model is trained on five different versions of the dataset: raw data, MinMax-scaled data, and Standard-scaled data, both before and after categorical mapping. Metrics like Accuracy, Precision, Recall, F1-Score, ROC AUC, and Matthews Correlation Coefficient are recorded for performance evaluation.

### ***Model Optimization***

After initial training, the next step is to optimize the models to improve their performance. Three main techniques are employed: hyperparameter tuning via Randomized Grid Search, and performance comparison across differently preprocessed datasets. The search summary, with the best results for each model with different datasets, are shown in the [📄 Project 2 - Random Search Summary](#) spreadsheet.

## **Task 6 – Model Evaluation [30 points]**


In the final phase of the project, the trained machine learning models are rigorously evaluated using a range of metrics. This step is crucial for understanding the efficacy of each model in a multi-faceted manner, which in turn informs the decision-making process for model deployment or further optimization. For comprehensive evaluation, five metrics have been selected: Accuracy, Precision, Recall, ROC AUC, and Matthew's Correlation Coefficient (MCC).

### ***Evaluation Metrics Used***

- **Accuracy:** Measures the proportion of correct classifications, ranging from 0 to 1.
- **Precision:** Measures the proportion of true positives among positive predictions, ranging from 0 to 1.
- **Recall:** Measures the proportion of true positives among all actual positives, ranging from 0 to 1.
- **ROC AUC:** (Receiver Operating Characteristic - Area Under Curve) Measures the performance of a binary classification model, ranging from 0 to 1.
- **MCC:** Measures the performance of a binary classification model, considering both under- and over-predictions, ranging from -1 to +1.

In the context of drinking detection, these metrics serve different yet complementary roles. Accuracy gives a quick snapshot of how well the model performs overall but might not be sufficient if the classes are imbalanced. Precision is crucial as we don't want to falsely label someone as a drinker, which could lead to unnecessary interventions or stigmas. Recall is also significant because failing to identify someone who is a drinker could mean missing an opportunity for timely intervention. ROC AUC helps us understand the trade-off between sensitivity and specificity, and a high AUC suggests good model performance across different classification thresholds. Finally, MCC offers a balanced view that considers all four quadrants of the confusion matrix, giving a more nuanced picture of model performance than any single metric could provide.

### ***Analysis and Observations***

The  Project 2 - Evaluation Metrics spreadsheet contains the evaluation metrics for all of the selected models with a detailed comparison across five different pre-processing steps (raw (0), MinMax-scaled, Standard-scaled, both before (4, 5) and after (2, 3) categorical mapping). Below is a table with the top models with the metrics of the best pre-processing steps presented:

Model	Accuracy	Precision	Recall	ROC AUC	MCC
Logistic Regression	0.755 (2)	0.748 (2)	0.77 (2)	0.828 (2)	0.51 (2)
Decision Tree	0.670 (2)	0.67 (2)	0.67 (2=3)	0.667 (2)	0.34 (2)
Random Forest	0.725 (2=3)	0.714 (2=3)	0.75 (2=3)	0.791 (2)	0.45 (2=3)
Extra Trees	0.685 (2=3)	0.679 (2=3)	0.70 (2=3)	0.757 (2=3)	0.37 (2=3)
SVM	0.765 (3)	0.726 (3)	0.85 (3)	0.835 (3)	0.54 (3)

## ***Implications***

1. *SVM + Standard Scalar + Categorical Mapping*
  - Highest in all metrics, particularly excelling in Recall and MCC.
  - Indicates a well-rounded and balanced performance.
2. *Logistic Regression + MinMax Scalar + Categorical Mapping*
  - Second in terms of balanced performance based on MCC.
  - Good alternative for scenarios where a balance between Precision and Recall is needed.
3. *Random Forest + MinMax/Standard Scalar + Categorical Mapping*
  - Decent performer, but slightly lower on MCC and doesn't stand out under any metric.
4. *Extra Trees + MinMax/Standard Scalar + Categorical Mapping*
  - Low in MCC, relatively poor performance compared to the top 3.
  - Not suitable for tasks requiring balanced classification for this specific dataset/goal.
5. *Decision Trees + MinMax Scalar + Categorical Mapping*
  - Low in MCC, relatively poor performance compared to the top 3.
  - Not suitable for tasks requiring balanced classification for this specific dataset/goal.

Upon considering all metrics, including the balanced measure of MCC, SVM stands out as the most effective model for this specific classification task, paired with StandardScalar and categorical mapping. Its highest MCC score of 0.54 underscores its balanced performance across all aspects of classification. For this reason, this combination of model and data preprocessing techniques can be prioritized for further tuning or deployment for this specific task if higher performance is required.