Project Walkthrough Guide

Objective:

To build a data-driven model that classifies whether an individual is a **drinker or non-drinker** based on clinical and biometric signals (e.g., liver enzymes, blood pressure, cholesterol, etc.).

How To Read the Notebook:

- Every key code block in SMKData.ipynb aligns with sections in this guide.
- Use SMKData.pdf as the visual reference (plots, tables, figures).
- For model explanations, refer to SHAP plots (page 45–47 and final SHAP block).

1. Data Understanding and Preparation

1.1. Dataset Context

The dataset contains **health examination data** from Korean adults. It includes lab results, physiological measurements, and smoking/drinking history.

Goal: Use these features to predict **DrinkNum** (1 = Drinker, 0 = Non-Drinker).

1.2. Preprocessing Steps

See these in the notebook around Cells **[35] – [44]** and in SMKData.pdf page **35-39**.

1.2.1. Handling Extreme Values and Skewed Features

Certain lab measurements (e.g., gamma_GTP, SGOT_ALT) showed extreme values and right-skewed distributions.

Fixes Applied:

- **Log Transformation** – Compresses high values.
- **Winsorization** – Caps outliers at 99th percentile.
- **Box-Cox Transformation** – Further normalizes shape where needed.

1.2.2. Feature Engineering

New features were derived:

- **BMI** from weight and height.
- **BMI_Smk** – an interaction term combining BMI and smoking status.

These steps appear in Cells **[43]–[46]** and page **38** of the PDF.

1.2.3. Feature Encoding

- Categorical variables like sex, SMK_stat_type_cd (smoking type) were **numerically encoded**.
- DrinkNum was kept as the target.

See encoding validations on **page 14–15** of SMKData.pdf.

Insights:

- Key clinical variables had significant skew (e.g., gamma_GTP, triglyceride) needing transformation to avoid model bias.
- Interaction features like BMI_Smk and LDL_Drk were retained for their potential to capture compounded risk.
- Drinking behavior (DrinkNum) was evenly distributed, allowing clean binary classification without major class imbalance issues.

Project Walkthrough Guide

2. Multicollinearity and Feature Selection

2.1. VIF (Variance Inflation Factor) Analysis

High VIF values signal redundancy between features. See **page 32** and Cells **[146]–[150]**.

Actions Taken:

- Dropped height, hear_right, hear_left, and sight_right due to high VIF.
- Retained key variables like SBP, DBP despite multicollinearity due to clinical relevance.

Insights:

- High VIF values (>40) identified multicollinear features like height, hear_right, hear_left.
- These were dropped to reduce redundancy without sacrificing clinically relevant predictors.
- Final retained features balanced both statistical rigor and domain importance (e.g., kept waistline and SBP/DBP).

3. Train-Test Splitting and Initial Modeling

3.1. Splitting Strategy

- 70% Training
- 15% Validation
- 15% Testing

Refer to Cell **[56]** and page **42** of SMKData.pdf.

3.2. Baseline Models

Three algorithms were evaluated:

- **Logistic Regression**
- **Random Forest**
- **Gradient Boosting**

Metrics: Accuracy, F1-Score, AUC
See summary table and confusion matrix analysis on **page 44–45**.

Best Performer: Gradient Boosting with F1 ~0.739 and AUC ~0.816.

Insights:

- Gradient Boosting emerged as the strongest base model with an F1 score of ~0.739 and AUC of ~0.816.
- All models showed similar performance (difference <1%), suggesting limited variance in initial feature discriminative power.
- Confusion matrices revealed **false negatives** were the most common error across models.

4. SHAP Explainability (Pre-Tuning)

4.1. SHAP Analysis

SHAP helps explain **how each feature contributes to the prediction**.
Run on GradientBoost before tuning. See Cell **[61]–[63]** and **pages 45–47**.

Top 5 Features by Importance:

1. age
2. SexNum
3. SMK_stat_type_cd
4. gamma_GTP

5. HDL_chole

Each of these had interpretable, clinically plausible impact on drinking prediction.

Insights:

- Top drivers included age, SexNum, gamma_GTP, HDL_chole, and SMK_stat_type_cd.
- gamma_GTP showed a steep risk increase—validating its known link to alcohol consumption.
- SHAP dependency plots suggested non-linear patterns for age and interactions with HDL and BMI.
- Some features (e.g., urine_protein, sight_left) had minimal impact and were deprioritized for future steps.

5. Interaction-Aware Modeling

5.1. Adding Interactions

Interaction terms (especially BMI_Smk) were added to capture **joint effects** of lifestyle and body metrics.
See Cell **[70]–[72]** and page **71**.

Insights:

- Adding BMI_Smk, LDL_Drk, and SBP_DBP improved the model's ability to differentiate drinkers and non-drinkers.
- GradBoost with interactions (GradBoost_Interact) slightly improved validation metrics across all key measures.
- SHAP validation showed BMI_Smk was among the top drivers, confirming its utility.

6. Hyperparameter Optimization (Optuna)

6.1. Gradient Boosting Fine-Tuning

Used **Optuna** to search for the best combination of model parameters to maximize F1-Score.
See the full optimization process and best parameters on **page 70–71**.

Best Settings Found:

- n_estimators: 248
- max_depth: 8
- learning_rate: ~0.013
- min_samples_split: 3
- min_samples_leaf: 5
- subsample: 0.68
- max_features: None

Insights:

- Best model had:
    - n_estimators: 248
    - max_depth: 8
    - learning_rate: ~0.013
    - min_samples_split: 3
    - min_samples_leaf: 5
    - subsample: 0.68
    - max_features: None
- Tuning improved F1 from ~0.739 to **0.74+**, and AUC to **0.819+**, indicating a performance ceiling had been nudged further.

- No signs of overfitting between train/validation scores.

7. Final Model Evaluation

7.1. Post-Tuning Validation Performance

Re-trained model with tuned parameters.

Validation Performance:

- Accuracy: 73.7%
- F1 Score: 74.0%
- AUC Score: 81.9%
  (See Cell **[81]**, page **71**)

7.2. Final Test Set Evaluation

Tested on unseen data (15% split).
See Cell **[83]**, page **72**.

Test Performance:

- Accuracy: 73.3%
- F1 Score: 73.6%
- AUC: 81.7%

Insights:

- Final test set F1 (~0.736) and AUC (~0.817) confirmed that model generalized well beyond validation.
- Slight drop from validation indicates **no data leakage**, and that the model remains robust.

8. SHAP Analysis on Tuned Model

This stage validates that the **model logic still aligns with clinical insights** post-tuning.

Insights:

- SHAP confirmed earlier logic: gamma_GTP, BMI_Smk, age, and HDL_chole remained dominant.
- Patterns were preserved despite parameter tuning—meaning model logic remained interpretable and clinically sound.
- New feature BMI_Smk continued to show additive or interaction-based value, especially in compound risk cases.

9. Conclusion

a. **The Goal**: Build an interpretable, clinically aligned model to predict alcohol consumption using routine medical data.

b. **What Was Done:**
   - Data cleaned and transformed.
   - Redundant variables removed.
   - Top features engineered and evaluated.
   - Multiple models tested and tuned.
   - Final model evaluated .

c. **What We Found:**
   - Age, liver enzymes, BMI-smoking interaction, HDL, and sex explain most risk.
   - Model balances accuracy and recall.