INDIANA UNIVERSITY

Prediction of Smoking and Drinking Behaviors Using NHIS Korea Dataset

By:-

Minju Kim

Vaishnavi Pawar

Saransh Singh

Table of Contents

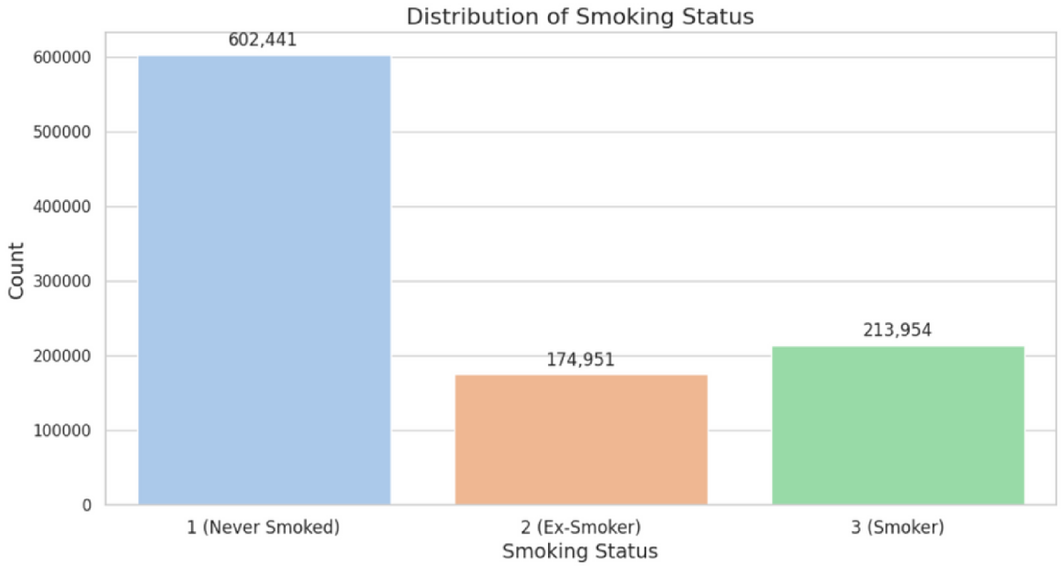
1. **About Dataset**
2. **EDA & Preprocessing**
3. **Machine Learning Models Applied**
4. **Conclusion**
5. **Limitations and Future Development**

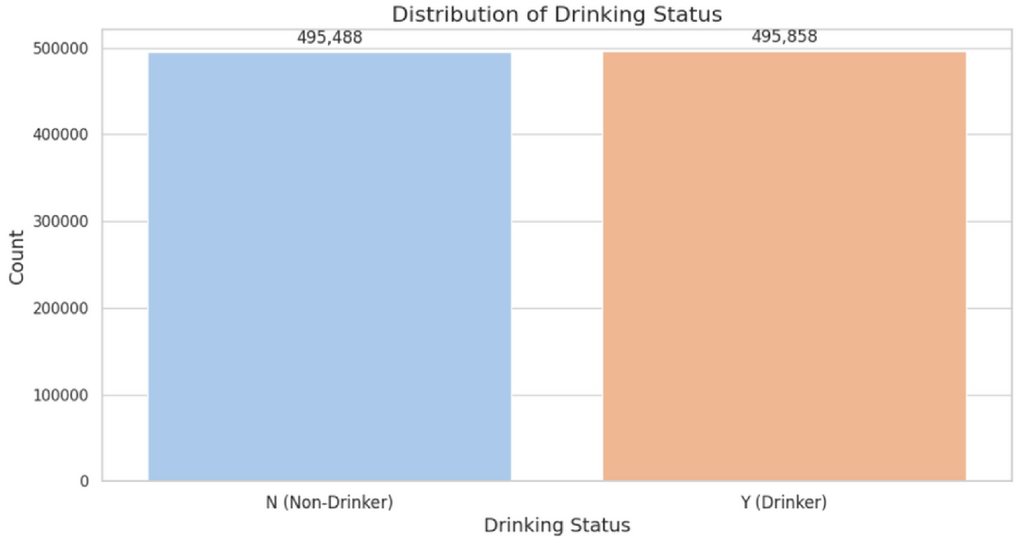
Utilizing Predictive Modeling in Public Health Research

- **Dataset:** "Smoking and Drinking Dataset with body signal" from Kaggle, sourced from NHIS Korea.
- **Purpose:** Create predictive models linking physiological signals to smoking and drinking behaviors.

Key Points:

- **Dataset Source:** Sourced from Kaggle, curated by a user who processed the original NHIS data.
- **Anonymized Data:** Personal and sensitive information has been rigorously removed from the dataset, making it suitable for public use and research applications.
- **Objective:** Develop predictive models to uncover relationships between physiological signals and smoking/drinking behaviors.
- **Impact:** Our research extends beyond Korea, potentially influencing global public health strategies.



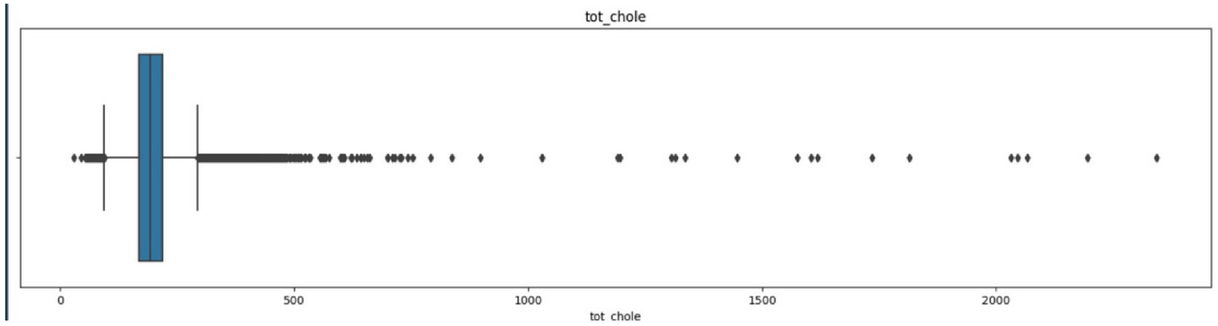


- The "Distribution of Smoking Status" bar chart shows:
  - Over 600,000 never smoked.
  - Approximately 175,000 ex-smokers.
  - Just under 214,000 current smokers.
  - The majority are non-smokers, with current smokers being the least.

- We have employed SMOTE to balance the smoking data by synthesizing records for the less prevalent smoker and ex-smoker categories.

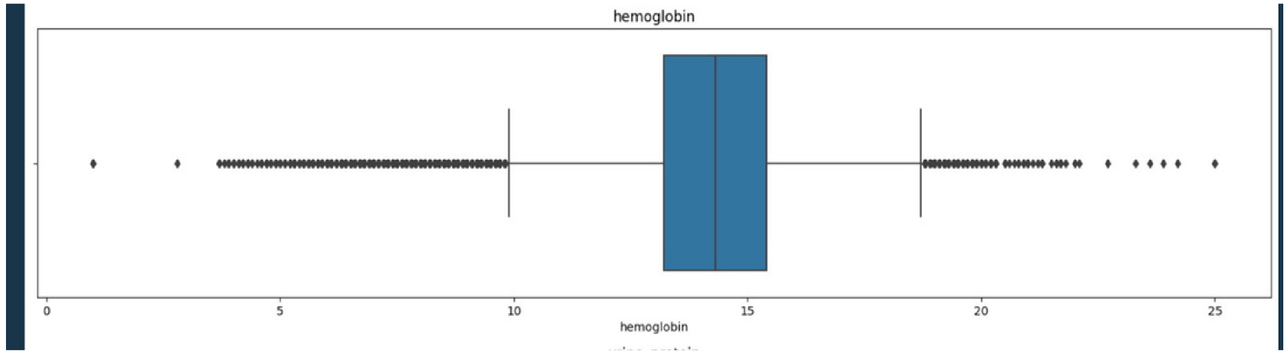
- There are two target variables, we have divided the dataset into two datasets and applied SMOTE to the smoking dataset.

- The "Distribution of Drinking Status" bar chart indicates:
- Close to 495,000 non-drinkers.
- Nearly an equal number of drinkers.
- The population is evenly divided between drinkers and non-drinkers.



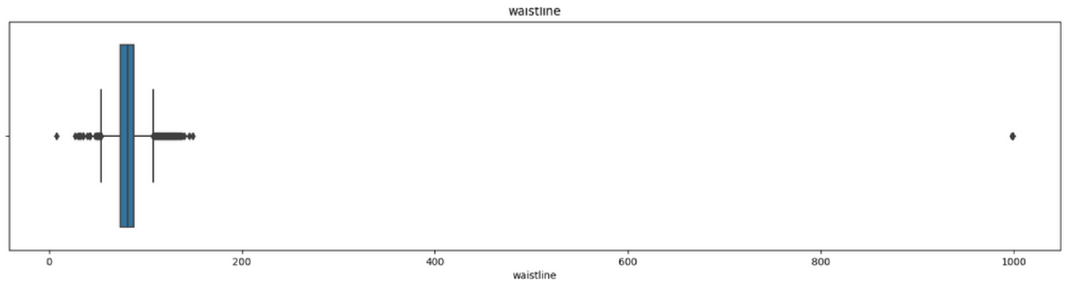
Preprocessing Steps:

Log transformation for right-skewed variables (e.g., BLDS, cholesterol).



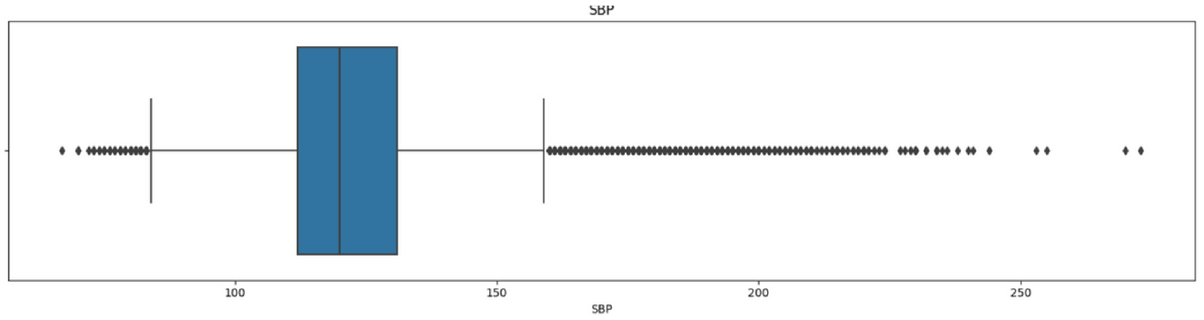
Preprocessing Steps:

No transformation for symmetric variables with few outliers.



Preprocessing Steps:

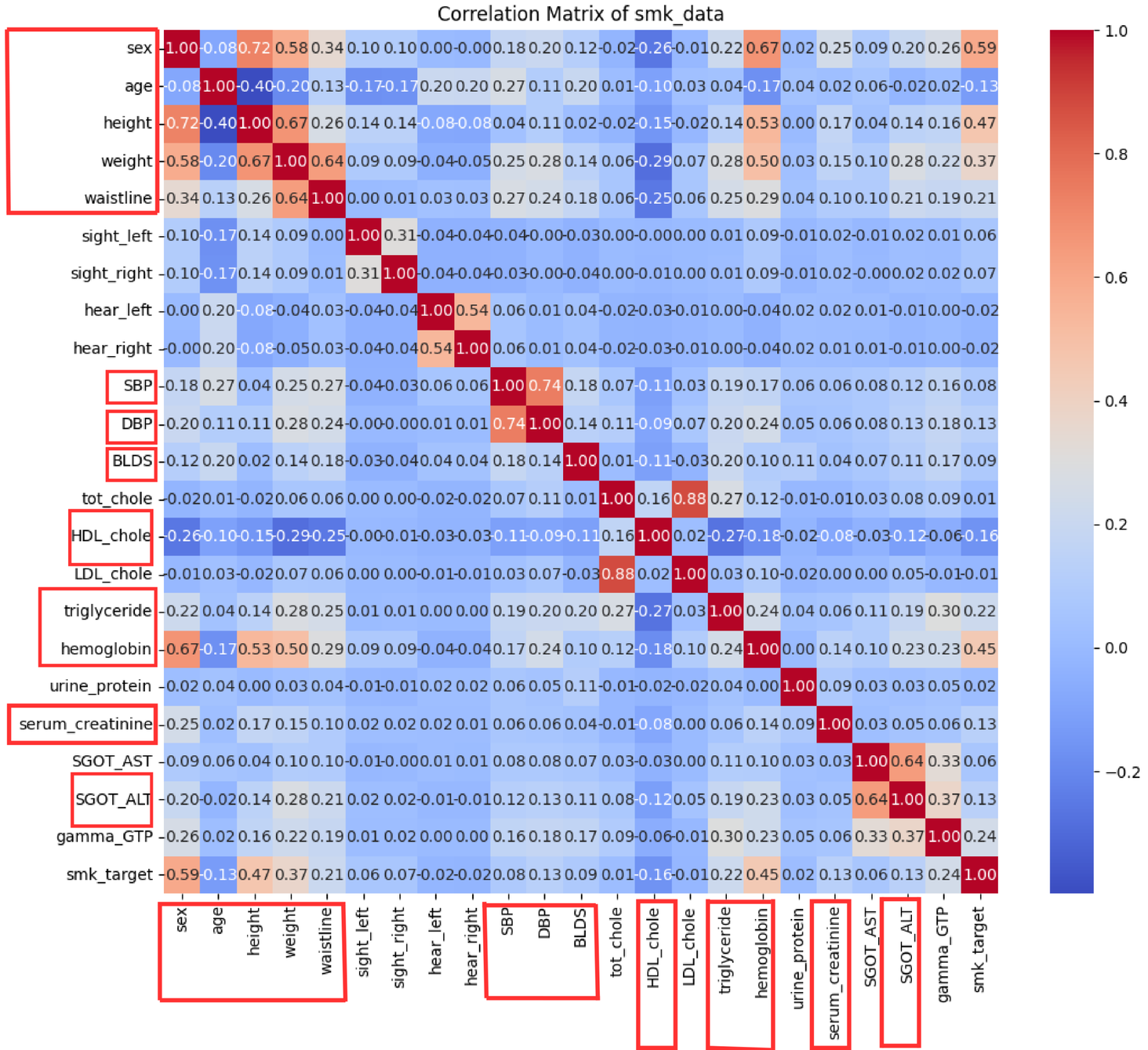
Square root and logarithmic transformations are considered for skewed variables (e.g., total cholesterol, weight, waistline).



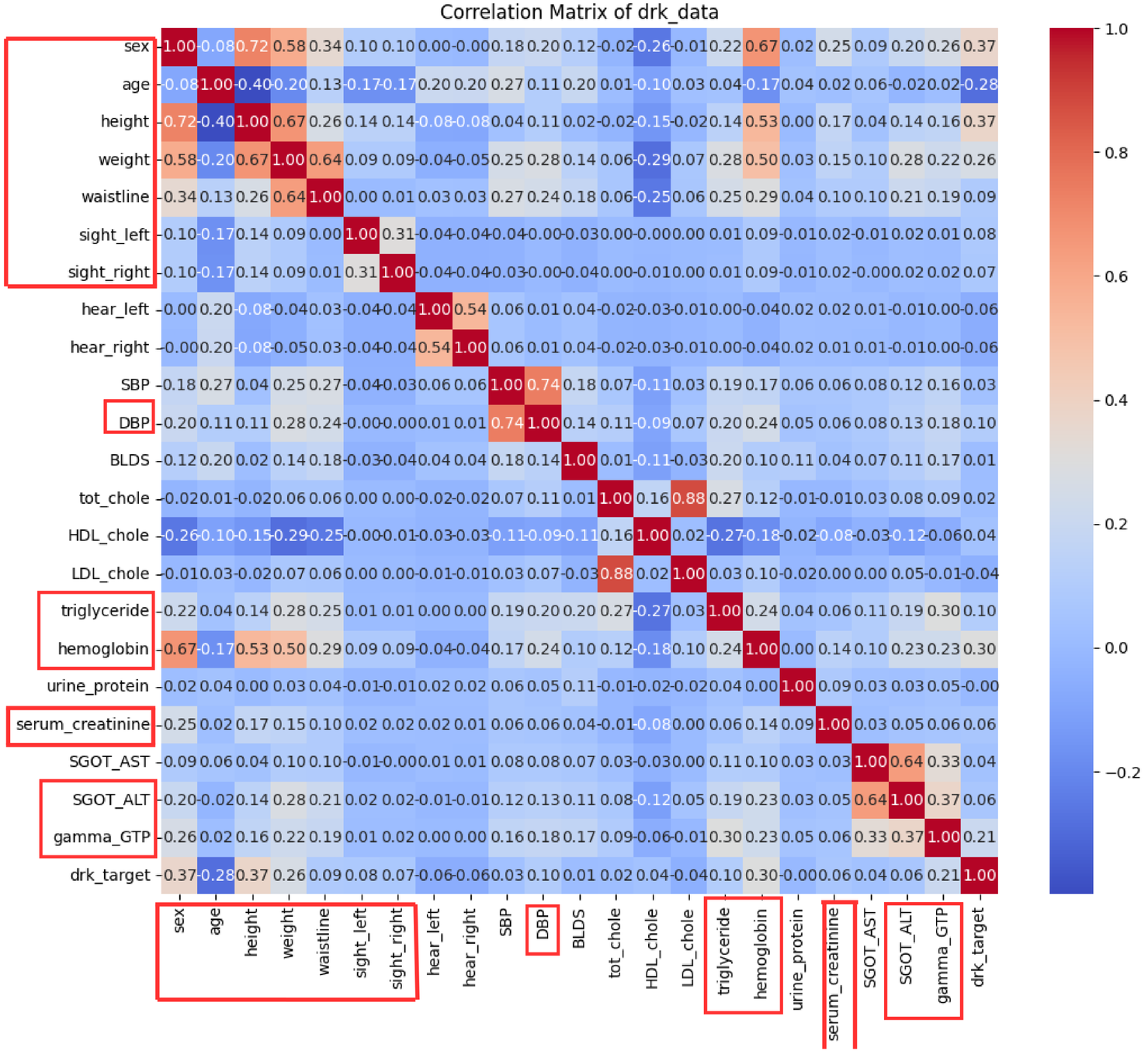
Preprocessing Steps:

Square root transformation chosen for SBP due to mild skewness.

Correlation Matrix of Smoking Dataset



Correlation Matrix of Drinking Dataset



For Smoking Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.6468	0.7005	0.7007	0.6239	0.5951	0.6893	0.6941
Precision	0.5963	0.7065	0.7066	0.5887	0.5699	0.7009	0.7007
Recall	0.6468	0.7005	0.7007	0.6239	0.5951	0.6893	0.6941
F1 Score	0.6006	0.7027	0.7029	0.6003	0.5798	0.6943	0.6954
ROC AUC	0.7642	0.8466	0.8471	0.7004	0.678	0.8368	0.8417

For Drinking Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.7132	0.7231	0.724	0.6584	0.6873	0.7178	0.7178
Precision	0.7135	0.711	0.7144	0.6546	0.7015	0.7065	0.7033
Recall	0.7124	0.7467	0.7464	0.6708	0.6461	0.7396	0.7479
F1 Score	0.713	0.7284	0.7301	0.6626	0.6727	0.7227	0.7249
ROC AUC	0.7826	0.8021	0.8022	0.6999	0.7427	0.7935	0.7958

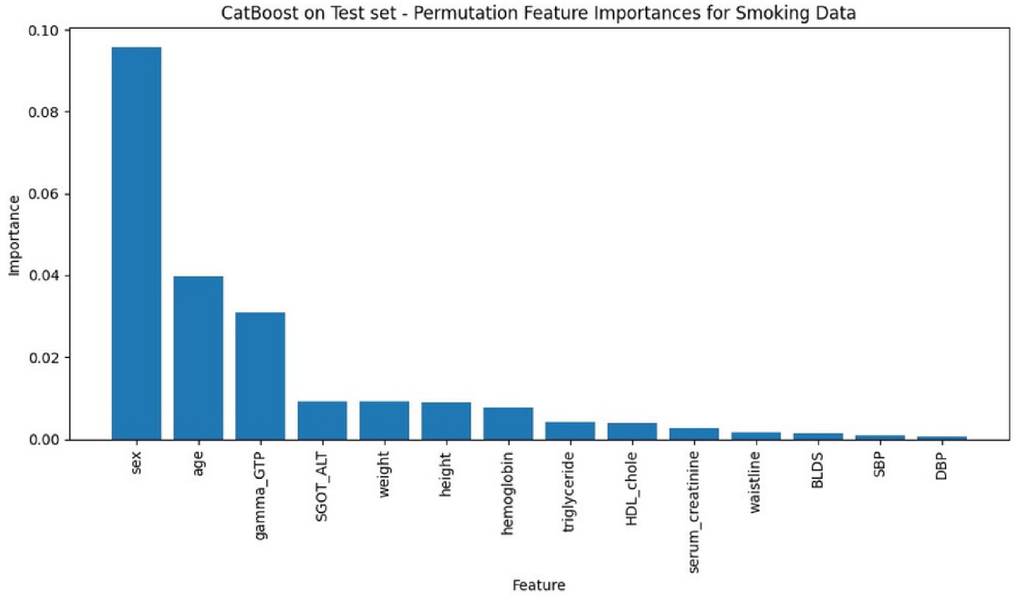
For SMOTE Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.5967	0.6779	0.6769	0.5552	0.5889	0.6165	0.6865
Precision	0.6579	0.7469	0.7471	0.62	0.5722	0.6288	0.7194
Recall	0.5967	0.6779	0.6769	0.5552	0.5889	0.6165	0.6865
F1 Score	0.6174	0.6996	0.6987	0.5782	0.5795	0.6222	0.6992
ROC AUC	0.9435	0.8377	0.836	0.6772	0.6771	0.6515	0.8341

For Smoking Dataset			
	Logistic Regression	Final CatBoost	Final LightGBM
Accuracy	0.6468	0.7005	0.698
Precision	0.5963	0.706	0.7035
Recall	0.6468	0.7005	0.698
F1 Score	0.6006	0.7025	0.7001
ROC AUC	0.7642	0.8474	0.8458

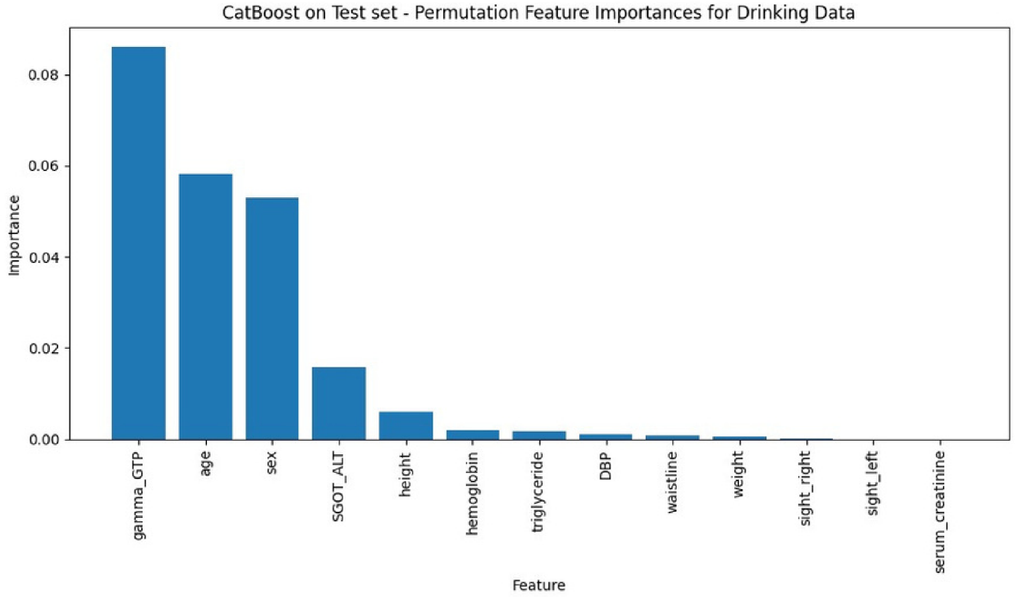
For Drinking Dataset			
	Logistic Regression	Final CatBoost	Final LightGBM
Accuracy	0.6468	0.7232	0.7229
Precision	0.5963	0.7116	0.712
Recall	0.6468	0.7474	0.7453
F1 Score	0.6006	0.729	0.7283
ROC AUC	0.7642	0.8023	0.8016

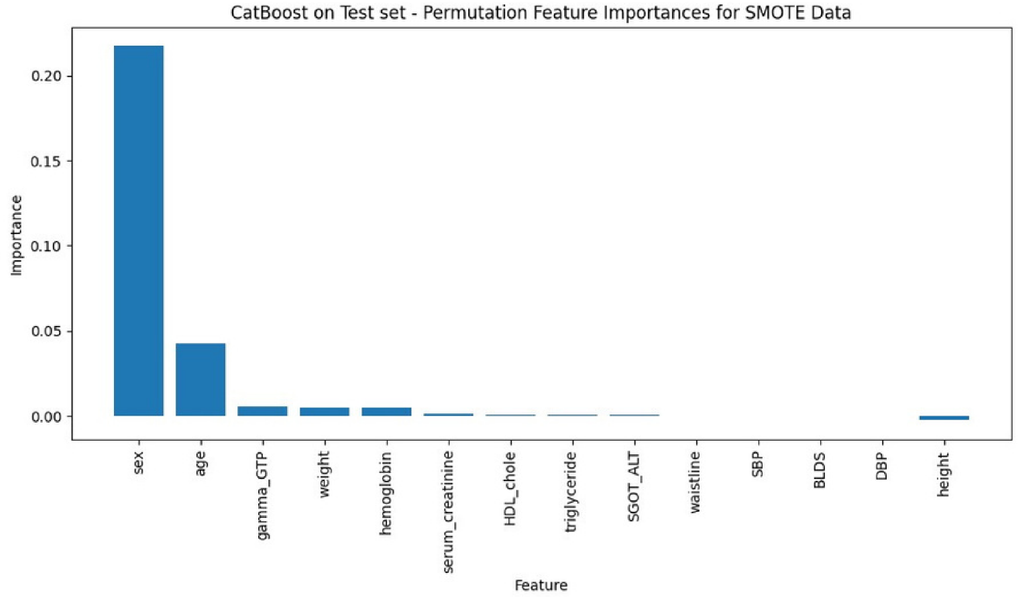
For SMOTE Dataset			
	Logistic Regression	Final CatBoost	Final LightGBM
Accuracy	0.5967	0.6771	0.6759
Precision	0.6579	0.7457	0.7455
Recall	0.5967	0.6771	0.6759
F1 Score	0.6174	0.6986	0.6977
ROC AUC	0.9435	0.8371	0.835

What are the top three features that most significantly influence smoking, drinking, and SMOTE datasets, according to the final CatBoost model, and how do they differ?



- **Smoking Data:**
  - **Sex: 9.565% - Primary determinant**
  - **Age: 3.964% - Indicates correlation with different age groups**
  - **Gamma-GTP: 3.088% - Reflects liver enzyme levels' significance**
- **Drinking Data:**
  - **Gamma-GTP: 8.596% - Liver function as a crucial indicator**
  - **Age: 5.820% - Age-related drinking patterns**
  - **Sex: 5.285% - Secondary but significant factor**
- **SMOTE Data:**
  - **Sex: 21.718% - Amplified impact in balanced dataset**
  - **Age: 4.263% - Consistent factor across datasets**
  - **Gamma-GTP: 0.550% - Reduced influence in balanced data**





How does the SMOTE technique work, and was it efficient in handling the imbalanced target feature in the smoking data?

Understanding SMOTE:

- Creates synthetic instances for minority classes.
- Works by interpolating between minority class samples and their nearest neighbors.
- Aims to achieve a more balanced dataset.

SMOTE's Impact on Smoking Data:

- **Balancing Effect:** This led to a more nuanced understanding of the data, as seen in the increased importance of the sex feature.
- **Model Accuracy:** Post-SMOTE accuracy was 0.6771, compared to pre-SMOTE accuracy of 0.7005.
- **Interpretation:** While SMOTE improved data balance, it did not significantly enhance the model's predictive accuracy for complex behaviors like smoking.

What are the differences in predictive performance between the smoking and SMOTE-augmented datasets?

----- Final Catboost of Smoking data -----
Accuracy: 0.7005, Precision: 0.7060, Recall: 0.7005, F1 Score: 0.7025, ROC AUC: 0.8474

----- Final Catboost of SMOTE data -----
Accuracy: 0.6771, Precision: 0.7457, Recall: 0.6771, F1 Score: 0.6986, ROC AUC: 0.8371

Predictive Performance Comparison:

- **Accuracy:** Reduced from 0.7005 (smoking) to 0.6771 (SMOTE).
- **Precision:** Increased, indicating improved true positive identification.
- **Recall & F1 Score:** Lowered, highlighting a trade-off in model balance.
- **ROC AUC:** Minor decrease, signifying reduced overall effectiveness.

Feature Importance Changes:

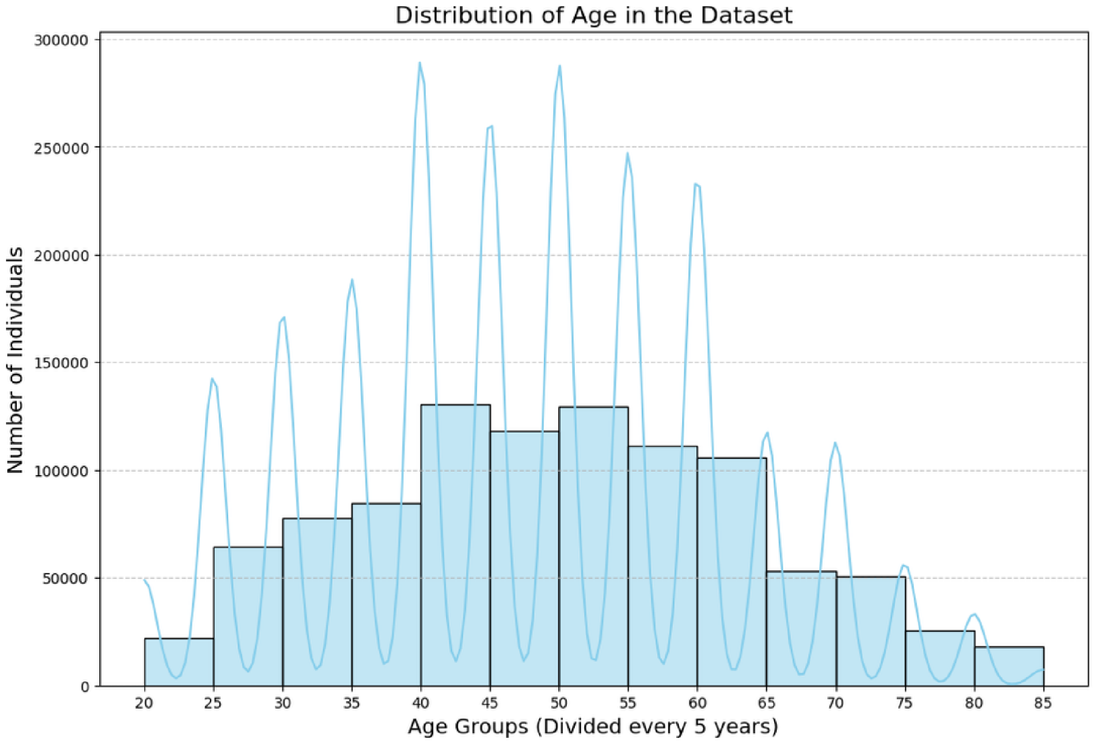
- **'Sex':** More pronounced in SMOTE, possibly due to synthetic overemphasis.
- **Class Imbalance Impact:** SMOTE may amplify features like 'sex'.

Practical Implications:

- SMOTE improves balance but may affect feature sensitivity.
- Essential to consider balance vs. accuracy in model application.

Limitations and future development

1. **External Validity:**
   - **The model based on NHIS Korea data may not apply to different demographics.**
   - **Future work: Validate with diverse global datasets for broader applicability.**
2. **Behavioral and Temporal Contexts:**
   - **Current model lacks psychosocial and time-series data.**
   - **Suggestion: Integrate behavioral variables and temporal data for richer insights.**
3. **Mid-Age Concentration in Data:**
   - **Data skews towards mid-age group, limiting age group representativeness.**
   - **Future data collection should aim for balanced age distribution.**
4. **Use of SMOTE for Class Imbalance:**
   - **SMOTE may introduce bias by artificially increasing minority class.**
   - **Alternatives: Targeted data collection or cost-sensitive learning for natural class balance.**



Thank you