BT5153 Group Project

GROUP 3

Tianyuan Jiang	A0280413N
Zihan Hou	A0280422N

Problem Statement:

As one of the leading causes of death, smoking has been an obstinate problem with ongoing conflict debates. The smoking exposes human body to more than 7000 chemicals and approximately at least 70 of them are categorized as toxic and potentially indument to cancer. According to some reports, smoking might tangle with many health risks, including lung cancer, heart disease, stroke and many chronic diseases. The global health burden of smoking is staggering, not only in terms of mortality but also in terms of medical costs and lost productivity associated with smoking-related diseases. Governments, communities and individuals should pay special attention to the dangers and addictive nature of smoking. Given the profound public health impact of smoking, there is an urgent need for interventions to effectively identify and support at-risk populations. More systematically and effectively reduce the prevalence of smoking and its associated health consequences.

Therefore, the primary goal of our exploration is to find out what health measurements or features might be most likely correlated to smoking status. By achieving this goal, different stakeholders will benefit from detecting those potential features. For example, if certain patients have both smoking history and diseases or abnormal health measurements that correlate to smoking status, the doctor could adjust their treatment plan accordingly by highly prioritising the importance of quitting smoke. The medical research agencies then could learn more about what factors might be highly correlated to smoking and do more precise education on smoking cohorts about the potential drawbacks of smoke. Overall, this experiment could help us understand more about the relationships between health conditions with smoking status and more importantly, the knowledge of what major features smoking cohorts tend to have will assist medical practitioners or individual health to a large extent.

Dataset:

the dataset titled “National Health Insurance Corporation_Health checkup information” encompasses a detailed compilation of health-related data for approximately 1 million subscribers of the National Health Insurance who underwent health checkups within the specific year. Managed by the Big Data Strategy Headquarters, this dataset includes a broad spectrum of information, categorizing basic demographic details such as province code, gender, and age range, alongside extensive health metrics covering height, weight, blood pressure, blood sugar, total cholesterol, and hemoglobin levels. The dataset represents a diverse demographic, including employed subscribers, dependents over the age of 40, and local subscribers who are either heads of households or above 40 years old, thus providing a comprehensive view of public health across various age groups and socioeconomic statuses. Updated annually and presented in CSV format for ease of use, the dataset was last corrected on April 23, 2024, to amend inaccuracies in age information, ensuring the reliability of subsequent data analyses. It is freely available for

download with no restrictions on use, encouraging broad academic and clinical application. Researchers are advised to pay attention to potential missing values and are recommended to refer to the National Health Information Data Health Checkup Information User Manual for a thorough understanding of the coding and data structure.

Data Preprocessing:

The dataset underwent rigorous preprocessing to ensure a focus on complete and relevant entries for our analysis.

The first step we take is to look at columns with null columns which we find three columns('Whether or not the defect is healed',' Tooth wear rate', 'Third molars (wisdom teeth) or more') with 0 non-null values, therefore, we directly drop those three columns. However, there are still some columns with a very large proportion of null values, which we selectively deal with those null values. The procedure we followed is that we either used a heatmap or histogram to explore the intro or inter-relationship between those features that have a large portion of null values, finally we decided to drop those features ('total cholesterol', 'triglycerides',' HDL cholesterol',' LDL cholesterol') that might have not any significant influence on our target variable.

The next step we take is to enrich our current features, which we did several transformations based on our features since they are all medical examination results, therefore, based on some WHO standards, we could divide them into different levels.

Selected Health Metrics and Their Categorizations:

- **Body Mass Index (BMI):** Calculated as weight in kilograms divided by the square of height in meters (kg/m^2). Categorizations follow WHO guidelines: underweight (BMI < 18.5), normal weight (BMI 18.5–24.9), overweight (BMI 25–29.9), and obese (BMI ≥ 30). (In the dataset: 0: underweight, 1: normal weight, 2: overweight, 3: obesity)
- **Waist Circumference:** According to WHO recommendations, waist measurements indicating increased metabolic risk are categorized as follows: Men > 94 cm and Women > 80 cm signify increased risk; Men > 102 cm and Women > 88 cm signify substantially increased risk. (In the dataset, 0: normal, 1: increased risk, 2: substantially increased risk)
- **Vision:** Although the WHO does not specify exact thresholds for vision impairment, it advocates for regular vision screening and appropriate corrective measures, which were documented for both eyes.

- **Blood Pressure:** Categorized into normal, elevated, and stages of hypertension according to the WHO's classifications: normal (Systolic < 120 mmHg and Diastolic < 80 mmHg), elevated (Systolic 120–129 mmHg and Diastolic < 80 mmHg), hypertension stage 1 (Systolic 130–139 mmHg or Diastolic 80–89 mmHg), and hypertension stage 2 (Systolic ≥ 140 mmHg or Diastolic ≥ 90 mmHg). (In the dataset, 0: normal, 1: Elevated, 2: Hypertension stage 1, 3: hypertension stage 2, -1: uncategorized)
- **Blood Sugar Levels:** Following American Diabetes Association (ADA) guidelines, which are often used in conjunction with WHO recommendations, fasting blood sugar levels are classified as normal (< 100 mg/dL), prediabetes (100–125 mg/dL), and diabetes (≥ 126 mg/dL). (In the dataset, 0: normal, 1: Prediabetes, 2: Diabetes 1, -1: uncategorized)
- **Hemoglobin Levels:** Typical value ranges were applied based on normative data, which can vary by age and sex, indicative of possible anemia or other blood disorders. (In the dataset, 0: normal, 1:Abnormal)
- **Serum Creatinine:** Used to assess kidney function with normal ranges typically varying based on lab and demographic factors but generally aligned with medical standards. (In the dataset, 0: normal, 1:Abnormal)
- **Liver Enzymes (Serum AST and ALT):** Normal values for AST (10 to 40 units/L) and ALT (7 to 56 units/L) are used, which fall within standard medical guidelines to assess liver health. (In the dataset, 0: normal, 1:Abnormal)
- **Gamma GT (Gamma-glutamyl transferase):** Normal values set at 0 to 65 U/L for men and 0 to 45 U/L for women are used to assess liver health and detect excessive alcohol use. (In the dataset, 0: normal, 1:Abnormal)

This detailed preprocessing approach ensures that each health indicator is evaluated within a framework established by WHO and other authoritative health guidelines. By adhering to these standards, our analysis is poised to provide robust insights into how smoking affects various health dimension.

Exploratory Data Analysis:

In the exploratory data analysis, we did different explorations on both numerical variables and categorical variables on our dataset.

Numerical variables (In total 15 numerical variables): ‘BMI’, 'Age code (5 years increments)',' Height (in 5cm units)',' Weight (in 5 kg units)',' Waist circumference', 'Vision (left)',' Vision

(right)','systolic blood pressure','diastolic blood pressure','Pre-meal blood sugar (fasting blood sugar)',' hemoglobin',' Serum creatinine',' Serum AST (AST)',' Serum GPT (ALT)',' Gamma GT'

For those numerical variables, the major exploration we did is in three aspects, the first is we explore their statistical features which include the mean, median, standard deviation, minimum, maximum and their first or third quantile value. Secondly, we explore the distribution of the feature through a histogram plot which we find most of the features roughly follow a normal distribution, except features like blood-sugar level or age code are a little right-skewed. And also we explored those features' relationship with our target variable, and we found that features like height, weight, BMI, diastolic blood pressure, hemoglobin are obviously positively correlated with smoking status and the age group has a negative relationship with smoking status.

Categorical variables (In total 16 categorical variables): All other variables

For categorical variables, our exploration is mainly conducted through histogram plots which could help us examine whether different categories potentially reflect different relationships between our features and target.

In this exploration, there are also many interesting facts are being revealed:

1. Male groups tend to have more smokers
2. People who drink alcohol are more likely to smoke as well
3. People who smoke tend to be more likely to have higher tartar level
4. When people have higher blood sugar levels, there are more smokers than non-smoke people
5. Some liver-related tests like Serum AST (AST), Serum ALT (ALT), and Gamma GT also indicate obvious differences between smokers and non-smokers, which smokers tend to have more abnormally.

Overall, in the above EDA process, we explore both statistical features and distributions, in addition, to their own inherited trend and their relationship with different features and many interesting facts to be founded

Model Training & Evaluation

In our study, we evaluated multiple machine learning models to predict health outcomes associated with smoking status, leveraging a preprocessed dataset. Each model was assessed based on its roc-auc score, which provides insights into the model's performance in terms of its ability to distinguish between classes, respectively.

The selected models are: Logistic regression, decision tree, random forest, XGBoost and LightGBM

Implementation&Evaluation:

Logistic regression: The training process for logistic regression will be very simple since logistic regression does not have many hyperparameters, therefore, what we did is just tune the class weight to a balanced weight. Logistic regression here serves as a purpose as a native model to compare with other models which helps us to learn both whether the dataset or features we did is effective and whether model difference really makes any difference.

Decision tree: The selection of a decision tree is due to its advantage of capturing more complex patterns rather than just linear relationships and the training process mostly spends the effort on hyperparameter tuning, since we have a comparatively large dataset, therefore, we decided not to use some automated hyperparameter tuning methods like Grid Search or Random Search and the hyperparameters we used are max_depth, min_samples_split and min_samples_leaf as well as the class weight we set to balance.

Random forest: The choice of random forest model then focuses on its ability to avoid overfitting and ability to uncover more patterns. Similarly to the decision tree, we also did manual hyperparameter tuning and besides those hyperparameters, we applied for the decision tree, we also added a number of estimators and maximum features.

XGBoost: The choice of XGBoost is due to its high efficiency when dealing with a large dataset and the major hyperparameter tuning is then we use its unique regularization proper and added both l1 and l2 regularization.

LightGBM: The choice of LightGBM is also largely due to its efficiency. But unlike XGBoost we did not add regularization since we did not find very severe overfitting issues.

The final model performance is listed below:

Model	Training roc-auc score	Test roc-auc score
Logistic Regression	0.729	0.728
Decision Tree	0.810	0.805
Random Forest	0.885	0.864
XGBoost	0.933	0.867
LightGBM	0.884	0.870

Neural Networks:

We also chose the neural networks as one of our candidate models after considering our dataset. Since we have more than 300K rows which is a large dataset, therefore, neural networks would be an ideal model to handle this large dataset, however, the neural networks might not bring proper interpretation in terms of our project aim which is to obtain more understanding the features of smoking cohorts.

Implementation & Result:

The implementation of our neural networks is basically the process of hyperparameter tuning, first of all, the first hyperparameter is the batch size of our data loader and after trial and error, the most ideal batch size is selected to be 512. Then there will be the design of our neural networks, since if we only use one hidden layer, it will basically have not much difference from a logistic regression, therefore, we decide to take the idea of building deep neural networks. Then we built three different hidden layer designs, the first design we built involved three hidden layers, with the first hidden layer containing 31 input neurons and 32 output neurons the second layer taking (32,64) design and the third one then taking (64,128) design and also three hidden layers followed by ReLU activation function which in this design we expect to expand more info through the hidden layer, then after three hidden layers we finally use the sigmoid function to get the final prediction probability. For the second design, we then want to restrict the feature size but still take a similar three-layer design, therefore we design all three hidden layers to be the size input and output size(31,31). The last design we take is to decrease the hidden layer size to two and this last design takes the same hidden layer design as the second design which has two

hidden layers of similar size (31,31). Overall, we experimented with all three designs we first found the best learning rate for each layer design and implemented the best learning rate to the validation process, meanwhile, we also adjusted other hyperparameters like epoch amount, number of iterations etc. However, all three deep neural network designs are not very useful for finishing our tasks compared to other models we built, which the best-performed model takes the first three-layer design, with the best epoch roc-auc score performance is 0.754, but on average, the performance measured by roc-auc score is around 0.5 even there is no obvious overfitting.

Final Model Recommendation:

The final model we picked is the Lightgbm model since it has the highest roc-auc score among all the candidate models and its final performance on the test set is 0.870 as well

Model Interpretation:

Random Forest

Global: We use both random forest inherited feature importance and permutational feature importance which both indicate a similar result in which the top features are height, hemoglobin, Gamma GT, Drinking status, Serum Creatinine etc. Which generally indicates that people's smoking status is largely related to their liver-related medical condition.

Local: We also use LIME to obtain some local explanation on some randomly chosen instances which are 0, 1687 and 8970, the first instance actually reflects similar features like height, drinking status hemoglobin, Gamma GT and also who is categorized as smoke. For instance 2 and 3, we actually do not find those features play a significant role when determining their class and they then being categorized into class 0 which is non-smoke.

XGBoost

Global: For the global explanation, we also use permutation importance as interpretation for our XGBoost model, in which the permutation importance by XGBoost also reflected similar feature importance as random forest, although, the XGBoost model ranks their a little different, which the model ranks hemoglobin as the most important feature and waist circumference is being weighted more in the XGBoost model.

Local: The LIME explanation for the above three selected instances in the XGBoost model also reflected a similar property which most of the instances, the model distinguishes them based on those top selected features.

LightGBM

Both the global and local explanations for LightGBM and XGBoost are almost similar, since the essence they both share a similar model design, therefore, we did not further discuss them here.

Conclusion:

Overall, this experiment is very successful, and we built a relatively strong classification model with a 0.870 AUC score. More importantly, the experiment actually provides two significant insights which are that height might be one of the most important features in distinguishing between smoking cohorts and non-smoking cohorts and the second important observation is that liver-related medical conditions should be specifically noticed when distinguishing smoking cohorts or non-smoking cohorts and also there might indicate very strong correlation or even causation between people who have abnormal liver medical conditions with smoke status. Several stakeholders might benefit from this experiment, the government or medical agencies could improve their intervention in smoke cohorts on both propaganda aspects and medical treatment aspects. The government could improve the authenticity of its anti-smoke intervention when doing anti-smoke campaigns by indicating more specifically the dangers that smoke cohorts might face which in our experiment is the potential danger of liver abnormal medical conditions and the medical treatment aspects is that the model we built could help medical agency to judge whether the patients are smoke or not even before patients verify that information more effectively, more importantly, the model's local interpretability could help doctors to pre-select which abnormal medical conditions that the patients most correlated to its smoke status and give more specified medical advice for those cohorts. In total, this experiment enables us to obtain more detailed insights into what factors smoking status might correlate with and also it provides an alternative way to help medical practitioners figure out potential major factors that correlate to a cohort's behaviour like drinking or smoking by employing machine learning methods.

Reference

국민건강보험공단_건강검진정보_20221231. (2024b, April 23). 공공데이터포털. https://www.data.go.kr/data/15007122/fileData.do (Dataset)

Indicators index. (n.d.). https://www.who.int/data/gho/data/indicators/indicators-index

GHO. (n.d.). https://www.who.int/data/gho

Health effects of smoking and tobacco use. (2022, March 30). Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm#:~:text=Smoking%20causes%20cancer%2C%20heart%20disease,immune%20system%2C%20including%20rheumatoid%20arthritis

Professional, C. C. M. (n.d.). *Smoking*. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/17488-smoking

Health effects of cigarette smoking. (2022, August 19). Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/

World Health Organization: WHO. (2023, July 31). *Tobacco*. https://www.who.int/news-room/fact-sheets/detail/tobacco

Github:

https://github.com/KwongTy/BT5153-Code-Dataset.git

Contributions of each member:

Tianyuan Jiang	DP+MTE+Report+pre
Zihan Hou	EDA+MTE+Report+pre