Francis Yang A15839617
Milka Waniak A16428340
Maya Beeri-Feldman A16650164

Introduction

Knowledge of cigarette smoking has come a long way from its earliest emergence. Since their first introduction in the U.S. in the 1970s cigarettes “have rapidly increased their market share” and subsequently exploded in their usage around the world (Fielding, 1985).

Even though their popularity was later pushed back with awareness of its harmful effects, it is still the cause of many diseases that can be made preventable. Above that, research shows that 8 million people die prematurely because of smoking, so that means that smoking is responsible for one in seven deaths” (Roser, 2021). Smoking not only negatively impacts human health, but smoking is also the reason for economic damage to over $ 500 billion USD every year (Ekpy & Brown, 2015).

Thus, though smoking awareness has come a long way, there is still much to be done to lower numbers, especially as it was through knowledge of smoking’s negative effects that numbers began to lower in the first place. Specifically, our group found a data set that worked closely with measures of blood work as well as a few other factors.

Blood pressure, uniquely, has “long been known” to “increase during smoking” as these effects are due to “an increase in cardiac output and total peripheral vascular resistance” (Omvik, 1996). However, it is also worth noting that “while smoking acutely increases blood pressure, a slightly lower blood pressure level has been found among smokers than non-smokers in larger epidemiological studies” (Omvik, 1996). So, we are hoping that this data set can give us further insight into this relationship as well as point us to more specific factors that we can link to smoking.

Technology-based interventions for smoking include practices like advertising media, telecommunication through different media like TV, radio, or the internet, and these effects are shown to effectively decrease smoking behavior (Ekpy & Brown, 2015). Despite its "long decline" in smoking, still, around 20% of adults smoke, and it takes a very long time to have any reduction in smoking rate" (Roser, 2021). By creating a model that effectively predicts if someone smokes or not based on health factors, one can provide further evidence that can be used in technology-based intervention and therefore reduce smoking behavior.

Data Set

The data set that we are using for our study can be found here: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking?select=smoking.csv (From here originally https://www.data.go.kr/data/15007122/fileData.do#/tab-layer-file)

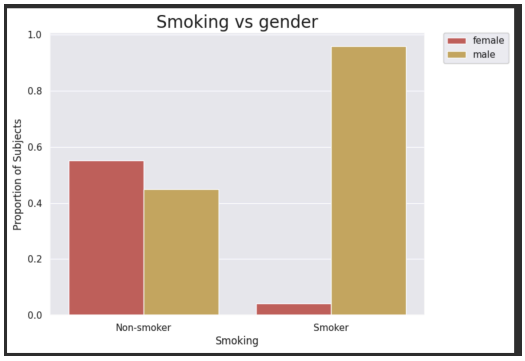
The dataset contains 25 total predictors for determining the presence of smoking in the body. At a cursory glance, a number of these predictors seem less like real predictors and more like just standard classification for the study participants such as gender, age, height, weight, eyesight, and hearing. The predictors that actually have to do with the body's biosignals such as systolic and relaxation blood pressure, cholesterol levels, urinary protein levels, etc., will likely be much more relevant as far as the final prediction of whether or not the person is a smoker. Because there was not much additional information pertaining to the dataset, it is impossible to determine whether or not there are any unseen factors that are affecting the data set quality. In that same vein, the website that the data was sourced from is completely Korean so no one in our group knows how exactly the data was sourced. However, the website itself seems fairly reputable with Google Chrome's auto-translate feature showing multiple contact methods as well as a list of sources, so at the very least we are confident in the credibility of all of the data.

EDA Results

To perform our EDA and determine which predictors would be most relevant, we divided the dataset into smokers and non-smokers in order to compare the distributions for every predictor and see which ones actually had relevant differences. We decided not to analyze eyesight, oral examination status, and urine protein, for varying reasons. First, it was impossible for us to determine what the eyesight values referred to so we had to remove that value due to us being unsure of what exactly the indicator meant. Second, oral examination status was the same for all participants so we removed that value (they all received examination). For urine protein, we were not able to understand what the values recorded corresponded to as the values did not seem to line up with typical measurements of urine protein so just to be safe we scrapped this value as well.

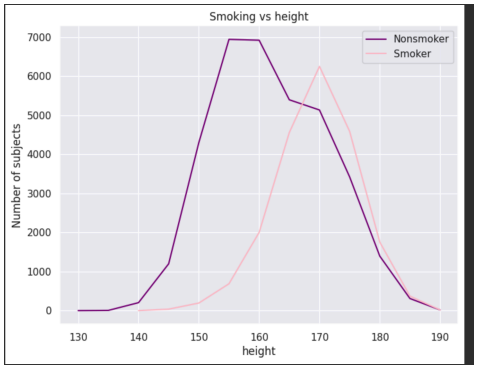
Furthermore, one of the variables was called ‘relaxation’ but from the values recorded and how close it was systolic on the CSV file, and how other users on Kaggle interpreted it, we believed this value to be diastolic and treated it as such. Similarly, one variable was called GTP whenever I searched it up would yield results for GGT. Since the inclusion of gamma and the values it had made sense for GGT as well, we decided to rename this variable.

Gender(binary: males and females): Gender appears to be very <u>important</u>, with one of the largest differences in distribution between males and females. The distribution of non-smokers was around a 60-40 split favoring females, but in the case of smokers, the distribution was massively skewed with males taking up upwards of 95% of the data.

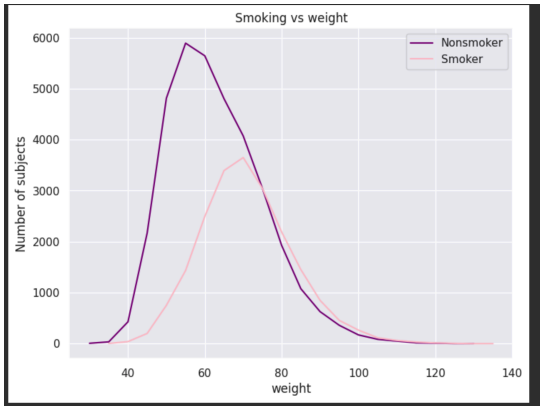


Age(years): Though we would normally expect age to play a role in predicting someone's smoking status, the age predictor does not appear to be relevant, outside of a difference in intercept the actual shapes of the distributions themselves are nearly identical.

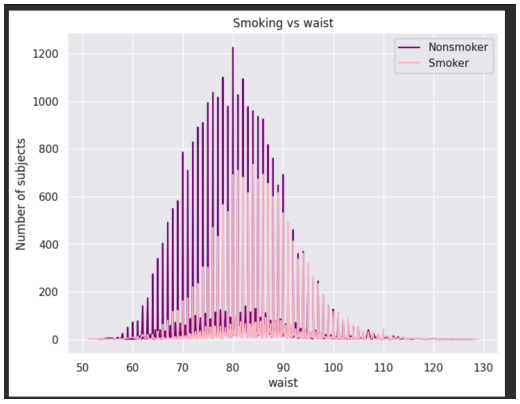
Height(cm): Although the shapes of the distributions are similar, where the peaks occur are slightly different. This is most likely a relevant factor in predicting smoking. It is worth noting that the difference though might just be a confounding variable of gender differences.



Weight(kg): Same as with height, the shapes of the distributions are the same but with different peaks, suggesting that this feature might have some effect in predicting whether someone smokes. Though this again could also be due to males naturally weighing more than females.



Waist(cm): The shapes of the distributions appear differently, suggesting that waist size does appear to be relevant. This again is probably a reflection of gender differences in smoking.



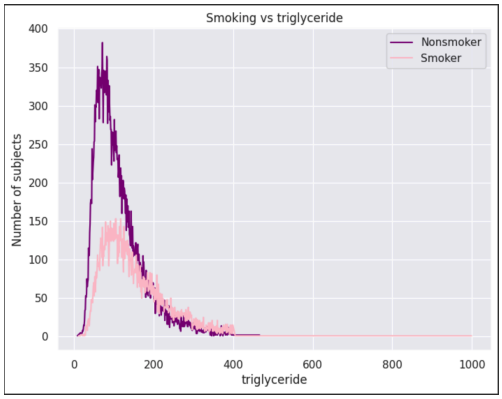
Hearing(1-normal, 2 - abnormal): Hearing was not relevant for both the left and right sides, with both smokers and smokers having the same distribution where less than 5 percent of participants had abnormal hearing.

Blood Pressure(mmHg): The same as with hearing, both systolic and diastolic blood pressure did not appear to be relevant since the distributions for smokers and nonsmokers were nearly identical.

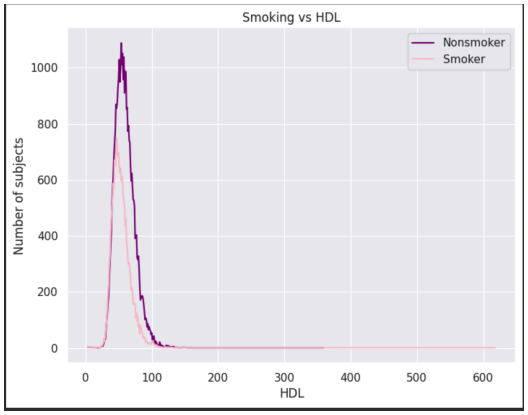
Fasting Blood Sugar(mg/dL): FBS is a measure of the amount of sugar in the blood, and is often used to measure diabetes. FBS was not relevant with the distributions being identical in shape.

Cholesterol(mg/dL): Cholesterol levels were the same as FBS where it was not relevant with the distributions being identical in shape.

Triglyceride(mg/dL): Triglycerides are a type of fat that is found within the bloodstream, known to be the most common type of fat. This variable might be important as the distribution for smokers seems to have a bigger variance and the center of the distribution is a little bit to the right.

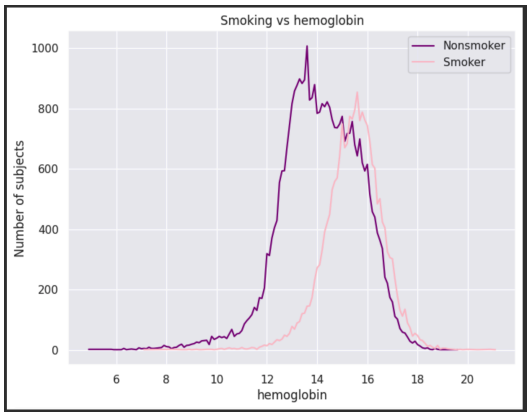


HDL(mg/dL): HDL refers to high-density lipoproteins, which bring cholesterol to the liver in order to remove it. The shapes of the distributions of smokers and nonsmokers are slightly different, with smokers having the center of distribution slightly to the left (lower mean HDL). That means that this variable might be important.



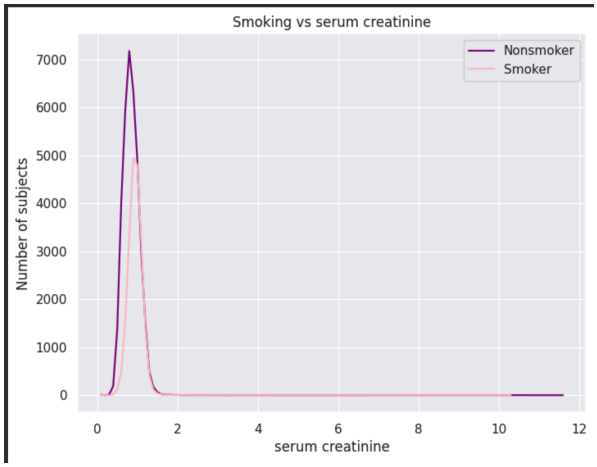
LDL(mg/dL): Low-density proteins that do not carry cholesterol like HDL and can actually lead to a clog of cholesterol at higher levels. This variable seems not to be important, as the distribution seems to have the same shape for smokers and nonsmokers. There is just a difference in the peaks between smokers and nonsmokers that is probably caused by the different number of subjects that smoke and don't smoke.

Hemoglobin(g/dl): For the typical human, a high hemoglobin level is necessary to ensure that there is sufficient tissue oxygenation throughout the body. This variable seems to be important as the shapes of the distributions are different and they also have different centers - smokers have higher hemoglobin scores.



Serum creatinine(mg/dL): An increased creatinine level can be a potential indicator of poor kidney function. The shapes of the distributions of smokers and nonsmokers are slightly different, with smokers having the center of distribution slightly to the right (higher mean for

serum creatinine). That means that this variable might be important.

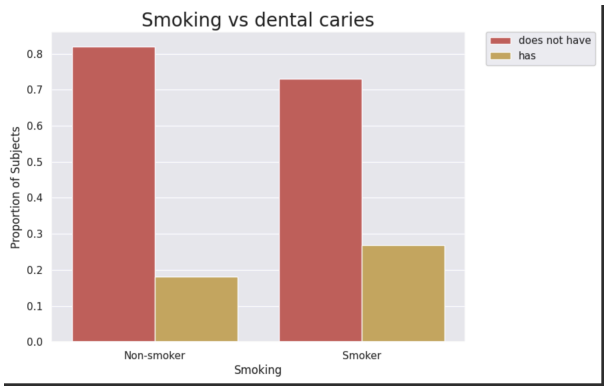


AST: glutamic oxaloacetic transaminase type(U/L): An enzyme found in high levels in the liver, heart, and muscles. This variable seems not to be important, as the distribution seems to have the same shape for smokers and nonsmokers. There is just a difference in the peaks between smokers and nonsmokers that is probably caused by the different number of subjects that smoke and don't smoke.

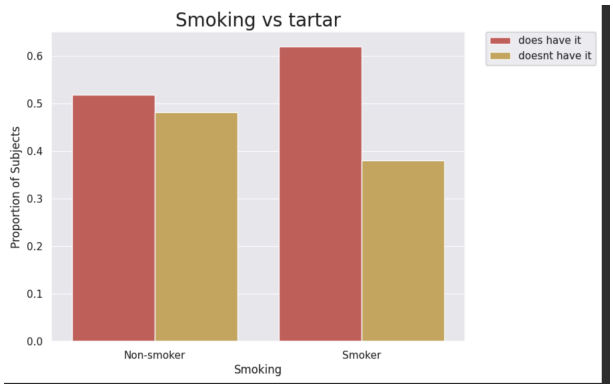
ALT: glutamic oxaloacetic transaminase type(U/L): Alanine transaminase is another enzyme found within the bloodstream. This variable seems not to be important, as the distribution seems to have the same shape for smokers and nonsmokers. There is just a difference in the peaks between smokers and nonsmokers that is probably caused by the different number of subjects that smoke and don't smoke.

γ-GGT: gamma-glutamyltransferase (U/L): An enzyme found within the liver, kidney, pancreas, heart, and brain. This variable seems not to be important, as the distribution seems to have the same shape for smokers and nonsmokers. There is just a difference in the peaks between smokers and nonsmokers that is probably caused by the different number of subjects that smoke and don't smoke.

Dental caries(binary): Binary variable that determines whether the person has any dental cavities/diseases. This variable seems to be important as the proportion of people who have dental caries is different for smokers and nonsmokers with smokers having a higher proportion of dental caries.



Tartar: tartar status(binary): Tartar refers to deposits found on the teeth that generally indicate some amount of decay. This variable seems to be very important as the proportion of those who have tartar is higher for smokers than nonsmokers.



Hypothesis

We predict that the variables, tartar, dental caries, serum creatinine, hemoglobin, HDL, triglyceride, height, waist, and gender, will have the most impact in predicting whether a person smokes or not.

Methods

Model Approach

For our data analysis, the method that we chose to go with was a KNN classifier, with different models being based on different values of k. On top of KNN naturally being simple to use, having high accuracy, and not being computationally taxing, KNN does not necessarily require a relationship between the predictors and the classification like something like linear regression, something that is useful in this case since not every one of our predictors is going to be relevant. We are specifically focusing on the prediction of smoking status from these predictions.

Something important to note in our data analysis is the curse of dimensionality. The curse of dimensionality, or Hughes Phenomenon, dictates that models become increasingly more difficult to make meaningful as the number of predictors increases. This problem becomes even further exacerbated when some of those predictors are not meaningful/have no relevance. Since our model contains not only a high amount of total predictors but also a high amount of non-relevant predictors, the curse of dimensionality is a very real problem that we could run into while trying to make our model.

Cross Validation

Additionally, for our cross-validation method, we chose a k-fold cross-validation method. Essentially, since we used a k value of 5, our data was split into 5 different sets. Each of the sets was then used as testing data for the rest of the observations not included in the test set (thus the training set). This was completed five times, once for each test set. Since we have 9 values of K

(where K is the number of neighbors used in our k-nn model), there will be a total of 45 total values of accuracy.

Furthermore, we will also be using a forward stepwise feature selection algorithm that will one at a time add features to see if they aid in better performance and accuracy and thus will include them accordingly. Because we have a higher amount of features, a lot of which we believe won't be relevant in prediction, we will be using this tool in order to help us pick the best features to aid in our predictions. Specifically, we will use this stepwise feature over general subset feature selection that would simply be too costly.

Results

Model Selection

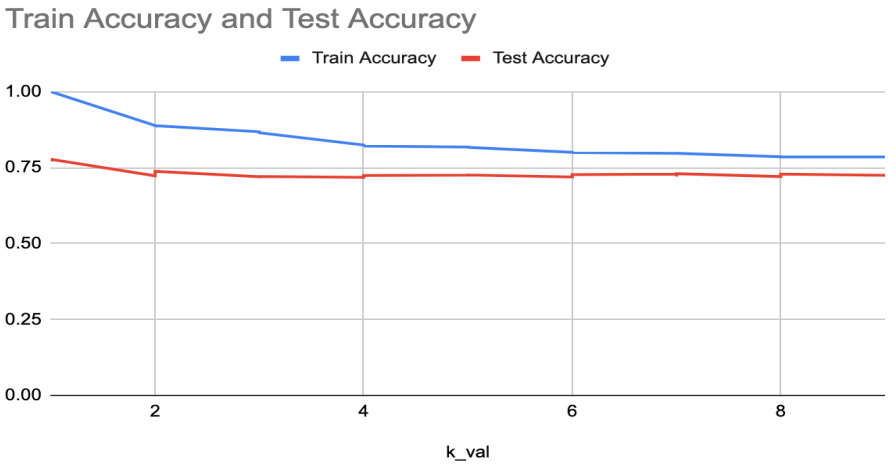
Though our value of K = 1 had the highest training and testing value this is not entirely surprising as a lower value of K means a more complex decision boundary that will be more susceptible to noise and possible outliers and thus have a higher chance of overfitting to our data. This also means a higher variance as each data point is even more weight and leads to the model being less robust. However, our higher values of K like 9, while being less exposed to possible overfit, will have a higher bias present as the decision boundary will be more prone to underfitting and thus there being a larger discrepancy between the model and true relationship.

Thus, to balance both of these extremes, we believe that choosing a K value of 5 will be our best model as it has the next highest testing score after K = 1 and 2 and is still not as simple of a decision boundary as the higher values of K.

Model Estimation

After running our K-NN model with different values of K and k-fold as cross-validation, we found that the K value of 1 had the highest training and testing accuracies of 1 and .78 respectively. Additionally, as our K's increased the value of training and testing accuracy seemed to steadily decrease and fluctuate around values of 70-73 for testing and .88-.78.

Using a value of K = 5, our final parameter, as chosen earlier, when tested on the whole data we got an accuracy of 81 overall.



Discussion

Our hypothesis was partially supported. We correctly predicted the importance of tartar, dental caries, serum creatinine, hemoglobin, HDL, waist, height, and triglyceride in predicting if someone is a smoker or not. However, gender turned out not to be important in the model. It is possible that this is a limitation for using stepwise selection instead of the best subset selection.

Above that, there were a few other variables that turned out to be predictive of someone's smoking habits, but in our EDA we classified them as not important: age, ALT, GGT.

For future work, it might be helpful to look into these variables to determine their influence on smoking to see if there was some issue with the forward stepwise or if this result is

replicable with new participants. Additionally, if we had had more time we could have tried other forms of subset selection like backward stepwise or trying out a more extensive parameter search with different values of k or even looking at other models other than KNN. Furthermore, it might also be helpful to look at other health factors that smoking impacts like lung health.

Another limitation is that our data comes from sources that we are not familiar with such that we are not sure if this data is accurate or representative of any population. Additionally, we cannot be sure how valid or unbiased any of the data was since we could not check to see if it was collected by a credible source due to our own lack of Korean. However, even if the data is representative of the South Korean population it might not generalize well to other populations. Thus in future studies, it would be crucial to include other populations and groups of people to make sure there is a representative sample of a broader population.

References

Ekpu, V. U., & Brown, A. K. (2015). The economic impact of smoking and of reducing smoking prevalence: review of evidence. *Tobacco use insights*, *8*, TUI-S15628.

Fielding, J. E. (1985). Smoking: health effects and control. *New England journal of medicine*, *313*(9), 555-561.

Omvik, P. (1996). How smoking affects blood pressure. *Blood pressure*, *5*(2), 71-77.

Roser, M. (2021). Smoking: How Large of a Global Problem Is It. *And How Can We Make Progress against It*.