CS 677 Final Project Analysis

SayoPillow is an edge device that analyzes the physiological changes that occur during human sleep and stores stress data. In layman terms, Sayopillow is a special pillow that measures sleeping features to detect and classify human stress levels. For my final project, I am interested in building a mathematical model that can predict snoring rate based on other sleeping features and simplify the provided stress levels into smaller groups or clusters.

Before I start using any machine learning techniques, I will explain in detail about the dataset I will be using. The dataset provided from Kaggle has six-hundred thirty rows of data with nine columns of features. The nine columns of features consist of snoring rate of the user, respiration rate measured in number of breaths per minute (normal respirate rate for an adult at rest is 12 to 20 breaths per minute), body temperature in Fahrenheit, limb movement rate measure in seconds, blood oxygen levels measured in percentages (90-100% is the normal range), eye movement, number of hours of sleep, heart rate measured in BPM and stress levels (0-4). There are no null values or missing values hence no need to do any data cleaning. In the histogram of the stress levels, the data is shown to be uniformly distributed meaning we will have no problems with having uneven proportions of each stress level to work with.

The first question is can I predict the snoring rate of a human given the other sleeping features I mentioned previously. To answer this question, I will be building a multi-linear regression model using snoring rate as the dependent variable. I did not include stress level as a feature since it is not considered an independent variable in the context of this problem. Plotting a correlation matrix on a heatmap, I noticed that a lot of the features are highly correlated. As shown later, the absolute correlation range is from 0.86 to 1. To build the multi-linear regression model, I split the data into a 60/40 train and test set and fit the training data. To determine how suitable the model is, I calculated the accuracy by comparing the original snoring rate vs the predicted snoring rate and the SSE. The accuracy was 96.86% which matches the SayoPillow documentation accuracy meaning this multi-linear regression model is a superb model. As for the SSE using all features (2841.42), we cannot determine how well the model is unless we do some feature removal.

Removing the most highly correlated features (correlation coefficient of 1) one-by-one or in pairs, I noticed that the accuracy increased at most 0.02% (from 96.86% to 96.88) (ex: removing respiration rate and body temperature) The SSE of removing respiration rate and body temperature did however decrease the SSE by about 15.6 (from 2841.42 to 2825.82). These slight differences are not worth removing any features hence I will use the original multi-linear regression model using all sleeping features to predict snoring rate y = (19183726823373.3)\*(respiration\_rate) + (1.55078125)\*(body\_temperature) + (1.9580078125)\*(limb\_movement) + (-1.067138671875)\*(blood\_oxygen) + (0.07177734375)\*(eye\_movement) + (-1.591552734375)\*(sleeping\_hours) + (-7673490729348.662)\*(heart\_rate) + 76734907293452.08).

The second question is given there are 5 stress levels, can I simply these stress levels by clustering. To reduce the amount of stress levels/ labels, we must figure out how to group similar labels by using k-means clustering. Since k-means uses Euclidean distance, I will also be testing whether I should scale the features or not. Both scaled and unscaled features’ elbow plots determined k = 3 as the optimal k. For these new clusters, I will label them as “Not Stressed”, “Stressed” and “Over Stressed”. In terms of which stress levels belong to which of the 3 clusters, scaled and unscaled k-means plot are very different. For scaled k-means plot of snoring rate vs blood oxygen, I notice that cluster 2 has the smallest count (126) compared to the other clusters that have double the amount of data points. For unscaled k-means plot of snoring rate vs blood oxygen, data points from stress level 3 belonged in both cluster 0 and cluster 2. I considered stress level 3 as cluster 0 since cluster 0’s stress level 3 counts are greater than cluster 2.

Before deciding on which version of k-means I want to use (scaled or unscaled), I need to look at the range of the features after being aggregated into the new clusters/ labels. The range of “Over Stressed” features is a lot smaller in the scaled features versus the unscaled features. Based on the k-means plot and ranges of the features, I should be using the scaled data, but I personally believe that using the unscaled features clustering is more reasonable. In the real world, I would rather have myself classified early on as “Over Stressed” rather than just “Stressed” For example, if my snoring rate is 85, I would be considered as “Stressed” using the scaled data, but in the unscaled data I would be considered “Over Stressed” causing the treatment to reduce my stress to differ depending on which data I used for k-means.

SayoPillow can be used to detect stress levels based on sleeping features, but it can also be used to predict snoring rate and diagnose stress levels in simplified categories. My project was able to create a multi-linear regression model that predicts snoring rate with 96.86% accuracy when considering all sleeping features. I was also able to simplify the 5 stress levels into 3 categorial labels consisting of “Not Stressed”, “Stressed” and “Over Stressed” using k-means clustering on the unscaled sleeping features.