Final Project Plan (COMP 4448)

**Megan Hoeksema**

Complete this document as part of your project plan about your final project. Your answers need to be short and brief. About one to three sentences for each question is fine. Write your answers under each question below. Make sure to maintain the question numbering.

Maximum Grade Points = 50 points

1. What is purpose of your project? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way).

**The purpose of this project is to compare different classification algorithms, KNN, decision trees and random forests and determine which algorithm is better at classifying whether or not an indoor worker is a smoker or not based on work ban, gender, age, education and race.**

1. What is the significance of your project? That is, why is this project important? What potential positive impact does your project have?

**The significance of this project is to evaluate each of the models and tune the parameters to produce the best model for each of the classification algorithms then determine the most appropriate classification model for determine smoking status of indoor workers.**

1. What is the research question? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way. A research question is the purpose of your project flipped into a question).

**Is KNN, Decision Trees or Random Forest a better model for predicting the smoking status of indoor workers based on race, gender, age, education and workplace smoking bans.**

1. What algorithms will you be comparing to answer your research question?

**KNN, Decision Trees, Random Forest**

1. What is your output variable?

**Smoking Status**

1. What are the input variables for your analysis? You should list at least five input variables in the dataset to be used for analysis/modeling.

**Ban, age, education, African American, Hispanic, and gender**

1. What kinds of data preprocessing are you anticipating to perform on the dataset before modeling?

**All columns except for age are categorical and will need to be replaced with numerical values. Potentially combine the African American and Hispanic columns together and assign 0 for African American and 1 for Hispanic, then relabel the column as race. Will need to probably scale the data using min-max scalar or standard scaler.**

1. What exploratory data analyses (on which variables) do you plan to do before modeling?

**Descriptive statistics will be performed on the age column.**

**Perform correlations between variables**

1. What visualizations (on which variables) do you plan to do before modeling?

**Visualize the number of indoor smokers vs non-smokers, can also show based on gender. Visualize the age range of the data set.**

1. What is the link to your data source? If your dataset is not from an internet data source, you will need to upload it to GitHub and share the link here. Avoid using common internet datasets such as Iris dataset, mushroom dataset, etc. Your dataset should be something interesting to you and uncommon). Feel free to use a dataset from this link: <https://vincentarelbundock.github.io/Rdatasets/datasets.html> or from some other source such as Kaggle competition but don’t present any copy and paste internet solutions for your final project. You will use the tools you have learned from this course to do your own project independently. Please, provide the link to your data source here:

**https://vincentarelbundock.github.io/Rdatasets/doc/AER/SmokeBan.html**