###**Project 2 Report**

####**Background and Results from Project 1**

####**The MONICA Project and Dataset**

#####The MONICA Project, or the *Multinational Monitoring of Trends and Determinants in Cardiovascular Disease*, began enrolling participants in the fall of 1984. The main objective of the study was to “measure the trends and determinants in cardiovascular mortality and coronary heart disease and cerebrovascular disease morbidity and to assess the extent to which these trends are related to changes in known risk factors, daily living habits, health care, or major socio-economic features measured at the same time in defined communities in different countries.”1 The study had two main null hypotheses, 1) “among the reporting units, there is no relationship between 10 year trends in serum cholesterol, blood pressure, and cigarette consumption and 10 year trends in coronary heart disease incidence rates,” and 2) among the reporting units, there is no relationship between 10 year trends in 28 day case fatality rates and 10 year trends in acute coronary care.” A total of 41 MONICA Collaborating Centers were established, which studied 118 reporting units, representing a total population aged 25-64 of approximately 15 million people. The researchers conducted annual surveys within each of the 118 reporting units to measure the trends and determinants of cardiovascular mortality.

####**Project 1 SMART Research Question and Study Aims**

#####What are the specific demographic and physiological indicators that are associated with the outcome of mortality from cardiovascular disease and what is the strength of those associations among a globally representative sample taken between 1985 and 1993?

#####**Study Aim 1**: To perform a Chi-square analysis to assess the independence of each demographic and physiological risk factor with mortality from cardiovascular disease.

#####**Hypothesis**: We expect to see mortality from cardiovascular disease to be significantly dependant for all physiological indicators and demographic indicators.

#####**Study Aim 2**: To perform a Chi-square analysis to assess the independence of each physiological risk factor with mortality from cardiovascular disease after stratifying by the sex of the participant.

#####**Hypothesis**: We expect to see the outcome of mortality from cardiovascular disease to be significantly dependant for all physiological indicators in each sex stratum.

#####**Study Aim 3**: To measure the strength of association between each demographic and physiological indicator with the outcome of mortality from cardiovascular disease.

#####**Hypothesis**: We expect that participants with specific physiological indicators will have a significantly increased odds of mortality from cardiovascular disease compared to participants without the same physiological indicators. Further, we expect to see a decreased odds of mortality for those diagnosed at a younger age compared to those diagnosed at an older age and decreased odds of mortality for those with a later year of onset compared to those with an earlier year of onset.

####**Exploratory Data Analysis**

#####Prior to performing tests to assess our study aims, we conducted an exploratory data analysis to better understand the MONICA dataset. The goal of our exploratory data analysis was to elucidate key characteristics of the dataset that would contribute to our interpretation of our main research findings. Key components of the dataset that we assessed, included: the frequency, average, and distribution of each indicator in regards to the outcome of mortality from cardiovascular disease, where applicable; and the amount of missing data for each explanatory variable.

####**Results of the Exploratory Data Analysis**

The MONICA dataset has a total of 6367 observations and 13 variables. The variables include a participant ID, the outcome variable, and 11 explanatory variables. A brief description of the variables is presented below:

####**ID and Outcome Variables**

#####**Participant ID**: categorical

#####**Outcome**: mortality from cardiovascular disease. Dichotomous, categorical variable (live/dead)

####**Explanatory Variables**

#####**Sex**: Sex of the participant. Dichotomous, categorical variable (f/m)

#####**Age**: Age of the participant at onset of cardiovascular disease. Continuous variable.

#####**Yronset**: Year during which the participant was diagnosed with cardiovascular disease. Discrete variable (19xx).

#####**Premi**: Previous myocardial infarction (MI) event. Categorical variable (y/n/nk not known)

#####**Smstat**: Smoking status. Ordinal, categorical variable (c current/x ex-smoker/n non-smoker/ nk not known).

#####**Diabetes**: Diabetes status (y/n/nk not known)

#####**Highbp**: High blood pressure status (y/n/nk not known)

#####**Hicho**: High cholesterol status (y/n/nk not known)

#####**Angina**: Agina status (y/n/nk not known)

#####**Stroke**: Stroke status (y/n/nk not known)

#####**Hosp**: Hospitalization status at the time of diagnosis of cardiovascular disease (y/n)

####**General Observations of the Exploratory Data Analysis**

#####Unknown Statuses

#####Of the 11 categorical variables that were assessed, only outcome, sex, and hospitalization at time of cardiovascular disease diagnosis had complete recording of known values. The remaining 7 explanatory variables had very high rates of unknown values recorded, all at greater than 10%. The 3 explanatory variables with the highest rates of unknown values were cholesterol status (19.36%), angina status (15.31%), and high blood pressure status (14.89%).

#####High rates of unknown statuses among the explanatory variables could greatly impact the results of our primary data analyses. First, missing data will reduce the statistical power of our sample, resulting in a higher probability of Type II error, or a greater likelihood that we would fail to reject a false null hypothesis. Secondly, missing data is a type of information bias that, depending on how the data was collected, may result in the differential or non-differential misclassification of our explanatory variables. If the unknown statuses of a certain explanatory variable were determined to be spread randomly across our study population, then we would conclude that there was a non-differential misclassification of the exposure, which would bias our measure of association toward the null. Alternatively, if the unknown statuses of a certain explanatory variable were not spread randomly across our study population, then we would conclude that there was a differential misclassification of the exposure, which would bias our measure of association upward or downward depending on the mortality rate among those with the misclassified exposure. Taken together, we’ll need to carefully consider how the high rates of unknown statues may impact our conclusions when interpreting the results our primary data analyses.

#####Non-normal Distribution of Demographic Explanatory Variables

#####Both of the continuous demographic explanatory variables follow a non-normal distribution. For both men and women, the age at which a participant was diagnosed with cardiovascular disease is skewed strongly to the left. This is expected since the incidence of cardiovascular disease increases with age. Thus, we would expect very few participants to have an age of diagnosis younger than 50. The opposite is true for the year of onset of cardiovascular disease, for which the distributions for both men and women are skewed to the right. This is expected since as time progresses, a greater number of efficacious measures are introduced to prevent cardiovascular disease. Thus, we would expect to see higher rates of cardiovascular disease diagnoses at earlier time points. We will need to be aware of these non-normal distributions in our Chi-square and logistic regression analyses.

#####Additionally, we see that the study sample is overwhelmingly male (72-28%). If we find that mortality from cardiovascular disease is dependent on the sex of the participant in our Chi-square analysis, then we will have to assess whether sex is a confounder or effect modifier when measuring the association of the various physiological explanatory variables in our logistic regression analyses. We suspect that sex may be a associated with mortality based upon the differing distrubtions that we saw.

####**Primary and Secondary Analyses**

####**Study Aim 1**

#####Our first study aim seeks to determine if there is a significant association between each of the demographic and physiological risk factors with mortality from cardiovascular disease. In order to statistically assess for an association, we performed a Pearson’s Chi-square test of independence.

####**Results**

#####The results from the Pearson’s Chi-square test with Yates' continuity correction show that the outcome of mortality from cardiovascular disease is significantly dependent upon the following variables:

#####Previous myocardial infarction (𝝌2=10.24, p=0.0014)

#####Diabetes (𝝌2=16.28, p=<0.0001)

#####High cholesterol (𝝌2=41.50, p=<0.0001)

#####Angina (𝝌2=26.91, p=<0.0001)

#####Stroke (𝝌2=52.69, p=<0.0001)

#####Hospitalization (𝝌2=3399.90, p=<0.0001)

#####Smoking status (𝝌2=6.90, p=0.0317)

#####Age at diagnosis (𝝌2=157.77, p=<0.0001)

#####Year of onset (𝝌2=21.295, p=<0.0001)

#####The variables for which we did not see a significant dependence are:

#####Sex (𝝌2=1.35e-7, p=0.9997)

#####High blood pressure (𝝌2=1.71, p=0.1915)

#####The results from the Pearson’s Chi-square test for independence mostly confirm our hypothesis for this study aim. Interestingly, in our sample, the outcome was not statistically dependent on a subjects high blood pressure status.

####**Study Aim 2**

#####Our second study aim seeks to determine if there is a significant association between each of the demographic and physiological risk factors with mortality from cardiovascular disease after stratifying by the sex of the participant. In order to statistically assess for this association, we performed the same Pearson’s Chi-square test of independence as for study aim 2, but we stratified the explanatory variables by sex.

#####Our hypothesis for this study aim is that we expect to see the outcome of mortality from cardiovascular disease to be significantly dependant for all physiological indicators in each sex stratum, or that sex is not modify the association between the predictors and outcome.

####**Results**

#####The results from the sex-stratified Pearson’s Chi-square test with Yates' continuity correction are listed below:

#####Sex

#####Female (𝝌2=20.06, p=<0.0001)

#####Male (𝝌2=53.21, p=<0.0001)

#####Previous myocardial infarction

#####Female (𝝌2=0.20, p=0.6571)

#####Male (𝝌2=11.54, p=0.0007)

#####Smoking status

#####Female (𝝌2=2.77, p=0.2501)

#####Male (𝝌2=9.63, p=0.0081)

#####Diabetes

#####Female (𝝌2=7.17, p=0.0074)

#####Male (𝝌2=9.43, p=0.0021)

#####High blood pressure

#####Female (𝝌2=0.17, p=0.6807)

#####Male (𝝌2=3.57, p=0.0587)

#####High cholesterol

#####Female (𝝌2=27.94, p=<0.0001)

#####Male (𝝌2=18.26, p=<0.0001)

#####Angina

#####Female (𝝌2=0.21, p=0.6433)

#####Male (𝝌2=33.67, p=<0.0001)

#####Stroke

#####Female (𝝌2=7.13, p=0.0076)

#####Male (𝝌2=48.25, p=<0.0001)

#####Hospitalization

#####Female (𝝌2=778.08, p=<0.0001)

#####Male (𝝌2=2631, p=<0.0001)

#####Age at Diagnosis

#####Female (𝝌2=25.90, p=0.0002)

#####Male (𝝌2=149.58, p=<0.0001)

#####Year of onset

#####Female (𝝌2=5.46, p=0.0194)

#####Male (𝝌2=15.60, p=<0.0001)

#####The outcome was not statistically dependent upon high blood pressure status for both females and males, which matches our results from the unstratified analysis in study aim 1. For the variables previous myocardial infarction, smoking status, and angina, we see significant dependence for males, but not females. These differences may in some part be attributed to the smaller sample size for females in the study, which would impact the power of our Pearson’s Chi-square test.

#####The results from the Pearson’s Chi-square test for independence mostly confirm our hypothesis for this study aim.

####**Study Aim 3**

#####Our third study aim seeks to measure the strength of association between the outcome of mortality from cardiovascular disease and each significantly dependent predictor, as determined from study aim 1. In order to statistically assess for the strength of association, we performed univariate logistic regression analyses for each dependent predictor. Logistic regression analysis is used to assess the strength of the association between a binary outcome and categorical and/or continuous predictors. In our study, we have a binary outcome of mortality from cardiovascular disease and 9 categorical predictors.

####**Results**

#####The results from the univariate logistic regression analyses are listed below:

#####Previous myocardial infarction

#####OR=1.22 (z=3.23, p=0.0012)

#####Interpretation of the OR: Subjects with a previous myocardial infarction have 1.22 times the odds (or 22% increased odds) of dying from cardiovascular disease compared to those subjects without a previous myocardial infarction. This result is significant with a z-value of 3.23 and a corresponding p-value of 0.0012.

#####Diabetes

#####OR: 1.37 (z=4.06, p=<0.0001)

#####Interpretation of the OR: Subjects with diabetes have 1.37 times the odds (or 37% increased odds) of dying from cardiovascular disease compared to those subjects without diabetes. This result is significant with a z-value of 4.06 and a corresponding p-value of <0.0001.

#####High Cholesterol

#####OR: 0.67 (z=-6.46, p=<0.0001)

#####Interpretation of the OR: Subjects with high cholesterol have 0.67 times the odds (or 33% decreased odds) of dying from cardiovascular disease compared to those subjects without high cholesterol. This result is significant with a z-value of -6.46 and a corresponding p-value of <0.0001..

#####Angina

#####OR: 1.36 (z=5.21, p=<0.0001)

#####Interpretation of the OR: Subjects with angina have 1.36 times the odds (or 36% increased odds) of dying from cardiovascular disease compared to those subjects without angina. This result is significant with a z-value of 5.21 and a corresponding p-value of <0.0001.

#####Stroke

#####OR: 1.91 (z=7.21, p=<0.0001)

#####Interpretation of the OR: Subjects with a stroke have 1.91 times the odds (or 91% increased odds) of dying from cardiovascular disease compared to those subjects without a stroke. This result is significant with a z-value of 7.21 and a corresponding p-value of <0.0001.

#####Hospitalization

#####OR: 0.0004 (z=-13.50, p=<0.0001)

#####Interpretation of the OR: Subjects being hospitalized have 0.0004 times the odds (or 9996% decreased odds) of dying from cardiovascular disease compared to those subjects not being hospitalized. This result is significant with a z-value of 4.06 and a corresponding p-value of <0.0001.

#####The predictor with the highest odds of cardiovascular disease mortality was stroke (OR=1.91), while the predictor with the lowest odds of cardiovascular disease mortality was hospitalization (OR:00004). Interestingly, we see a decreased odds of cardiovascular disease mortality in those with high cholesterol, which refutes are hypothesis for this study aim. A possible explanation for the extreme OR value for hospitalization is that the effect is confounded by the severity of cardiovascular events - perhaps those experiencing more extreme cardiovascular events died prior to making it to the hospital. The results from the univariate logistic regression analyses largely confirm our hypotheses for this study aim.

####**Summary of Key Findings from Project 1**

#####Of those risk factors that were associated with the outcome, most lead to an increased odds in mortality.

#####Interestingly, having high cholesterol lead to a decreased odds of mortality from cardiovascular disease.

#####Hospitalization status produced a very extreme odds ratio - 0.0004.

#####For those who died, many likely never made it to the hospital following the event. We suspect that the severity of the cardiovascular event may be confounding the relationship between hospitalization status and mortality from cardiovascular disease.

#####Mortality status was independent of a participant’s sex, however, when stratified by sex, some differences in independence for certain predictors - suspected interaction.

####**Project 2**

#####As described above, in Project 1 we sought to identify the explanatory variables that were associated with cardiovascular disease, the strength of these associations, and whether the associations changed depending upon the sex of the subject. Based upon our results and conclusions, the next phase of our analysis is to determine which combination of explanatory variables best predicts the out of mortality from cardiovascular disease. In Project 2, we used the following analyses to optimize the predicting capability of our model: forward selection logistic regression, K nearest neighbor (KNN), and decision trees and random forest using Bayesian Information Criteria BIC selection. Each of these methods and the results from the analyses will be detailed in subsequent sections.

####**Project 2 SMART Research Question and Study Aims**

#####After stratifying by sex, what combination of demographic and physiological predictors most accurately predict mortality from cardiovascular disease among a globally representative sample taken between 1985 and 1993?

#####**Study Aim 1**: To use a forward selection multivariable logistic regression model to identify the combination of demographic and physiological risk factors that will most accurately predict mortality from cardiovascular disease.

#####**Hypothesis**: We expect the four explanatory factors with the highest odds of mortality to form the combination of predictors that will most accurately predict mortality from cardiovascular disease.

#####**Study Aim 2**: To perform a KNN analysis to predict the outcome of mortality from cardiovascular disease.

#####**Hypothesis**: We expect to see that our prediction of the outcome will improve as the number of clusters increases.

#####**Study Aim 3**: Using Bayesian Information Criteria to find the most important features. Subsetting the data using these features and using the same for decision trees and random forrest analysis.

#####**Hypothesis**: We expect the combination of explanatory variables that result in the lowest BIC value to most accurately predict mortality from cardiovascular disease.

####**Study Aim 1 - Multivariable Logistic Regression**

#####Our first study aim seeks to identify the combination of demographic and physiological risk factors that will most accurately predict mortality from cardiovascular disease. In order to statistically assess how well the multivariable logistic regression models predict the outcome, we used a forward stepwise selection method and measured the McFadden pseudo R2 values, further referred to as the R2 value. Our forward stepwise selection process started with hospitalization status, since it was the most significantly associated predictor for mortality. Individual predictors were then added to the multivariable logistic regression model in successful order based upon maximizing the R2 value. The R2 value is calculated by using the following formula:

#####R2 = 1 - [loglik(fitted model)]/[loglik(intercept)]

#####where loglik is the log-likelihood. The R2 value measure the total variation in the outcome that is predicted by the explanatory variables in the regression model. Thus, the higher the R2 value, the more accurately we’re able to predict the outcome.2

####**Results**

#####The results from our stepwise selection multivariable logistic regression analysis are presented below:

[TABLE FOR R2 VALUES]

#####The stepwise selection process results in non-meaningful increases in R2 values after adding 3 predictors to the stratified models. Even with all explanatory variables included in our stratified analyses, the full models are still not fitting the data very well, with approximately 55-65% of the variance in the outcome predicted by the independent variables. When comparing the stratified models, the male subset fits the data better than the female subset. We expected this outcome due to the higher power obtained from the skewed sample (men n=4605; women n=1762).

####**Study Aim 2 - KNN**

#####The KNN analysis is a non-parametric, lazy learning algorithm used for classification and regression. The non-parametric nature of the analysis means that this method does not make any assumptions about the distribution of the dataset. Further, the lazy learning nature of the analysis means that this method does not learn any discriminative functions from the training data. Thus, the algorithm memorizes the training set and keeps all of the training data to make predictions.3 In our KNN analysis, we looked at increasing odd K values to determine the number in which we would most accurately predict our outcome of mortality from cardiovascular disease in our stratified sample. For the MONICA dataset, we would have a 38% chance of predicting an outcome of ‘dead’ and a 65% chance of predicting an outcome of ‘live’ if selected at random. After stratifying the data by sex, we took a proportion of 70% for the training sets and a proportion of 30% for the test sets. For the purposes of our analysis, we grouped the two numeric explanatory variables into categories.

####**Results**

#####The results from KNN analysis are presented below:

[TABLE FOR KNN PREDICTION VALUES]

#####The results from our KNN model show that it performs well for classifying ‘live’ but has limited accuracy for classifying ‘dead’. When analyzing the entire dataset, as the K value increases, so does the predictive quality for ‘live,’ but not for ‘dead.’ For the unstratified dataset, 7 K values were optimal for predicting ‘live’ (99.5%), while 3 K values were optimal for predicting ‘dead’ (66.1%). For the female subset, the optimal K value was 3; and for the male subset, the optimal K value was 13.

####**Study Aim 3 - Decision Trees and Random Forest**

#####Decision trees can be used in statistics to identify which explanatory variables are most important for predicting the outcome.4 In our analysis, we used a classification decision tree due to our categorical, dichotomous dependent variable of mortality from cardiovascular disease. The dataset was first stratified by sex, then we found the most significant factors using Bayesian Information Criteria. The decision tree was then built using these factors. Lastly we used random forest to find the accuracy of predicting death.

[CODE FOR DECISION TREES AND RANDOM FOREST]

#####The results of the decision tree analysis that was generated after feature selection for females found hospitalization and angina to the most important factors. When we generated a tree using the entire dataset we found that hospitalization and high blood pressure were the most important features. The decision tree that was generated after feature selection for males found hospitalization and high blood pressure to the most important factors. When we generated a tree using the entire dataset we found that hospitalization, diabetes, high blood pressure and age of diagnosis were the most important features. We used 500 trees for the random forest analysis. It was used to predict death. The error rate for females was higher than that for males.

####Summary Conclusions

#####Full model does not fit the data very well. Hospitalization status was the strongest predictor of cardiovascular mortality. Analysis would have been more robust if quantitative variables were collected. In comparing the different analytical methods used, we conclude that the KNN method produced the most accurate predictions of the outcome. While the KNN method did poorly with predicting death, the model was nearly perfect in predicting which participants would live. The KNN method outperforms the multivariable logistic regression model, which produced McFadden R^2 values in the moderate 55-66% range.

Study validity - Moderate internal validity: + strong sampling methods, - high rate of missing values. Moderate external validity: representative to a European male population

Major Findings:

1. Logistic Regression
   1. Logistic Regression run for every variable, stratified by sex
   2. Can’t do PCA because the data is categorical- but we used BIC for feature selection
   3. With 2 - 3 variables still has a low level of precision (about 60%) each increasing factor increased relevance slightly.
   4. Higher relevance included all six categorical variables

***FEMALES***

**Full model all 8 variables:**

Significant variables: hosp, hichol, stroke

McFadden = 0.585 HIGHEST AF ALL

With the McFadden value of 0.585, about 58.5% of the variation in y is explained by this model.

**1-variate of all variables:**

Variable with highest McFadden is hosp: 0.395

**2-variate model with highest McFadden:** Hosp + hichol

McFadden = 0.541

**3-variate model with highest McFadden:** Hosp + hichol + yronset

McFadden = 0.562

**4-variate model with highest McFadden:** Hosp + hichol + yronset + angina

McFadden = 0.564

**5-variate model with highest McFadden:** Hosp + hichol + yronset + angina + stroke

McFadden = 0.572

**6-variate model with highest McFadden:** Hosp + hichol + yronset + angina + stroke + age

McFadden = 0.578

**7-variate model with highest McFadden:** Hosp + hichol + yronset + angina + stroke + age + premi

McFadden = 0.583

**Diabetes is last to add, then we have the full model and highest McFadden: 0.585**

**Since the highest variation that can attain with Logistic Regression is roughly 58.5%, we will try other models that could be potentially be a best fit.**

***MALES***

**Full model all 8 variables:**

Significant variables: hosp, stroke, yronset, angina, premi

McFadden = 0.660

**1-variate of all variables:**

Variable with highest McFadden is hosp: 0.511

**2-variate model with highest McFadden:** Hosp + hichol

McFadden = 0.615

**3-variate model with highest McFadden:** Hosp + hichol + angina

McFadden = 0.633

**4-variate model with highest McFadden:** Hosp + hichol + angina + age

McFadden = 0. 643

**5-variate model with highest McFadden:** Hosp + hichol + +angina + age + premi

McFadden = 0.651

**6-variate model with highest McFadden:** Hosp + hichol + angina + age + premi + stroke

McFadden = 0.657

**7-variate model with highest McFadden:** Hosp + hichol + angina + age + premi + stroke + yronset

McFadden = 0.661

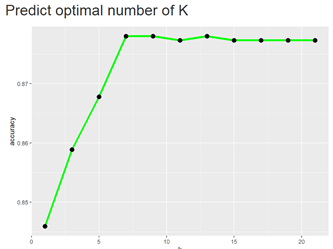
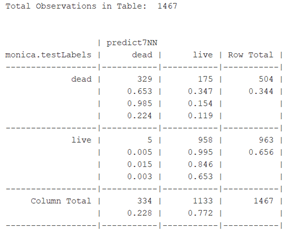
**If we add diabetes, then we have the full model with McFadden: 0.660 which is slightly smaller. Therefore, the 7-variate model is better.**

**Since the highest variation that can attain with Logistic Regression is roughly 66%, we will try other models that could be potentially be a best fit.**

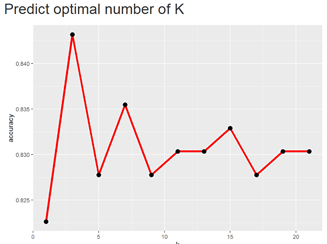
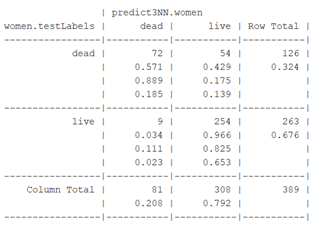
1. **K Nearest Neighbor**

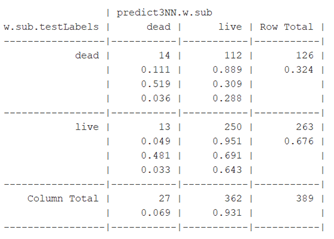
* 70% training set and 30% test set
* Year of onset and age of diagnosis were binned and included in the analysis.

**Classify outcome based on age, sex, and all risk predictors**

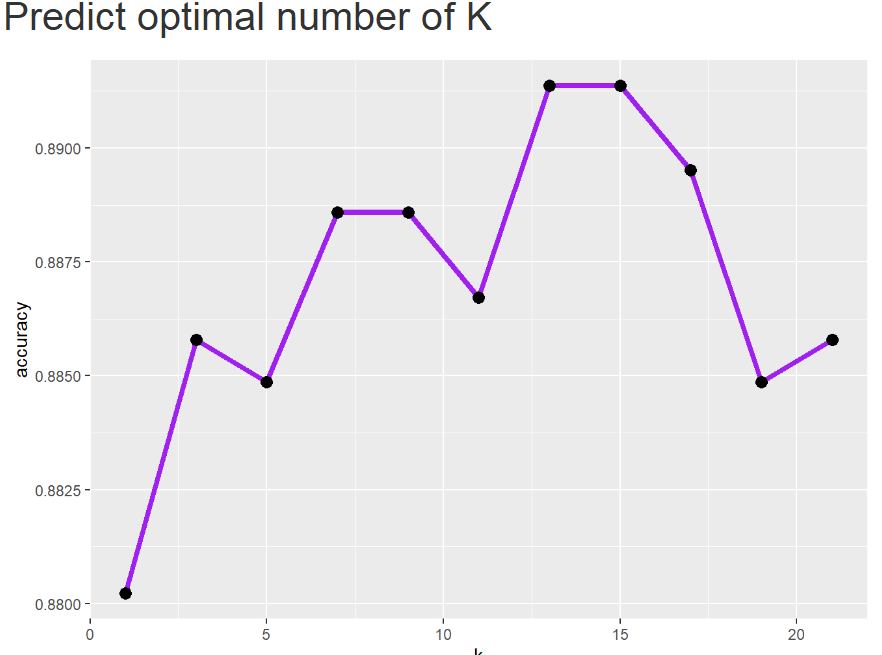
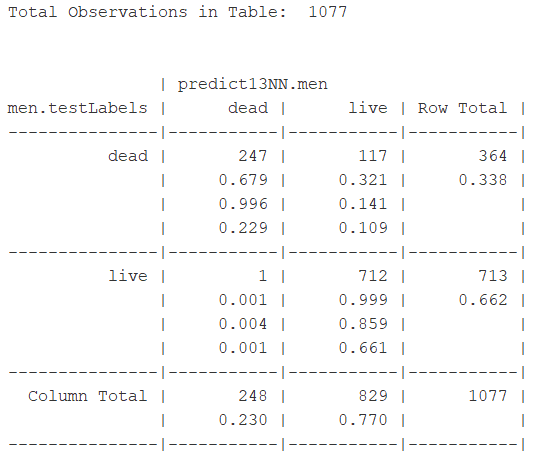
* With **7-nearest neighbors**, the classifier has a predictive accuracy of 87.73%.
* The classifier is accurate for predicting outcome = live: 99.5% was classified correctly.
* The classifier is not accurate for predicting outcome = dead: 65.3% of “dead” cases classified correctly. This suggests that other factors, that are not considered in the dataset, might have an impact on mortality from cardiovascular disease. 

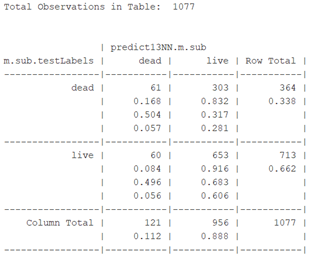
**Classify outcome based on age and all risk predictors for WOMEN**

* With **3-nearest neighbors,** the classifier has an accuracy of 83.8%
* The classifier is accurate for predicting outcome = live: 96.1% was classified correctly.
* The classifier is not accurate for predicting outcome = dead: 57.1% classified correctly
* *Subset significant predictors for women* (diabetes, high cholesterol, stroke, hospitalization, year of onset, and age of diagnosis): 11.1% of outcome=dead was classified correctly; and the predictive accuracy of the classifier dropped to **67.9%:**

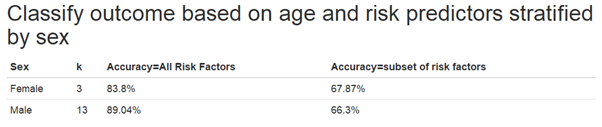
****

**Classify outcome based on age and all risk predictors for MEN**

* With 13-nearest neighbors, the classifier has an accuracy of 89%
* The classifier is accurate for predicting outcome = live: 99.9% was classified correctly.
* The classifier is not accurate for predicting outcome = dead: 67.9% classified correctly.
* *Subset significant predictors for men* (previous myocardial infarction, smoking status, angina, hospitalization, year of onset, and age of diagnosis): 16.75% of outcome=dead was classified correctly; and the predictive accuracy of the classifier dropped to **66.3%:**

****

Summary of analysis stratified by sex:



* 1. **We can predict live and but not dead. What factors are we missing from the data that will help us predict death? Poverty, geography, country of origin, access to health care, etc**

1. Decision Trees
   1. Feature selection for performed stratified by sex:
   2. For females it was found that yronset, angina and hosp are the most significant.
   3. For males it was found that yronset, premi, smstat, highbp, hichol, stroke and hosp are the most significant.
   4. Since the trees use the most significant variables are root node and stop once the data has been perfectly fit, we can see that according to decision trees:
      1. Hosp and angina are the most significant features females
      2. Hosp and stroke are the most significant for males.
   5. While running the decision tree without feature selection it is seen that hosp and highbp are the most significant for females.
   6. And Hosp, age, diabetes and highbp are the most significant in males when the model is run without feature selection.
   7. Accuracy:
      1. Females: 85%
      2. Males: 54%
2. Random Forest: number of trees used in each model: 500
   1. Runs on multiple trees: I used the entire dataset for random forest
      1. Females: error rate: 14.8%
      2. Males: error rate: 10.9%

Bibliography

1. The World Health Organization. Background, Development and Organization of MONICA. WHO MONICA Project e-publications. January 4, 2013.

<https://www.thl.fi/publications/monica/index.html>

1. Lo, Edwin. (2019) Data Science Generalized Linear Models: Logistic Regression. Lecture. [PowerPoint slides]. Retreived from <https://blackboard.gwu.edu/webapps/blackboard/execute/content/file?cmd=view&content_id=_9827630_1&course_id=_322369_1>
2. Lo, Edwin. (2019) K-Nearest Neighbor (KNN) K Means. Lecture. [PowerPoint slides]. Retreived from <https://blackboard.gwu.edu/webapps/blackboard/execute/content/file?cmd=view&content_id=_9837923_1&course_id=_322369_1>
3. Lo, Edwin. (2019) Decision Trees. Lecture. [PowerPoint slides]. Retreived from <https://blackboard.gwu.edu/webapps/blackboard/execute/content/file?cmd=view&content_id=_9853452_1&course_id=_322369_1>