D209 Performance Assessment Task 1

October 30, 2021

**Part A**

1. Using the medical dataset, a question that an organization could ask is to ask what factors within the dataset are affected by Age and how does Age affect other factors within the dataset.
2. The main goal of this data analysis is using a Naïve Bayes classification to determine what factors affect Age the most and how the data is affected by Age.

**Part B**

1. I have chosen to use the Naïve Bayes theorem because Bayes theorem demonstrates high accuracy when it comes to multi-class classification and is a very simple premise of comparing an independent variable with each predictor variable in order to check exactly how related each predictor is to the independent. The hope with this data mining project with the use of naïve bayes theorem is to figure out factors that could be influenced by the age of the patient.
2. One of the big assumptions that is needed in a Naïve Bayes theorem test is that all the factors need to be independent of each other.
3. For this analysis I chose to use python because that’s the best language I can use, plus it has a wide arrange of libraries that allow for various machine learning, data mining, and analysis tools and visualizations. Pandas was selected since it’s needed in order to work with databases. Numpy was used for calculations and numerical functions such as absolute values. Matplotlib was used in order to create figures and save them into a jpg format. Scipy was used for calculating zscores and standard deviations to remove outliers. Seaborn was imported to add graphical capabilities to the charts being created. Sklearn imported LabelEncoder, train\_test\_split, StandardScaler, GaussianNB, confusion\_matrix, accuracty\_score, and classification\_report all to help train, test, and fit the model that I was creating. They also helped create a confusion matrix to help understand the results of this analysis better and to get an overall accuracy on the results. I also imported the warnings library because I would get a lot of false warnings that were annoying so I could hide them.

**Part C**

1. One of the first things that I had to do to help clean the data was to replace all categorical variables with numerical variables in order to perform an analysis on them. I created a dictionary and replaced certain values in columns with numerical ones in order to analyze them. I also created a heatmap in order to see different relationships between certain variables and target the variables I wanted to use as my predictor variables in my analysis for Age.
2. For my analysis, I decided to throw everything but the kitchen sink into my analysis in order to really see what was operating under the hood. I threw in City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Children, Age, Income, Marital, Gender, ReAdmis, VitD\_levels, Doc\_visits, Full\_meals\_eaten, vitD\_supp, Soft\_drink, Initial\_admin, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Services, Initial\_days, TotalCharge, and Additional\_charges. I checked for duplicates, outliers, and then used a dictionary to replace certain categorical columns with numbers to allow for a proper analysis. I then created a heatmap and saw what factors stood out and reduced the table into Zip, Lat, Lng, Population, Age, Initial\_days, TotalCharge, and Additional\_charges, which were all continuous variables, while also taking away ReAdmis and HighBlood as categorical variables for my analysis.
3. For my analysis, I first created a data file with all the columns that I wanted to check. I then checked for duplicates or null values, followed by removing all rows that were outside 3 standard deviations, or 99.7 percentile. Next, I had to create a histogram to check the distribution of all the variables followed by a heatmap to see all variables compared to every other variable in a bivariate analysis. Using the heatmap, I took out the categories that were highlighted and then used them as my dependent variables and stuck to using Age as my dependent variable. I then created a gaussian model which I then trained my data using a train\_test\_split method. After creating my model and training it, I transformed the data and created a classifier for predicting the outcome of the model as well as calculate the accuracy. After taking the accuracy of the model, it came out with a 2.5% accuracy, which is abysmal. I then ran a classification report to check the precision, recall, f1-score, and support scores of the confusion matrix that I produced from this analysis which is attached.
4. The cleaned dataset is attached

**Part D**

1. This analysis split the training and prediction data into 2 csv files which have been attached.
2. The analytical technique I used for this analysis is a Naïve Bayes theorem because it seems to be used more often in machine learning models which is what I’m interested in. Since it works predominantly with independent variables and comparing them to the dependent variables, it seemed to make sense to use Naïve Bayes in order to check on what factors seem to be affected by Age the most. The main equation is as follows.

P(c|x) is the probability of the target given the predictor, P(c) is the prior probability of the target, P(x|c) is the likelihood of the predictor given the target, and P(x) being the probability of the predictor. An example calculation could be as follows regarding my dataset

The above equation would help to determine if a patient’s high blood pressure is being caused by the age of the patient. Further calculations in the classification report and confusion matrix can be seen in the attached file.

1. See attached file for code implementation.

**Part E**

1. For my analysis the accuracy was abysmal. I got a 2.5% accuracy when comparing the effects of Age on other variables within the dataset. An accuracy this terrible means that this analysis pretty much showed absolutely nothing of value unless the finding that Age doesn’t influence all the variables that were measured with it. The area under the curve demonstrates what kind of analysis should be used for this data set. The information is drawn from the confusion matrix and determines all possible outcomes and compares them. This area under the curve shows that there is very little accuracy within this analysis’ confusion matrix when compared to all the false positive/negative rate and true positive/negative rates.
2. After going through this full analysis, I have determined that Age doesn’t seem to influence increasing or decreasing anything. With an accuracy of only 2.5% however, it’d be hard to make any sort of claim that this is conclusive. One way that our question is answered could be by saying that after analyzing Age using a Naïve Bayesian method with multiple other independent variables, shows that Age doesn’t seem to have much of an effect on anything.
3. One major limitation of this analysis is how to exactly decide which variables to reduce the model down into in order to analyze with a Naïve Bayes theorem. I ended up using a heatmap and selected variables based on that. Considering the nature of some of this data, it could have made sense to perform a logistic regression on the data first in order to pull out variables that would have more of a direct impact on the results.
4. The recommended course of action for this analysis would simply be to stay put. The analysis doesn’t seem to show any kind of increased or decreased probability of Age causing problems down the line or affecting/being affected by the variables I tested. I think it would be important to reassess the factors that could influence age, while also taking time to test only categorical and only continuous variables in order to provide a cleaner analysis.

**Panopto video:** <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c2db3e42-18b4-4794-b5e9-adca01729d2b>

**References:**

<https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html> sklearn.naive\_bayes.GaussianNB, scikit-learn

<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> 6 Easy Steps to Learn Naïve Bayes Algorithm with codes in Python and R, Analytics Vidhya

<https://www.saedsayad.com/naive_bayesian.htm> Naive Bayesian, saedsayad

<https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/> A Guide to the Naïve Bayes Theorem, Analytics Vidhya