Megan Hoeksema

Comp448: Data Science Tools 2

Final Project

**Introduction**

The SmokeBan dataset (<https://vincentarelbundock.github.io/Rdatasets/doc/AER/SmokeBan.html>) will be utilized. The output variable utilized in this dataset will be smoker which is a categorical variable in which the participant has answered yes or no. The input variables include whether there is a workplace smoking ban, age, highest level of education, whether the individual is African American, whether the individual is Hispanic, and the individual’s gender. The only numerical variable is age, with all others being categorical. Descriptive statistics were performed on the age variable.

Of the following algorithms, KNN, Decision Trees, Random Forest, and Naïve Bayes, which will produce a better model for predicting the smoking status of indoor workers based on race, gender, age, education, and workplace smoking bans?

**Data Pre-Processing**

Data pre-processing included importing the data, checking for missing values, checking for uniqueness among each variable, converting the categorical variables to numerical, scaling the numerical variables, then splitting the data. To import the data, the data was downloaded via csv from the data repository, then imported into jupyter notebook using a pandas data frame. After importing the data, the data set was evaluated for missing values then each variable was checked for uniqueness. The goal of checked in column for unique values is to determine if there may be multiple inputs for the same value, for instances specifying male could also be Male, M, or m. This would create more categories than present in the data set and would produce incorrect results. The smokeBan data set contained no missing values and there were no variables that needed extra cleaning. The categorical variables were then converted to numerical using dummy variables. Age was then scaled using a MinMaxScaler.

Table

Description automatically generated

Using the raw pre-transformed data, several visualizations were created to further understand what was contained within the dataset. As age was really the only numerical value, descriptive statistics were performed on the age column. The median age of the entire dataset is 37 years old.

Graphical user interface, text, application, email

Description automatically generated

An age visualization was also created to show the average age per education and smoking statuses. Based on this visualization, each group had an average age between 35-45 years old. The goal of looking at this is to show that these groups have roughly the same age distribution.

Chart, bar chart

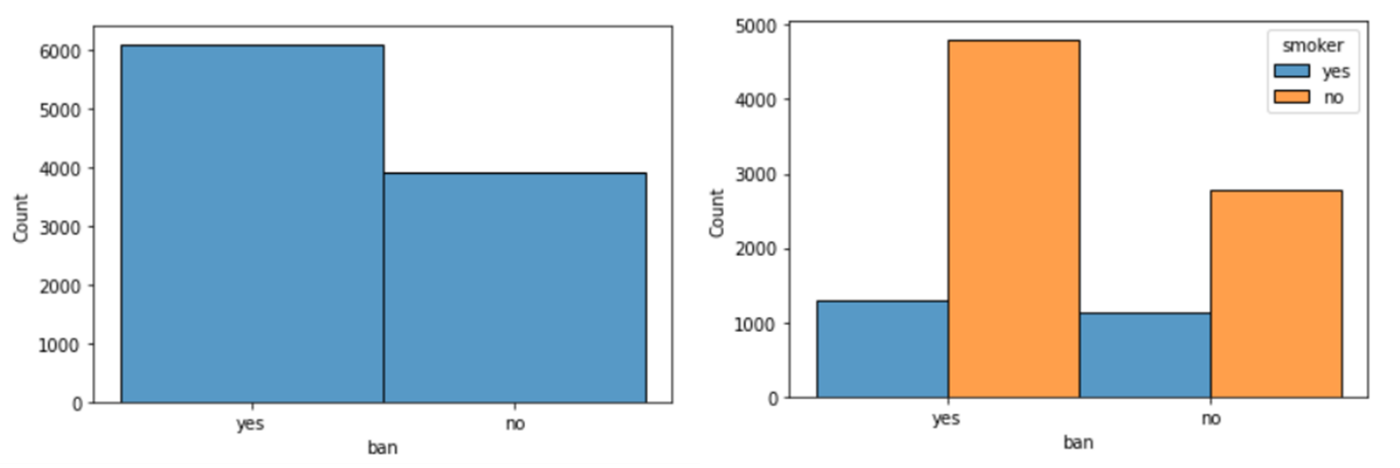
Description automatically generated

Based on a histogram plot of the counts of each gender, we can conclude that more women participated in this data collection than men.

Chart, histogram

Description automatically generated

Finally, we can visualize the number of participants that worked in a place with and without a ban. Overall, around 6000 participants worked in a place containing a smoking ban, and around 4000 do not work at a place containing a smoking ban. A second visualization was graphed to show the number of smokers and nonsmokers per ban status. This graph shows that there are many more participants who are nonsmokers with a workplace ban.



After the dataset was visualized as a whole, train\_test\_split was used to split the dataset 70/30. This ratio was used because it is the most used ratio to split data into training and test sets. With this data having 10,000 instances, 7,000 instances were randomly split into the training set and 3,000 instances were randomly split into the test set.

Text

Description automatically generated

**Model building and Evaluation**

Four classification models were then set up: Decision Trees, Random Forests, KNN and Naïve Bayes. The packages for each of these models were from sklearn. After the model was set up, the model was used to predict on both the training and test sets. Accuracy scores were printed out for each model on each set.

Text

Description automatically generated

Text

Description automatically generated

Text

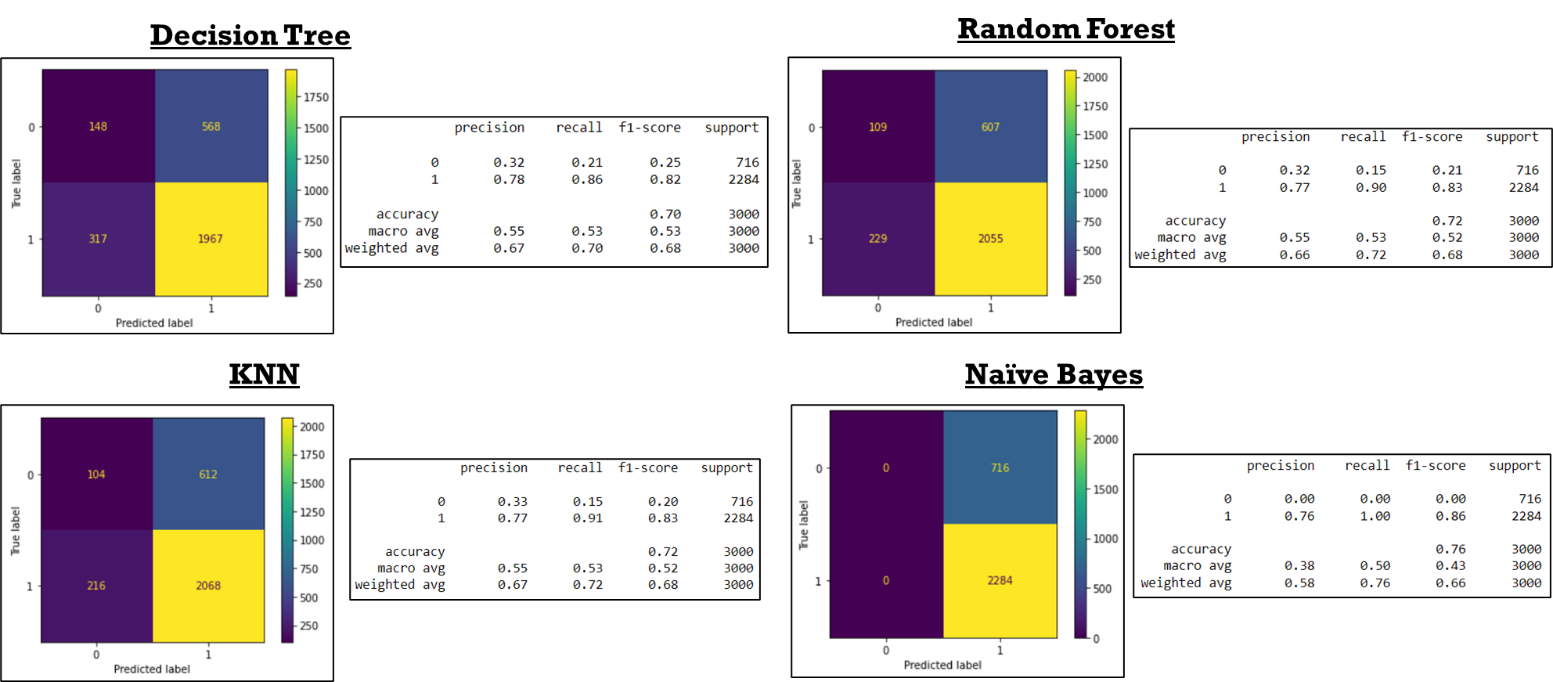
Description automatically generated

Text

Description automatically generated

Based on the accuracy of the training set, both Decision Tree and Random Forest perform equally well, with KNN and Naïve Bayes slightly less accurate. However, when looking at the accuracy scores for the test set, Naïve Bayes outperforms all other models.

For each of the models, a classification report and confusion matrix were printed out for each model.



Based on these data, Naïve Bayes would be the best model to use to predict the smoking status of indoor workers. However, we further refined the models by tuning some of the hyperparameters. For Decision trees, we chose to tune max\_depth, max\_features, and criterion. The best estimator proved to be max\_depth = 4, max\_features = 0.8, criterion = entropy. This increased the accuracy score of the training set from 0.705 to 0.76.

Graphical user interface, text, application, email

Description automatically generated

After tuning the max\_depth, max\_features, and n\_estimators, the best estimators were max\_depth = 7, max\_features = 0.8, n\_estimators = 30 which improved the accuracy of test set from 0.72 to 0.77.

Graphical user interface, text, application, email

Description automatically generated

For KNN, the default n\_neighbors are 5, so we tuned that parameter which resulted in the best estimator being 39 neighbors. This improved the accuracy of the test from 0.724 to 0.76.

Graphical user interface, text, application, email

Description automatically generated

Tuning of Naïve Bayes consisted of tuning the alpha, fit\_prior, and min\_categories parameters. After tuning, the best parameters were alpha = 0.01, fit\_prior = True, min\_categories = 18. Naïve Bayes was the only model that did not improve after tuning the hyperparameters. The accuracy of the model remained at 0.76.

Graphical user interface, text, application, email

Description automatically generated

Table of Accuracy Scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Test | Test (Tuned) |
| Decision Trees | 0.804 | 0.705 | 0.76 |
| Random Forest | 0.804 | 0.721 | 0.77 |
| KNN | 0.771 | 0.724 | 0.76 |
| Naïve Bayes | 0.756 | 0.761 | 0.76 |

**Conclusion**

In conclusion, based on the data pre-processing and data exploration, there were no missing values present in this dataset. We found the median age of the individuals in this dataset was 37. Based on the graphical representation of gender per education group and smoking status is between 35-45 for each group. We also found that more women participated in this data collection than men. Based on graphical representation of the number of smokers per ban category, the number of nonsmokers is much higher in workplaces that have a ban than those that do not have a ban. Based on the graphical representation of gender per education group and smoking status is between 35-45 for each group. We also found that more women participated in this data collection than men. Graphical representation of the number of smokers per ban category revealed that the number of nonsmokers is much higher in workplaces that have a ban than those that do not have a ban. Based on the accuracy of the training sets, both Decision Tree and Random Forest perform equally well, with KNN and Naïve Bayes slightly less accurate. When comparing the accuracy scores for the test set, Naïve Bayes outperforms all other models. We further looked to refine our models by tuning various parameters for each of the models. Tuning Naïve Bayes did not increase the accuracy score. After tuning the other three models, the accuracy of Decision Tree and KNN rose to 0.76. Random Forest appeared to perform the best (slightly) with an accuracy score of 0.77 after tuning parameters, however, with all of the accuracies being almost identical, any one of these models would be a good option to use for determining the smoking status of indoor workers.