Lovedeep Bajaj

Which Individual Is at Risk of Getting Diabetes

**Abstract:**

Approximately34.2 million people in America suffer from diabetes. My goal was to try and predict whether someone is likely to get diabetes before they even have it. The dataset that I worked with allows me to do exactly that. The features it contains includes sudden weight loss, visual blurring, weakness, and many more that can help classify whether someone will have diabetes or not. Since my goal was classification, I wanted to choose between the decision tree and random forest in terms of which one will yield the best results. Once the analysis was complete, the random forest was found to be the best machine learning technique to utilize for this dataset.

**Related Work:**

The dataset was obtained from the Kaggle website. By searching for the user, Akash Kumar, one can find the dataset, since he is the individual who uploaded it. Also, on the same website, one can see that few people have conducted an analysis on this dataset. Those who have worked on the dataset mostly used logistic regression and ensemble methods, which is not my primary focus. I want to compare the decision tree to the random forest and determine which is the best classification technique to use with this dataset.

**Data:**

The dataset only contained 520 rows and most of the features contained no and yes as the values. The only attributes that did not have no or yes as their values were age and class. For th4 age attribute, the values were in years, while the class attribute contained the values as positive or negative. A positive class value indicated that the person has diabetes and a negative class value indicated that the individual does not have diabetes. The summary of the data frame was reviewed to check for any missing values in the data. After finding no missing values, the number of duplicates was determined. There were 269 duplicated rows, but they were not removed due to making up 51% of the dataset. The dataset was already so small, so removing the duplicates would not have been helpful.

Next, a function was created that would produce the visualizations. The same type of chart would be created for each feature. The only attribute that the function was not used for was the class attribute, since in the function, I used it to fill the bars of the chart. The charts were created to get a better understanding of the data distribution.

**Figure 1.**

**Chart, bar chart

Description automatically generated**

Figure 1. Represents the age range distribution of the individuals in the dataset, and the number of individuals that have diabetes in each age range. People between the ages of 30 and 69 are the individuals more likely to be positive for diabetes.

**Chart, bar chart

Description automatically generatedFigure 2.**

Figure 2. Represents the gender distribution in the dataset in which most of the people are male. Also, females in this dataset make up most of the individuals who have diabetes.

**Figure 3.**

**Chart, bar chart, histogram

Description automatically generated**

Figure 3. Represents the number of individuals who have polyuria in the dataset. Most of the individuals who have polyuria also have diabetes.

**Chart, bar chart, histogram

Description automatically generatedFigure 4.**

Figure 4. Represents the number of people who have polydipsia and diabetes. If an individual has polydipsia, they are more likely to have diabetes.

More charts were created, which can be found in the analysis. Next, the attribute values were transformed. All the attributes were converted to character types. For the gender attribute, males and females were replaced with 0 and 1, respectively. The features that contained the values no and yes were converted to 0 and 1, respectively. Lastly, the class attribute values were converted to 0 and 1, where 0 represented negative and 1 represented positive.

**Methodology:**

After the data was converted, it was scaled due to the age feature containing values that were greater than 0 and 1. A function was created that conducts min-max scaling to convert the age values to fall in the range of 0 and 1. Next, all of the attributes were converted to be of type factor. Before determining whether the decision tree or random forest was the best to use, the data was split into training and testing. Since the data was not imbalanced, there was less of a worry about the training or testing containing no individuals that have diabetes. A 70/30 split was performed, in which 70% of the data was utilized for the training set and 30% was utilized for the testing set.

Determining the testing and training sets allowed me to move onto model comparison. The decision tree was the first model to be created. By training the model, the confusion matrix was predicted and examined. The confusion matrix displayed that the decision tree has an accuracy of 92% and a sensitivity of 93%. On the other hand, the random forest had an accuracy and sensitivity of 96% and 91%, respectively.

**Results:**

By obtaining the results of the two models, I concluded that the best model to use for the dataset is the random forest. Even though the dataset contains medical data, it is not imbalanced, and contains more individuals with diabetes than without. This is why my main focus is involving accuracy rather than sensitivity.

**References:**

1. <https://www.kaggle.com/singhakash/early-stage-diabetes-risk-prediction-datasets>
2. <https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html#:~:text=34.2%20million%20Americans%E2%80%94just%20over,Asians%20and%20non%2DHispanic%20whites>.
3. <https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#>
4. <https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>
5. <https://ggplot2.tidyverse.org/>