Report

Dataset Description

I have chosen to analysis the Diabetes dataset. The dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of the dataset is to predict if a patient has diabetes or not. The patients are all women who are at least 21 years old with Pima Indian heritage. The variable we want to predict is Outcome. Outcome consists of two classes, 0 and 1 where 0 is denoted as a patient not having diabetes and 1 is denoted as a patient having diabetes. The rest of the variables are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age. All of the variables are numerical datatype. The dimensions of the dataset are 768 rows by 9 columns.

Model Description

K-Nearest Neighbour

The variable we are predicting only has 2 classes, which makes KNN a useful model for binary classification problems. The Outcome variable has no distribution, which is perfect to predict with KNN because the model is nonparametric. Since there are 8 features and 768 instances in the dataset, which means the model performs best when a low number of features and instances are used because a low computation power is required. The model doesn’t consider the distribution of the datapoints in a scatterplot. Finally, the run time of building the model with training data is much faster compared to other classification models.

Classification and Regression Trees (Decision Trees)

The variable we are predicting only has 2 classes, which makes Decision Trees a useful classification model for binary classification problems. The Outcome variable has no distribution, which is perfect to predict with the model because the model is nonparametric. The model is easy to understand and interpret because it is similar to human level of thinking. All the variables in the dataset are numerical. This makes it possible for internal nodes to check a value from a feature against a particular threshold. The Gini Index is the best Attribute Selection Measure to use because the Outcome variable has two classes. This makes the Classification and Regression Tree a great algorithm to use. We only have to do a little amount of preprocessing on the dataset.

Experiments

Main Questions

* *Do patients have diabetes or not?*
* *What is the relationship between Outcome and the other variables?* This question was created because all the variables in the dataset are numerical. The question was answered with a correlation heatmap. The correlations of each variable with Outcome are Pregnanices=0.22, Glucose=0.47, BloodPressure=0.065, SkinThickness=0.075, Insulin=0.13, BMI=0.29, DiabetesPedigreeFunction=0.17 and Age=0.24.
* *What is the size of each classes in the Outcome variable?* A barplot was created to answer this question. There are about 500 patients who are not diabetic (denoted as 0) and about 275 patients who are diabetic (denoted as 1).

Choice of Parameters

* The KNN model began with the parameters K = 3 and Euclidean Distance.
* The Decision Tree model began with the parameters max\_depth = None and Attribute Selection Measure selected is the Gini Index.

Tuning

* Grid Search Cross Validation with cv = 5 determined that K = 8 is the best parameter for the KNN model. The Euclidean Distance parameter was selected again.
* Grid Search Cross Validation with cv = 5 determined that max\_depth = 4 and Gini Index are the best parameters for the Decision Tree model.

Results

What is the relationship between Outcome What is the size of each class in

![Chart, bar chart

Description automatically generated]()![A picture containing chart, timeline

Description automatically generated]() and the other variables? the Outcome variable?

![Table

Description automatically generated]()Table

Description automatically generatedKNN outputs before tuning KNN outputs after tuning

![Table

Description automatically generated]()![Table

Description automatically generated]() Decision Tree Outputs before tuning Decision Tree Outputs after tuning

Discussion

Results

In the correlation heatmap, the three variables that are most correlated with Outcome were chosen to be used as features for both models. They are Glucose, BMI and Age. In the barplot, the size not diabetic class is almost double the size of the diabetic class. This could affect the evaluation metrics of both the KNN and the Decision Trees models due to the possibility of the models being bias in classifying new instances into the larger sized class.

Evaluation of Accuracy

Tuning the parameter had definitely increased the Accuracy, Precision, Recall and F1-Score in both models. In KNN, the confusion matrix shows that the True Positive had increased but the True Negative had decreased. After tuning, the model became better at predicting one class but become worse at predicting the other class. Looking at the Precision, Recall and F1-Score for both classes, we can see that the three metrics have increased for class 0 but Recall and F1-Score had decreased for class 1. In Decision Tree, the confusion matrix shows that both the True Positive and the True Negative had increased. After tuning, the model became better at predicting both classes. The Precision, Recall and F1-Score for both classes have increased.

In both models, Precision, Recall and F1-Score are higher in class 0 compared to class 1. This suggests that both models were biased towards classifying the new instances as belonging to class 0. In other words, both models are more likely to predict a patient not having diabetes. This is due to class 0 being a larger sized class than class 1.

Algorithm with better outcome

The model that is best at predicting whether a patient has diabetes or not is the Decision Tree model because the model has a higher True Positive, True Negative, Accuracy, Precision, Recall and F1-Score than the KNN model.

References

*Pima Indians Diabetes Database* 2016, Kaggle, accessed 23 October 2020, <<https://www.kaggle.com/uciml/pima-indians-diabetes-database>>