A comparison of Decision Tree (DT) and Naïve Bayes (NB) for prediction of diabetes

INM431 Machine Learning Coursework | Vighnesh Mane| City University of London

Introduction

The purpose of this study is to compare decision tree classification to Naïve bayes classification. These strategies are used to the UCI Diabetes Pima Indians dataset. The primary goal of this work is to compare the performance of these two models to that of different research papers while using various machine learning approaches.

Context and Content

- The Pima Indian diabetes dataset was derived from the National Institute of Diabetes and Digestive and Kidney Diseases and is available through the UCI Machine Learning Repository.
- The collection contains data on 768 female patients who are at least 21 years old and their nine unique characteristics. Table 1 contains a description of the dataset's properties.
- Eight of the nine traits are independent/feature variables, while the ninth is dependent/target variable.
- Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome are all predicted.
- The result property has a binary value between 0 and 1, indicating that if the value is 0, the patient does not have diabetes, and if the value is 1, the patient does [fig b.].

Exploratory data analysis

• Firstly, the operation to check missing values is executed and after finding no missing values normalization is performed on feature variables as it was observed that the features were having different ranges.
• As certain medical parameters such as BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function cannot be zero. Therefore, the zero values are inspected and replaced with the median of it corresponding to the outcome attribute.
• Then outliers are examined and replaced the outliers with the mean of the previous value of the attribute.
• Furthermore, the relationship of the variable is investigated by performing correlation. The value above 0.5 and below 0.5 is displaying some correlation and the attributes having zero value are not correlated [fig a.].

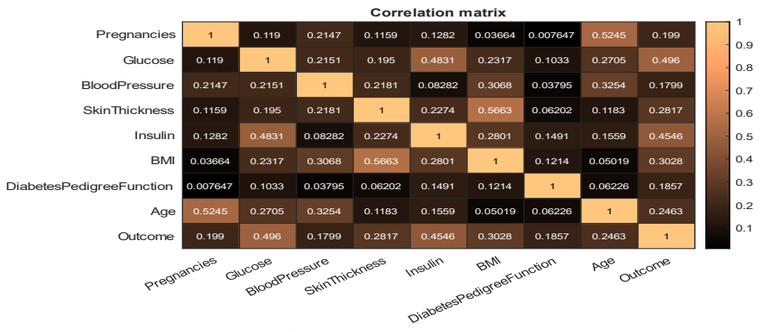

fig a. Correlation matrix

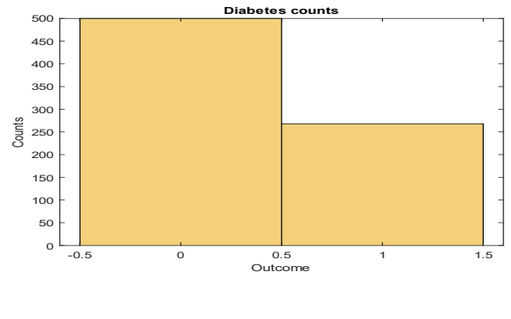

fig b. Diabetes counts

Attributes	Mean	Median	Std
Pregnancies	0.2220	0.1764	0.1926
Glucose	0.6114	0.5879	0.1530
Insulin	0.1559	0.1211	0.0712
Blood pressure	0.5929	0.5901	0.0930
Skin thickness	0.2914	0.2828	0.0843
DPF	0.1591	0.1227	0.1200
BMI	0.4806	0.4769	0.0976
Age	0.2088	0.1333	0.1805

table a. Statistics after normalization

Decision Tree Classification

- The decision tree is composed of multiple nodes, each of which carries a branch that is reliant on a distinct circumstance.
- The data is separated into branches until a predefined cutoff value is reached. A decision tree is composed of three types of nodes: root nodes, child nodes, and leaf nodes .
- To begin the process of developing a decision tree model, features are selected and each feature is converted to a decision node. The tree is then constructed by repeating the feature selection process for each child until the last attribute is reached.

Advantages

• Beneficial for decision-making difficulties.
• A decision tree is capable of recognizing both numerical and categorical input.
• Achievement of a feasible solution to the presented challenge is possible.

Disadvantages

• The decision logic can be altered in response to minute changes in the training data.
• Because it is difficult to interpret huge trees, the counter may become spontaneous.
• The computational complexity escalates if the class labels are increased.

Hypothesis

Using Decision Tree Classification and Naïve Bayes Classification, I believe that we can get results that go above chance and have a moderate to high prediction accuracy. Since feature interactions are non-parametric, I believe Decision Tree Classification will exceed Naïve Bayes Classification in prediction. Almost all research publications compare models based on particular factors, and our goal is to attain the same findings. Naïve Bayes Classification may not perform as well as Decision Tree Classification when it comes to predicting diabetes because of the quadratic decision limit.

Methodology

• K-fold cross-validation and holdout validation method on the data set.
• The data is separated into K folds before being tested using the K cross-validation technique. The first K-1 fold is used for validation/testing, whereas the following K-1 folds are utilized for training. The method will be repeated until every single K fold is tested.
• In order to do holdout validation on 768 observations, holdout generates a random, non-stratified partition. Divides observations into two groups: a training set and a holdout group (testing set).
• 80 percent of the data is allotted for training set, while only 20 percent of the data is given for testing set.
• From this we will obtain ROC, accuracy, recall, and other performance metrics and match it with the results of various research papers
• Hyperparameter optimization and feature selection are key to improving models.
• Based on performance measures, determine which models are the most effective.

Parameter selection and Experimental results

Decision Tree Classification

- Optimization parameters such as F-measure, recall, precision are selected and compared to the same parameters used in the paper [1] before they are optimized. It was found that the results of paper [1] were also similar.
- In addition, the decision tree model was optimized minleafsize, which improved the accuracy of the model and reduced the error.

Naïve Bayes Classification

- An F-measure, recall and precision comparison is made before any optimization is carried out in the work [1]. Paper [1]'s results were also determined to be similar.
- Adding to this, the kernel Naïve Bayes enhanced the model's accuracy and reduced error.

Analysis of the results and discussion

Decision Tree Classification

The holdout method was used on a decision tree model in which the data was divided into an 80/20 ratio. 80% of the data was used for training, while the remaining 20% was used for testing purposes. The F-measure for the holdout technique was 0.7451, and the model achieved an accuracy of 83.0065%, which is better than the study [1] and the error computed is 16.9934%. Training and testing accuracy are also excellent, with 95.4471 and 83.0065 percent, respectively. For the K-fold approach, the model employed had an accuracy of 85.5263 percent, compared to the holdout method. Also, the training accuracy and testing accuracy were good, with 95.4471 percent and 83.0065 percent respectively. However, for K-fold, training accuracy was improved to 95.6647 percent and testing accuracy was elevated by 2%. Minleafsize, a parameter that sets the minimum number of leaves in the tree structure, improved the tree structure, boosted accuracy by 3%, and minimized error. ROC curve is shown in [fig e.].

Naïve Bayes Classification

As part of the Naïve Bayes model, the data were split into an 80/20 ratio. 80% of the data was used for training, while the remaining 20% was used for testing.. According to the holdout technique, the F-measure was computed as 0.6600, and the model produced an accuracy of 76.77778% and an error of 22.2222 percent, which is better than the paper [1]. 79.0243 percent and 73.2026 percent are the training and testing accuracy, respectively. The model used for the K-fold approach has an improved accuracy of 81.5789 percent when compared to the holdout method, which was previously used. Training and testing accuracy are also excellent, with 95.4471 and 83.0065 percent respectively. For K-fold, 79.0462 percent of training accuracy and 46.0526 percent of testing accuracy were achieved. The model's accuracy increased to 84.9673% after applying hyperparameter optimization to the kernel parameter.

Naïve Bayes Classification

• Bayesian classification is based on the assumption that all predictors are independent. Using a Naïve Bayes classification technique, it is assumed the presence of a given feature in a class is unrelated to the presence of any additional feature in the class.
• As an example, if a fruit is red in color, round, and roughly 3 inches in diameter, it is an apple. Naïve is a term used to describe a fruit that has all of these traits, even if they depend on each other or on the existence of additional features.
• For huge datasets, the Naïve Bayes model is a great option. Additionally, Naïve Bayes has been proven to surpass even the most advanced classification systems in terms of accuracy.

Advantages

• To be competitive with other classification algorithms such as logistic regression, a Naïve Bayes classification requires less training data if the assumption of independence holds.
• Compared to numerical variables, category input variables perform better (s). The assumption of a normal distribution for a numerical variable is made (bell curve, which is a strong assumption).
• Predicting the class of a test data set is simple and rapid. It's also good at multi-class prediction..

Disadvantages

• Categorical variables that have a category in the test data set but not in the training data set will be given a 0 (zero) probability by the model, and it will be unable to generate any predictions as a result of that. "Zero Frequency" is a common term for this. Using the smoothing technique can help us. Laplace estimation is one of the simplest methods for smoothing data.
• The assumption of independent predictors is another drawback of Naïve Bayes. It is very hard to get a group of predictors that is fully unrelated in real life.

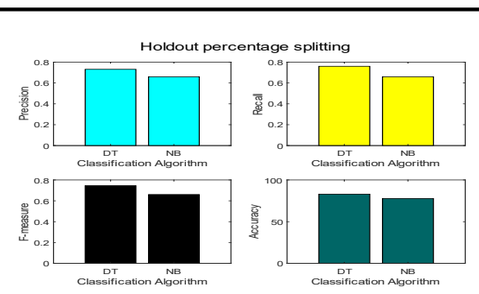

fig c. 3. Holdout percentage splitting

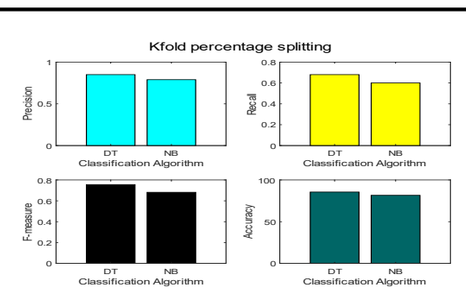

fig d. K-fold percentage splitting

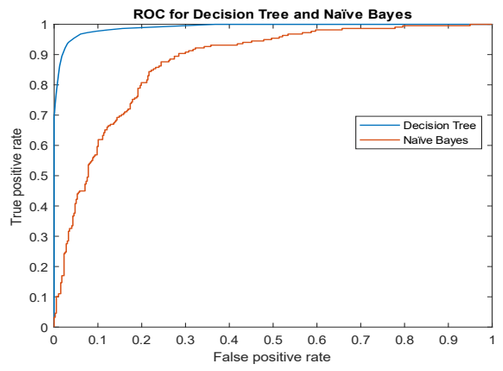

fig e. ROC curve

Model	Precision	Recall	F-measure	Accuracy
DT-holdout	0.7308	0.7600	0.7451	83.0065
NB-holdout	0.6600	0.6600	0.6600	77.7778
DT-K-fold	0.8500	0.6800	0.7556	85.5263
NB-K-fold	0.7894	0.6000	0.6812	81.5789

table b. Performance measures

Lessons Learned

Missing data and outliers are no problem for decision trees since they include characteristics that make them simple to utilize. Overfitting could have been a problem for decision tree but as the data was limited so the decision tree performed well. When data was optimized the Naïve Bayes performed well as compare to decision tree.

Future Directions

• The technique may be refined to automate diabetes analysis using various machine learning methods [3].
• The outcome variable 'Yes' might be further studied as type 1 diabetes, type 2 diabetes, or other diabetes. Diabetes comes in many forms. Type 1 diabetics cannot produce insulin. Type 2 diabetics have trouble generating insulin as well as using it. Type 2 diabetes is the most prevalent and affects people in their forties or older. Gestational diabetes mellitus is a kind of diabetes that occurs in pregnant women. A more precise categorization of the target variable may disclose the medical predictor-outcome link. And additional data could be gathered. Daily food consumption, dietary composition, and physical activity are all linked to diabetes. Regarding data selection, we may increase the age and gender of patients to include children and men [2].

References

1. Khanam, J. ., & Foo, S. Y. (2021, February 20). *A Comparison Of Machine Learning Algorithms for Diabetes Prediction - ScienceDirect.* A comparison of machine learning algorithms for diabetes prediction - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2405959521000205.
2. Huang, R. (2021, March 19). *Prediction Of Pima Indians Diabetes With Machine Learning Algorithms.* Prediction of Pima Indians Diabetes with Machine Learning Algorithms. https://escholarship.org/uc/item/6rh07945.
3. Sisodia, D., & Sisodia, D. (2018, June 8). Prediction Of Diabetes Using Classification Algorithms - ScienceDirect. Prediction of Diabetes using Classification Algorithms. https://www.sciencedirect.com/science/article/pii/S1877050918308548 .
4. Abedini, M., Bijari, A., & Banirostam, T. (2020, July 7). Classification Of Pima Indian Diabetes Dataset Using Ensemble Of Decision Tree, Logistic Regression And Neural Network. https://ijarcce.com/wpcontent/uploads/2020/07/IJARCCE.2020.9701.pdf.
5. Ahuja, R., Sharma, S. C., & Ali, M. (2019, July 14). *(PDF) A Diabetic Disease Prediction Model Based On Classification Algorithms | Annals Of Emerging Technologies In Computing (AETiC) - Academia.edu.* (PDF) A Diabetic Disease Prediction Model Based on Classification Algorithms | Annals of Emerging Technologies in Computing (AETiC) - Academia.edu. https://www.academia.edu/39729407/A_Diabetic_Disease_Prediction_Model_Based_on_Classification_Algorithms.