A comparison between Decision Tree Classifier & Support Vector Machine (SVM) Classifier on predicting stroke tendencies among patients

Description and Motivation

We will solve the binary classification problem of predicting if a patient had a stroke or not, based on the information available from the dataset from kaggle, using Decision Tree classifier and SVM classifier. We aim to compare the results from the two machine learning models and use performance metrics to assess model performance.

Exploratory Data Analysis

- Dataset- Stroke Prediction Dataset from Kaggle.
- This dataset provides relavant information about the patients, which can be used to predict whether a patient is likely to get stroke based on the independent variables like gender, age, various diseases, and smoking status.
- The original dataset consists of 5110 rows and 12 columns. There are 11 predictors and 1 target where the latter has only two values, 1: patient had a stroke; 0: patient did not have a stroke
- 201 missing values of 'bmi', filled with the mean value.
- Table 1 shows information about the numeric variables with max, min, mean and standard deviation.
- Figure 1 shows there is a positive trend of patients who have had stroke according to there age. It shows the density of people above the age of 50 have suffered stroke more.
- Only 0.5% of patients under 45 have had a stroke.
- Figure 2 shows the density of people having glucose level less than 150 have suffered stroke more.
- Figure 3 shows the correlation between various attributes.
- Figure 4 shows that most of the population consists of Private workers, and that maximum proportion of the stroke patients have Private as their work type.
- Figure 5 & 6 suggest that patients with Hypertension and Heart Disease are more likely to have a stroke.

	mean	std	min	max
age	43.226614	22.612647	0.08	82.00
hypertension	0.097456	0.296607	0.00	1.00
heart_disease	0.054012	0.226063	0.00	1.00
avg_glucose_level	106.147677	45.283560	55.12	271.74
bmi	28.893237	7.854067	10.30	97.60
stroke	0.048728	0.215320	0.00	1.00

Table 1

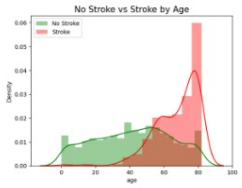


Figure 1

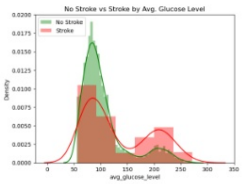


Figure 2

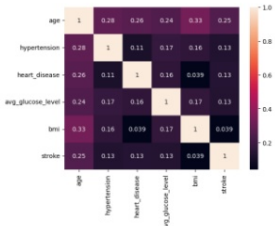


Figure 3

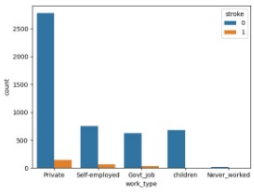
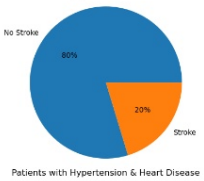
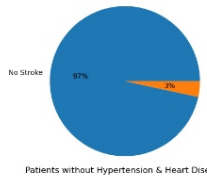


Figure 4



Patients with Hypertension & Heart Disease

Figure 5



Patients without Hypertension & Heart Disease

Figure 6

Decision Tree

A decision tree in machine learning is a versatile, interpretable algorithm used for predictive modelling. It structures decisions based on input data, making it suitable for both classification and regression tasks. Decision tree is derived from the independent variables, with each node having a condition over a feature. The nodes decides which node to navigate next based on the condition. Once the leaf node is reached, an output is predicted. The right sequence of conditions makes the tree efficient.

Advantages

- No preprocessing needed on data- Minimal data preparation is required for Decision Trees. Since the training procedure in CART deals with each input feature independently, at each node in the tree, data scaling and normalisation are not required. In addition, it is also possible to implement Decision Trees that can handle missing values and categorical features
- No assumptions on distribution of data.
- Handles colinearity efficiently.
- Decision trees can provide understandable explanation over the prediction.

Disdvantages

- Chances for overfitting the model if we keep on building the tree to achieve high purity. decision tree pruning can be used to solve this issue.
- Prone to outliers.
- Tree may grow to be very complex while training complicated datasets.
- Looses valuable information while handling continuous variables.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification and regression tasks. The main idea behind SVM is to find the best boundary (or hyperplane) that separates the data into different classes. In the case of classification, a SVM algorithm finds the best boundary that separates the data into different classes. The boundary is chosen in such a way that it maximizes the margin, which is the distance between the boundary and the closest data points from each class. These closest data points are called support vectors.

Advantages

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient
- SVM models have generalization in practice, the risk of over-fitting is less in SVM.

Disdvantages

- SVM algorithm is not suitable for large data sets.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
- Long training time for large datasets.

Hypothesis Statement

- From some literature reviews, the accuracy between the Desision Tree and SVM for categorical data, the former is likely to perform moderately better.
- The SVM seems to spend longer training time than the Decision Tree.
- SVM uses kernel trick to solve non-linear problems whereas decision trees derive hyper-rectangles in input space to solve the problem.Decision trees are better for categorical data and it deals colinearity better than SVM. (Danny Varghese, Towards Data Science, Dec 8, 2018)

Methodology

- Fill the missing data with the respective variable's moving mean value with a window of 10
- Split data into a 75:25 split for train and test data. using cvpartition function. The test data remains unseen to models until end.
- Fit the Decision tree model and SVM Classifier model with the training data.
- Evaluate which are optimal models based on performance metrics.
- Predict the unseen test set set.
- Measure and compare predictive performance of the optimized models.
- To find a good fit, meaning one with optimal hyperparameters that minimize the cross-validation loss, I have used Bayesian optimization. A list of hyperparameters to optimize by using the OptimizeHyperparameters name-value argument and optimization options by using the HyperparameterOptimizationOptions name-value argument have been specified.'OptimizeHyperparameters' as 'auto'. The 'auto' option includes a typical set of hyperparameters to optimize.
- fitcsvm finds optimal values of BoxConstraint, KernelScale, and Standardize. Set the hyperparameter optimization options to use the cross-validation partition c and to chosen the 'expected-improvement-plus' acquisition function for reproducibility.
- fitctree finds optimal values of MinLeafSize.
- The results for the SVM model and Decision Tree model are as follows-

SVM

```
Best estimated feasible point (according to models):
    BoxConstraint    KernelScale    Standardize
    _____________    ___________    ___________

       0.0013328        0.021955          true

Estimated objective function value = 0.050688
Estimated function evaluation time = 0.41928
```

Decision Tree

```
Best estimated feasible point (according to models):
    MinLeafSize
    ___________

        64

Estimated objective function value = 0.050092
Estimated function evaluation time = 0.050047
```

Analysis of Results

- Model Comparison

DT	Score	SVM
97.460000	Training Accuracy	95.040000
93.730000	Testing Accuracy	95.370000
0.240000	Precision	0.130000
0.2	Recall	0.92
0.220000	f1 Score	0.040000
0.570000	AUC	0.500000

Table 2

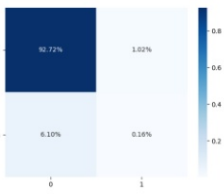


Figure 7

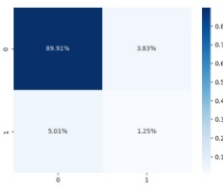


Figure 8

- Table 2 shows the comparison of the two models, Decision Tree & Support Vector Machine with various score metrics, namely Precision,Recall,f1,AUC.
- From the scores, it can be inferred that both the models's performances are quite accurate, although the Decision Tree Classifier does slightly better.
- Figure 7 shows the confusion matrix for Decision Tree Classifier, which represents the classification accuracy with the percentage values of each case of TP,TN,FP,FN.
- Figure 8 shows the confusion matrix for SVM Classifier.
- From both the confusion matrices, it's clear that true negeative sections for both the models are leading with the highest percentages, while the true positive is quite low, which is because the dataset itself has a very less proportion of true positives, i.e. patients who have had stroke.
- Regardless, since the False positives & negatives for both the models are quite low in proportion compared to the true positives & negatives, it can be concluded that both the models perform reasonably good.
- Before setting the optimized hyperparameters, the accuracy score of our SVM model was 94.51% . After setting the hyperparameters to the values resulted from the hyperparameter optimization, the accuracy score increased to 95.22%.
- Before setting the optimized hyperparameter MinLeafSize, the accuracy score of our DT model was 91.62%.
  After setting the hyperparameter to the value resulted from the hyperparameter optimization, the accuracy score increased to 94.44%.
- Francis Sahngun Nahm had mentioned in his research paper on "Receiver operating characteristic curve: overview and practical use for clinicians", about interpreting ROC curves for comparing model performances. As he states, " The AUC is widely used to measure the accuracy of diagnostic tests. The closer the ROC curve is to the upper left corner of the graph, the higher the accuracy of the test because in the upper left corner, the sensitivity = 1 and the false positive rate = 0 (specificity = 1). The ideal ROC curve thus has an AUC = 1.0."
- Figure 9 shows the ROC curves of both the models. Since the ROC curve of SVM model is more closer to the upper left corner of the graph, the SVM model has higher accuracy of predicting on the test data.

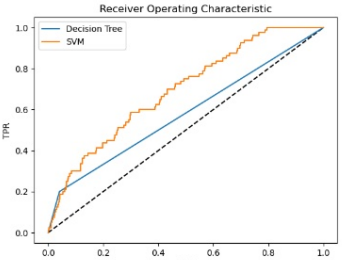


Figure 9

References

- Machine Learning: Decision Tree Learning- Bao Tram Duong (https://baotramduong.medium.com/machine-learning-decision-tree-learning-7ef7a0134112)
- Comparative Study on Classic Machine learning Algorithms, Danny Varghese (https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222)
- Top 4 advantages and disadvantages of Support Vector Machine or SVM, Medium Article Jun 13, 2019
- A Complete View of Decision Trees and SVM in Machine Learning, Hailey Huong Nguyen,
- Towards Data Science Article,Jan 8 2019, RESEARCH ARTICLE 'A Comparison of Support Vector Machine and Decision Tree Classifications Using Satellite Data' by Langkawi Island H.Z.M. Shafri and F.S.H. Ramle (https://scialert.net/fulltext/?doi=itj.2009.64.70)
- Receiver operating characteristic curve: overview and practical use for clinicians- Francis Sahngun Nahm https://doi.org/10.4097/kja.21209