CS339: Artificial Intelligence
Spring 2021

Stroke Prediction Classification Accuracy using Decision Trees

By Amna Khalid

1. ABSTRACT

This paper will use a stroke prediction dataset to research how accurately a decision tree classifies each observation in the dataset. This is done by first performing an exploratory data analysis on the stroke prediction dataset. The target attribute (i.e. 'stroke') is removed from the dataset to see the actual accuracy of the decision tree classifier. Accordingly, the dataset is split into five training and testing sets. The accuracy for classifying the observations in each split set is measured and recorded using the F1 score. Results show that Decision Trees predict an occurrence of a stroke with an average of 91.4% accuracy.

2. INTRODUCTION

2.1 Supervised Machine Learning

Supervised machine learning is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into a model, it adjusts its weights through a reinforcement learning process, which ensures that the model has been fitted appropriately. Supervised learning helps organizations solve a range of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs as wells as correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through *the loss function*, adjusting until the error has been sufficiently minimized. Supervised learning can be separated into two types of problems when data mining:

- Classification: uses an algorithm to accurately assign test data into specific categories.
- Regression: used to understand the relationship between dependent and independent variables.

2.2 Decision Trees

Decision Trees are a supervised learning technique that predicts values of each entry in the data by learning decision rules derived from features. They can be used in both regression and a classification context. These particular models work by partitioning the feature space into a number of simple rectangular regions, divided up by axis-parallel splits. In order to obtain a prediction for a particular observation, the mean or mode of the training observations' responses within the partition that the new observation belongs to, is used.

2.3 Model in Machine Learning

An *algorithm* in machine learning is a procedure that is run on data to create a machine learning "model." Machine learning algorithms perform *pattern recognition* to learn from the data or can be "fit" onto a particular dataset. Similarly, a *model* in machine learning is the output of a machine learning algorithm run on data. A model represents what was learned by a machine learning algorithm. In the case of the decision tree algorithm, a model is comprised of a tree of if-then statements with specific values.

For decision trees, the model f(x) is given by:

$$f(x) = \sum_{m=1}^{M} w_m \phi(x; v_m)$$

Where w_m is the mean response in a particular region, and R_m, and v_m represents how each variable is split at a particular threshold value. These splits define how the feature space in R^p is split into M separate regions.

Thus, given p features, the decision tree partitions the p-dimensional feature space into M mutually distinct regions (R_1,..., R_m) that fully cover the subset of the space and do not overlap. Any new observation that falls into a particular partition R_m is the mean of all training observations within the partition and is denoted by w_m.

As with all classification regimes, this research will be predicting a categorical, rather than continuous, response value. In order to actually make a prediction for a categorical class, we have to instead use the mode of the training region to which an observation belongs, rather than the mean value. Thus, to split the trees, the Gini Index is used.

2.4 Gini Index

The Gini Index is an error metric that is designed to show how "pure" a region is. "Purity" in this case means how much of the training data in a particular region belongs to a single class. If a region R_m contains data that is mostly from a single feature in the dataset, then the Gini Index value will be small.

$$G = \sum_{c=1}^{C} p(i)\,(1 - p(i))$$

Where *c* is the particular attribute in the dataset and *p(i)* is the probability of picking the observation is that particular attribute.

This research will thus be using a stroke prediction dataset to predict how likely one is to get a stroke based on the input parameters like gender, age, etc. However, the goal of this research is to understand how accurately a decision tree classifies each entry in the stroke prediction dataset.

3. METHODOLOGY

To first understand the dataset, exploratory data analysis is performed on our Stroke Prediction Dataset. Further, after understanding the data, necessary adjustments are made so that the dataset can now be utilized for the *DecisionTreeClassifier()*.

3.1 Cross-validation

The stratified K-Folds cross-validator is used as a cross-validation object for the dataset. It is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class.

To test the accuracy of the decision trees as a classifier, the dataset is first split into training and testing subsets. To do this, *StratifiedKFold()* from the Scikit-learn machine learning library is used. The data is split into five individual training and testing datasets.

3.2 Missing values

To account for any missing values without losing data, *KNNImputer()* from the Scikit-learn machine learning library is used. Each sample's missing values are imputed using the mean value from *n_neighbors nearest neighbors* found in each training set. Two samples are close if the features that neither is missing are close.

3.3 Classification Accuracy

To test the accuracy of the decision tree as a classifier for each entry in the stroke prediction dataset, the F1 score for each split of the dataset is recorded.

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula that the F1 score uses is:

$$F1 \;=\; \frac{2\ x\ (precision * recall)}{(precision * recall)}$$

In the multi-class and multi-label case, this is the average of the F1 score of each class with the weight depending on the average parameter.

4. RESULTS

4.1 Data Analysis

According to the World Health Organization (WHO) stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient. The data contains 5110 observations with 12 attributes. The attributes and their descriptions are given in the below.

Attribute	**Description**
ID	Unique identifier
Gender	"Male", "Female" or "Other"
Age	Age of the patient
Hypertension	0: the patient does not have hypertension 1: the patient has hypertension
Heart_disease	0: the patient does not have any heart diseases 1: the patient has a heart disease
Ever_married	"Yes" or "No"
Work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

Residence_type	“Rural" or "Urban"
Avg_glucose_level	The average glucose level in the patient’s blood
BMI	Body Mass Index
Smoking_status	"Formerly smoked", "never smoked", "smokes" or "Unknown"
Stroke	0: the patient did not have a stroke 1: the patient had a stroke

*Note: "Unknown" in Smoking_status means that the information is unavailable for this patient.

Fig. 1.1

Upon further analysis of gender in relation to stroke predictability, the data shows that females are more likely to have a stroke as shown by Fig. 2.1. Results also shows that a patients smoking habits do not have any corelation to causing a stroke as shown by Fig. 2.2. Additionally, Fig. 2.3 shows how a patients residence type does not correlate to causing a stroke in a any significant way.

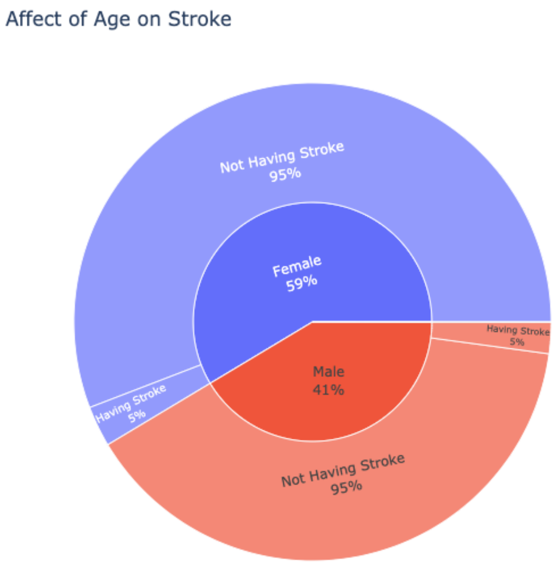


Fig. 2.1

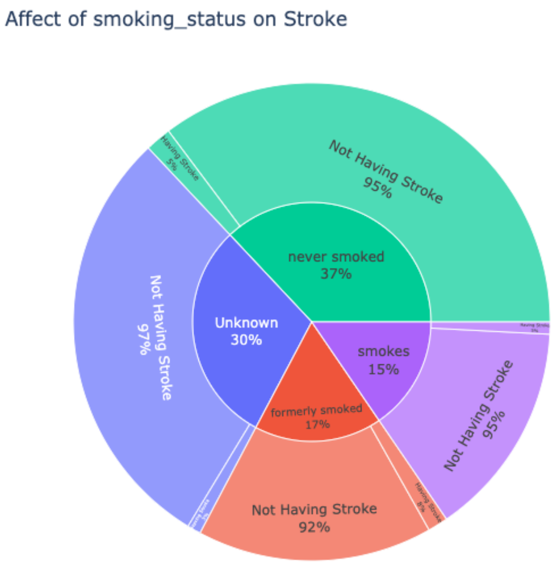


Fig. 2.2

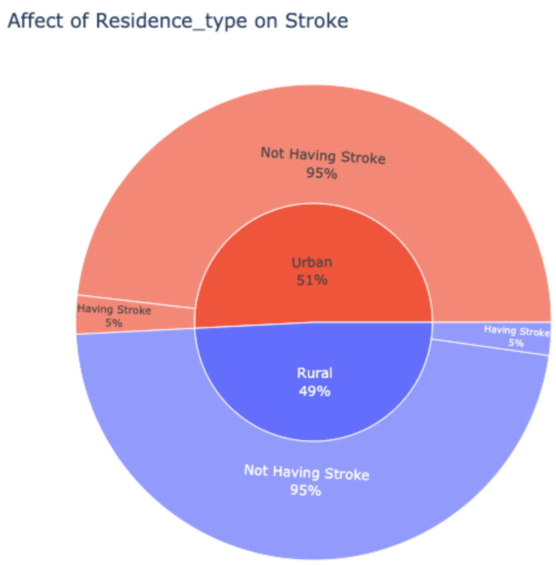
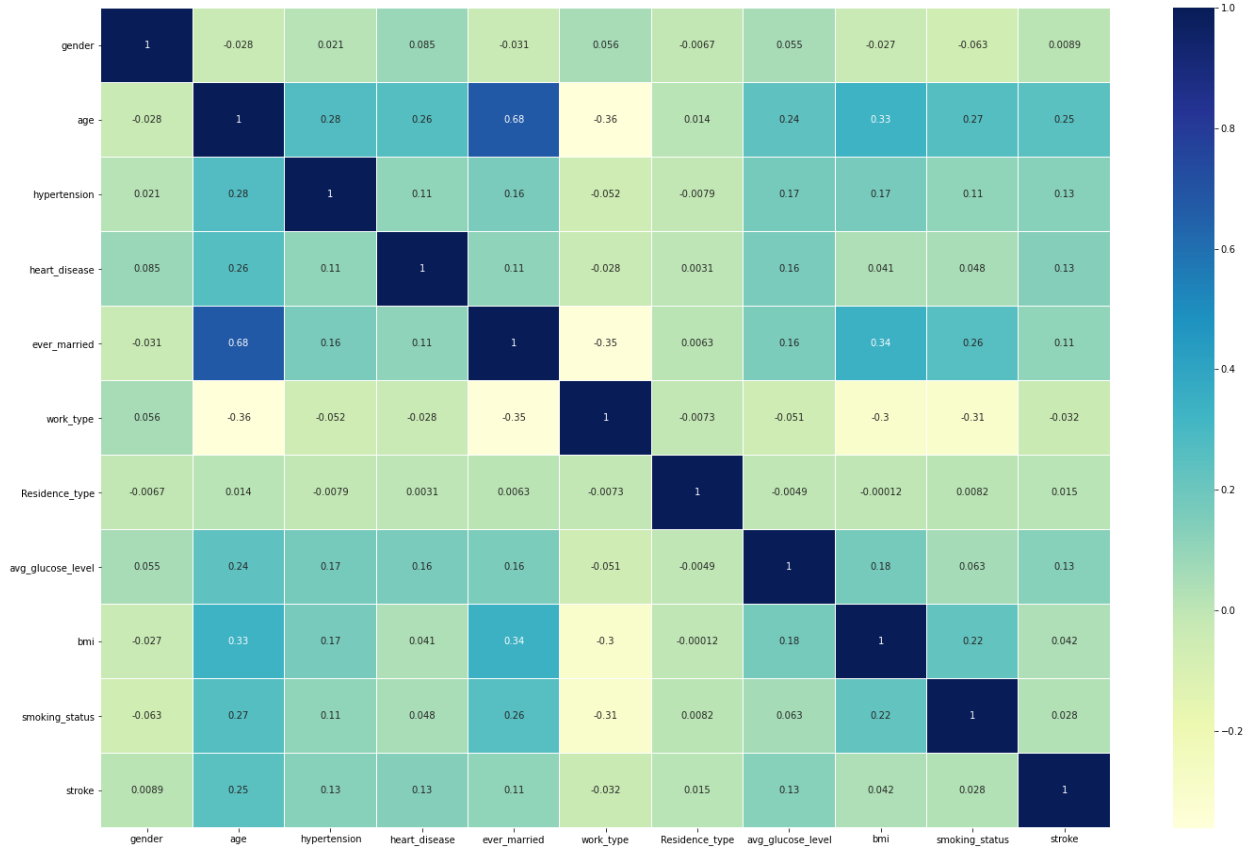


Fig. 2.3

To see the correlation coefficients between each attribute in the dataset, a correlation matrix is generated. Each cell in the table shows the correlation between two the two attributes.



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
gender	1	-0.028	0.021	0.085	-0.031	0.056	-0.0067	0.055	-0.027	-0.063	0.0089
age	-0.028	1	0.28	0.26	0.68	-0.36	0.014	0.24	0.33	0.27	0.25
hypertension	0.021	0.28	1	0.11	0.16	-0.052	-0.0079	0.17	0.17	0.11	0.13
heart_disease	0.085	0.26	0.11	1	0.11	-0.028	0.0031	0.16	0.041	0.048	0.13
ever_married	-0.031	0.68	0.16	0.11	1	-0.35	0.0063	0.16	0.34	0.26	0.11
work_type	0.056	-0.36	-0.052	-0.028	-0.35	1	-0.0073	-0.051	-0.3	-0.31	-0.032
Residence_type	-0.0067	0.014	-0.0079	0.0031	0.0063	-0.0073	1	-0.0049	-0.00012	0.0082	0.015
avg_glucose_level	0.055	0.24	0.17	0.16	0.16	-0.051	-0.0049	1	0.18	0.063	0.13
bmi	-0.027	0.33	0.17	0.041	0.34	-0.3	-0.00012	0.18	1	0.22	0.042
smoking_status	-0.063	0.27	0.11	0.048	0.26	-0.31	0.0082	0.063	0.22	1	0.028
stroke	0.0089	0.25	0.13	0.13	0.11	-0.032	0.015	0.13	0.042	0.028	1

Fig. 2.4

4.2 Decision Tree Accuracy

The target attribute (which in this case would be ‘stroke’) is first removed from the dataset to be able to see the accuracy of the decision tree classification. Further, after spliting the dataset using *StratifiedKFold()*, the F1 scores for each split is recorded. The results show that the Decision Tree classifies each split with an average of 91.4% accuracy. The scores for each split is given in Fig. 3.1.

N_split	**Respective F1 Score**
0	0.9097980262363825
1	0.9169055196452457
2	0.9113864011567206
3	0.9109751784206975
4	0.9203554554397944

Fig. 3.1

5. CONCLUSION

The research shows that Decision Trees are able to predict strokes, give the stroke prediction dataset, with above 90% accuracy. We find that the decision tree model can handle categorical and continuous features in the same data set. Further, the construction method for DT model automatically selects attributes, rather than having to use subset selection. Thus, the models are able to scale effectively on large datasets.

While DT models suffer from poor prediction performance in comparison to other machine learning algorithms, they are extremely competitive when utilised in an ensemble setting, via bootstrap aggregation ("bagging"), random forests or boosting.

In the future, the accuracy as well performance comparison of bagging, boosting, and random forests should be researched. This will further demonstrate the effectiveness of these different supervised learning algorithms in comparison to the other machine learning algorithms present nowadays.

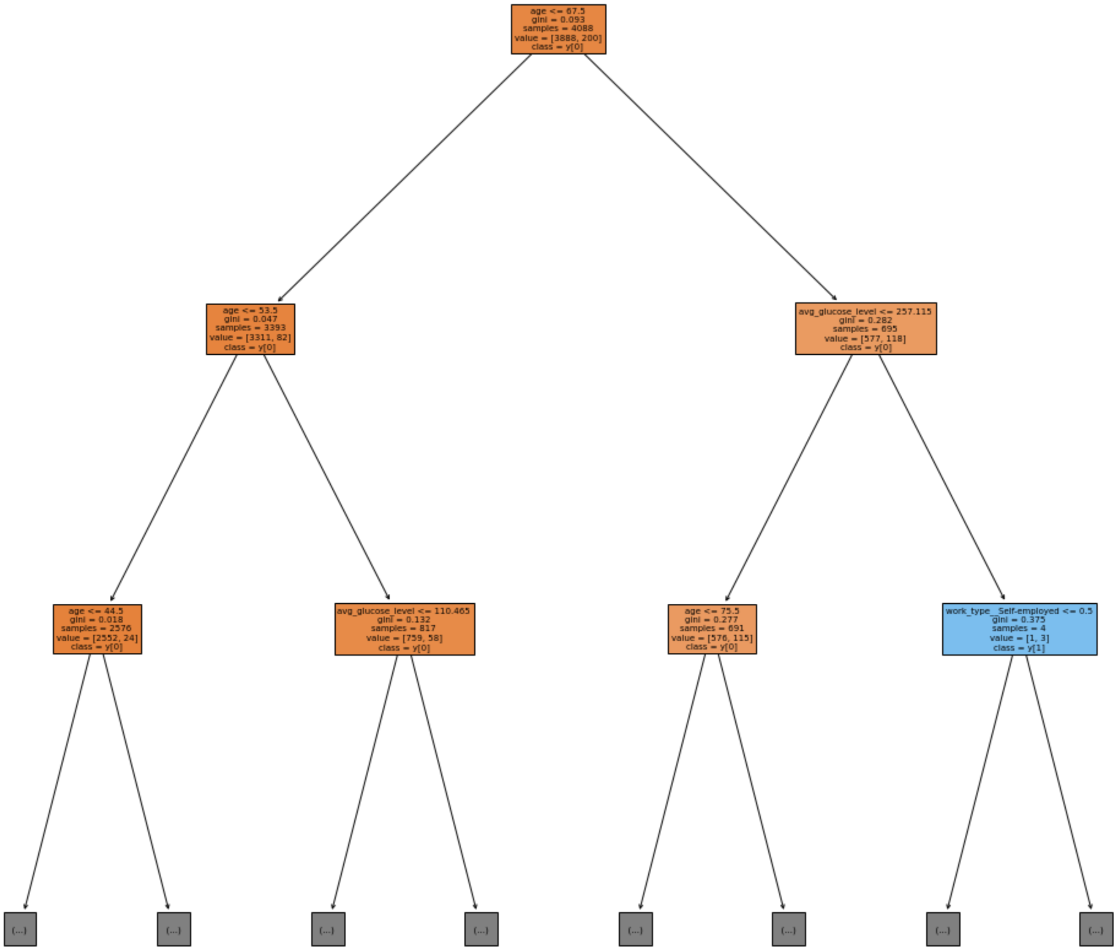


Fig. X: head of the decision tree

6. BIBLIOGRAPY

Fedesoriano. (2021, January 26). Stroke Prediction Dataset. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv.

Ivanmsiegfried. (2021, April 22). *Kaggle*. Clinical Features Model to Predict Stroke.