24.05.2020

Mammographic Mass Dataset Classification ()

Student: Tahani Fennir

The work in this small report deals with classifying Mammographic masses as benign or malignant (binary classification problem) using different Machine Learning algorithms, including SVM, Logistic Regression, Decision Trees, Random Forest, Naive Bayes and Artificial Neural Network. ROC curves are plotted for each to identify the best classification algorithm for the problem.

Dataset

I choosed the dataset "Mammographic masses" which is a public dataset from UCI repository (*https* : *//archive.ics.uci.edu/ml/datasets/Mammographic+Mass*), to build a classification model that Predicting whether a tumour is benign or malignant from BI-RADS attributes and the patient's age. Attribute Information:

	Age	Shape	Margin	Density	Severity
0	67.0	3.0	5.0	3.0	1
1	43.0	1.0	1.0	NaN	1
2	58.0	4.0	5.0	3.0	1
3	28.0	1.0	1.0	3.0	0
4	74.0	1.0	5.0	NaN	1

Figura 1: top 5 rows of the dataset

In figure 1 you notice that there are some missing values in the dataset. Dealing with missing data is something very important in data preprocessing. I dropped the null values from the data.

Decision Tree Classifier :

It is a Supervised Machine Learning algorithm, where the data is continuously split according to a certain parameter with output like "benign.ºr "malignant".

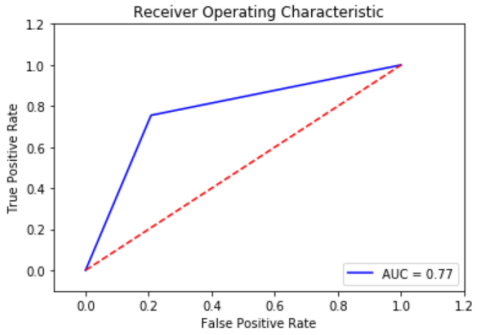


Figura 2: ROC Curve Plot

The accuracy performance of decision tree classifier is 0.77 as shown in ROC Curve Plot above (Figure 2).

Random forests Classifier:

it is a supervised, flexible and easy learning algorithm. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.
By implementing the random forests algorithm the accuracy of performance is 0.76 as show in figure 3.

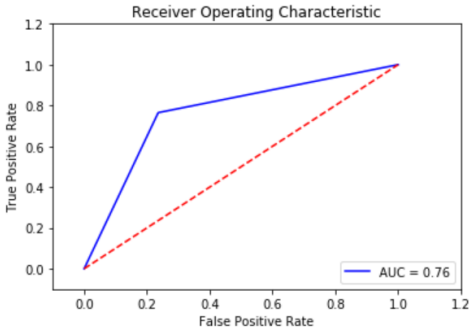


Figura 3: ROC Curve Plot

SVM (Support Vector Machine):
It is a supervised machine learning algorithm that can be used for both classification and regression purposes, but it more generally used in classification situations.SVM is based on the idea of finding a hyperplane(a line separating and classifying linearly a set of data) that best divides a dataset into two classes.
I build a classifier using SVC class and kernel methods.

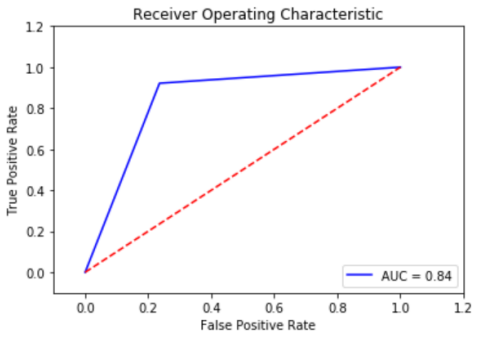


Figura 4: LINEAR KERNEL

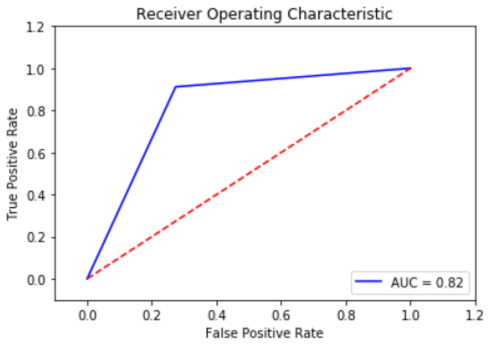


Figura 5: POLY KERNEL

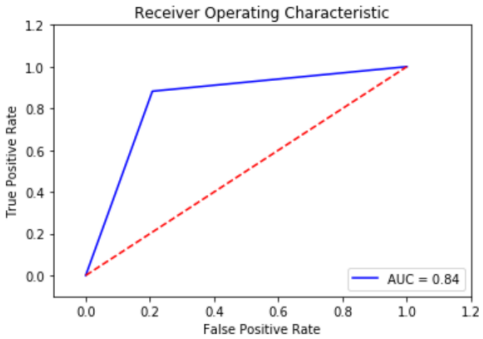


Figura 6: RBF KERNEL

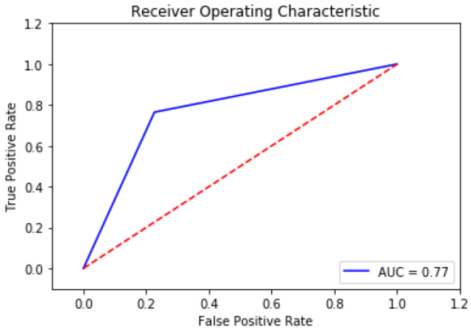


Figura 7: SIGMOID KERNEL

In machine learning, kernel methods are a class of algorithms for pattern analysis, whose best known member is the support vector machine (SVM). The task of pattern analysis is to find and study general types of relations (for example classifications) in datasets. Polynomial kernel, Radial basis function kernel(RBF), Sigmoid kernel and linear kernel are a kernel function, their accuracies of performance are presented in the plots above, 0.84 for linear kernel, 0.82 for poly kernel, 0.84 for RBF kernel and 0.77 for

sigmoid kernel.

Naïve Bayes:
Naïve Bayes classifiers are a family of simple "probabilistic classifiers"(is a classifier that is able to predict, given an observation of an input) based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. The accuracy of performance I got by using naive bayes is 0.79 as shown in the plot below.

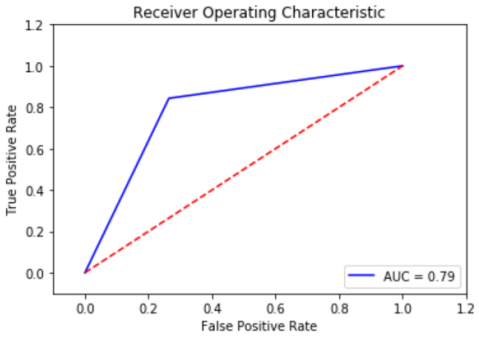
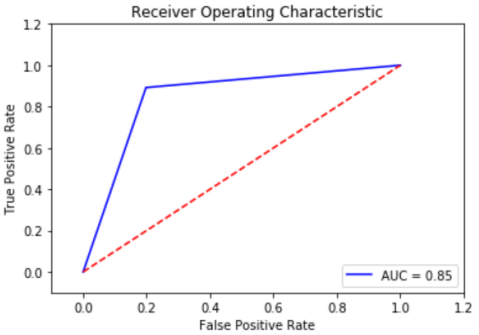
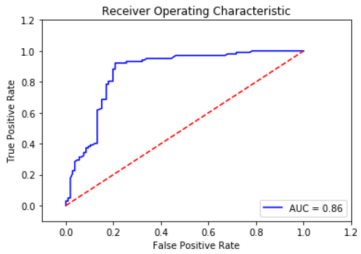


Figura 8: ROC Curve Plot

Logistic regression:
Logistic regression is a supervised classification learning algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, meaning there would only be two classes. The performance accuracy I got is 0.85 which is the best accuracy among the previous algorithms.



Artificial Neural Network:
I used **Keras** library to build the model, it is a powerful and free open source Python library for developing and evaluating deep learning models. Keras allows to define and train neural network models in just a few lines of code. ANN performs the best in this problem with AUC of 0.86 as shown below:



Conclusion:
Artificial Neural Network performs the best in this problem with accuracy of 0.86, and next is the Logistic Regression with accuracy of 0.85 .