Department of Computer Science
Technical University of Cluj-Napoca

Intelligent Systems

Exam activity 2019-2020

Name:Trif Gheorghe Andrei
Group:30235
Email:trifandrei@yahoo.com

Assoc. Prof. dr. eng. Adrian Groza
Adrian.Groza@cs.utcluj.ro

Contents

Chapter 1

Neuronal Networks

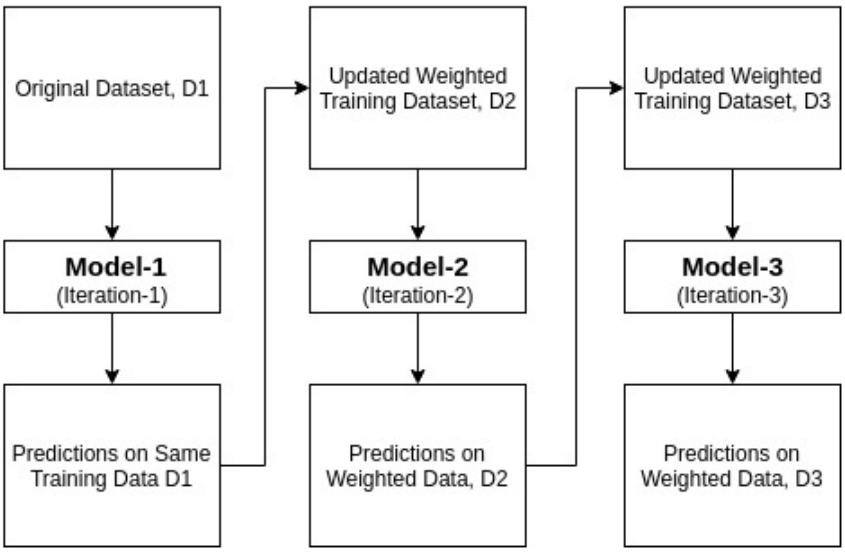
1.1 Problem definiton

This project is try to implement and compare different type of classification algorithm.I managed to implement the following algorithms: Ada Boost,Decision Tree, Lazy IBK(KNN), Logistics Regression, Multiclass Classifier(RidgeClassifier) ,Multilayer–Perceptron, Naïve Bayes and Random Forest.I use Breast Cancer Wisconsin data set from the UCI repository and all algorithm are analyzed on this data set.

1.2 1.Ada Boost

Ada-boost or Adaptive Boosting is an ensemble boosting classifier.AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.It works in the following steps:

Initially, Adaboost selects a training subset randomly.It iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training.It assigns the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification.Also, It assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.This process iterate until the complete training data fits without any error or until reached to the specified maximum number of estimators.To classify, perform a "vote" across all of the learning algorithms you built.



1.3 Decision Tree

A Decision Tree Algorithm is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome.The basic idea behind any decision tree algorithm is as follows:

1.Select the best attribute using Attribute Selection Measures(ASM) to split the records.

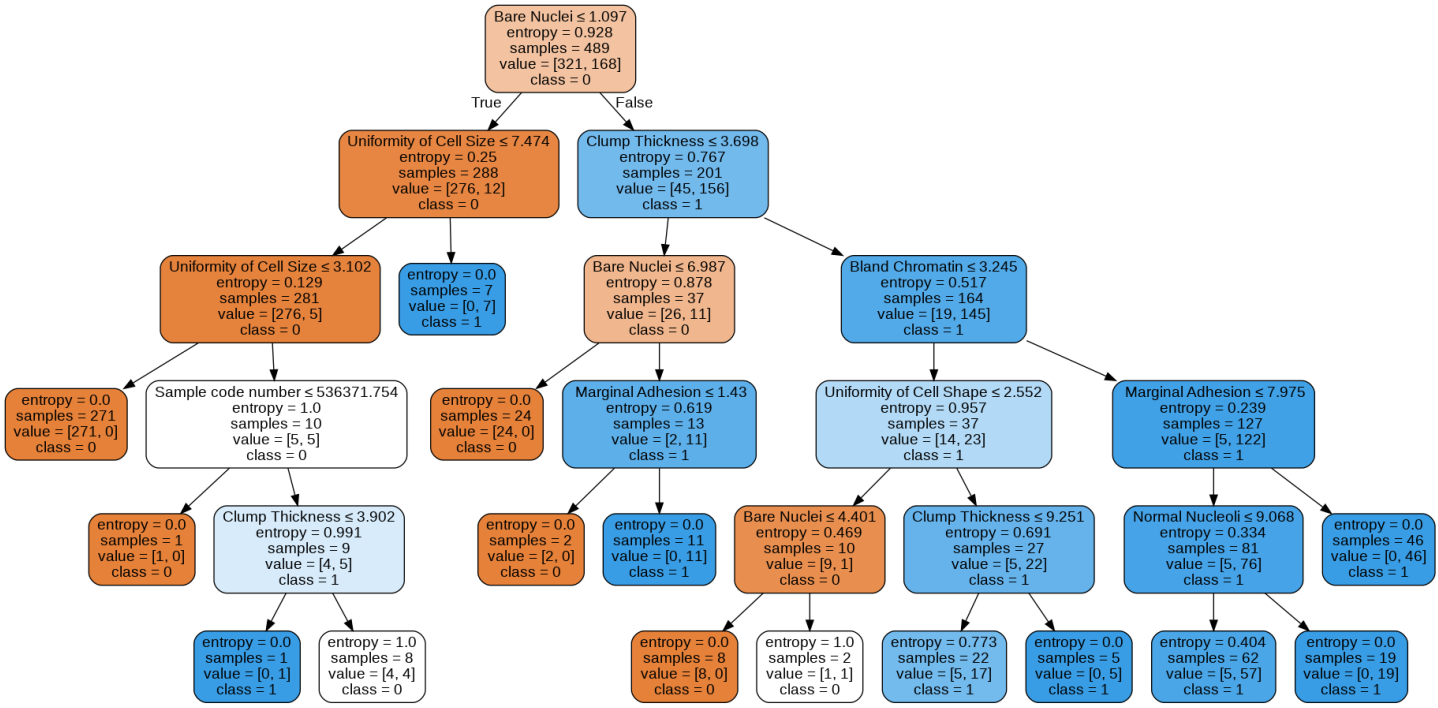
2.Make that attribute a decision node and breaks the dataset into smaller subsets.

3.Starts tree building by repeating this process recursively for each child until one of the condition will match:all the tuples belong to the same attribute value or there are no more remaining attributes or there are no more instances.

For this project i run the breast-cancer-wisconsin dataset and the algorithm return me for am max-depth of 5 the next flowchart:

Some advantages :

Decision trees are easy to interpret and visualize.It can easily capture Non-linear patterns.It requires fewer data preprocessing from the user, for example, there is no need to normalize columns.It can be used for feature engineering such as predicting missing values, suitable for variable selection.



1.4 Naive Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle.Here i implement the Multinomial Naive Bayes Classifier.For the Multinomial Naive Bayes Classifier the distribution formula is:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Advantages of Naive Bayes

If the independence assumption holds then it works more efficiently than other algorithms.

It requires less training data.

It is highly scalable.

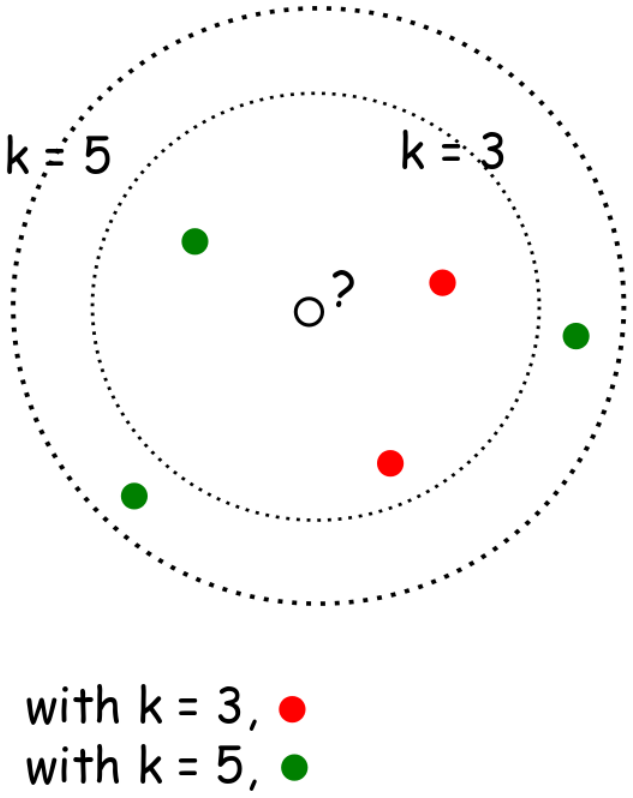
It can make probabilistic predictions.

1.5 K-Nearest Neighbor(Lazy IBK)

KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution.Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. This makes training faster and testing phase slower and costlier.

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm.

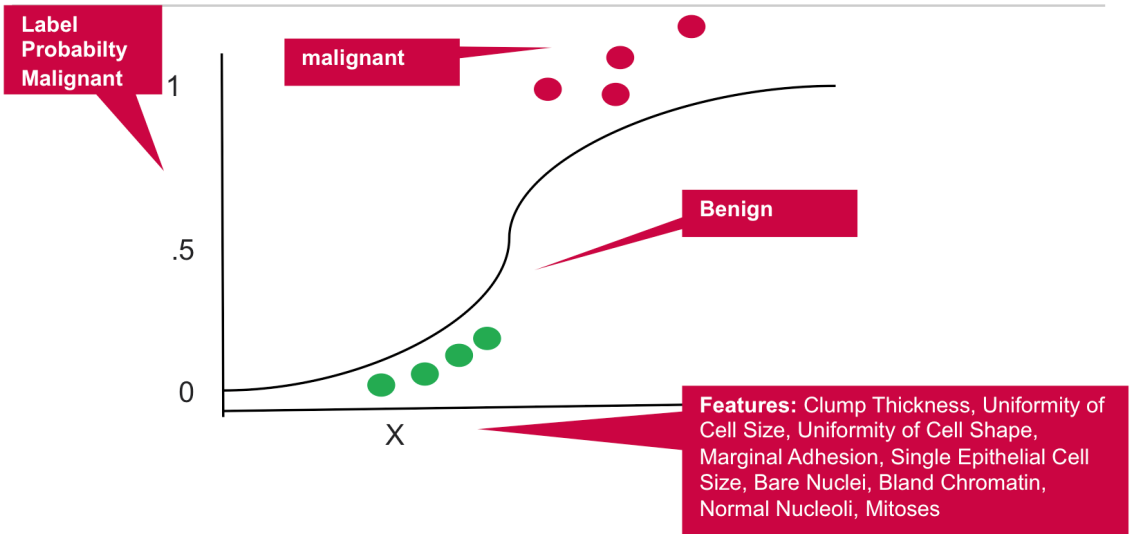
KNN has the following basic steps: calculate distance, find closest neighbors, vote for labels.



Some advantages: The training phase of K-nearest neighbor classification is much faster compared to other classification algorithms. There is no need to train a model for generalization.That is why KNN is known as the simple and instance-based learning algorithm.

1.6 Logistic Regresion

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes.Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable. This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.Logistic Regression is one of the most popular ways to fit models for categorical data, especially for binary response data in Data Modeling.Logistic Regression is used when the dependent variable (target) is categorical.In our case M = malignant, B = benign.



Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

Difference between Linear Regression and Logistic Regression:

Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.

1.7 Random Forests Classifiers

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

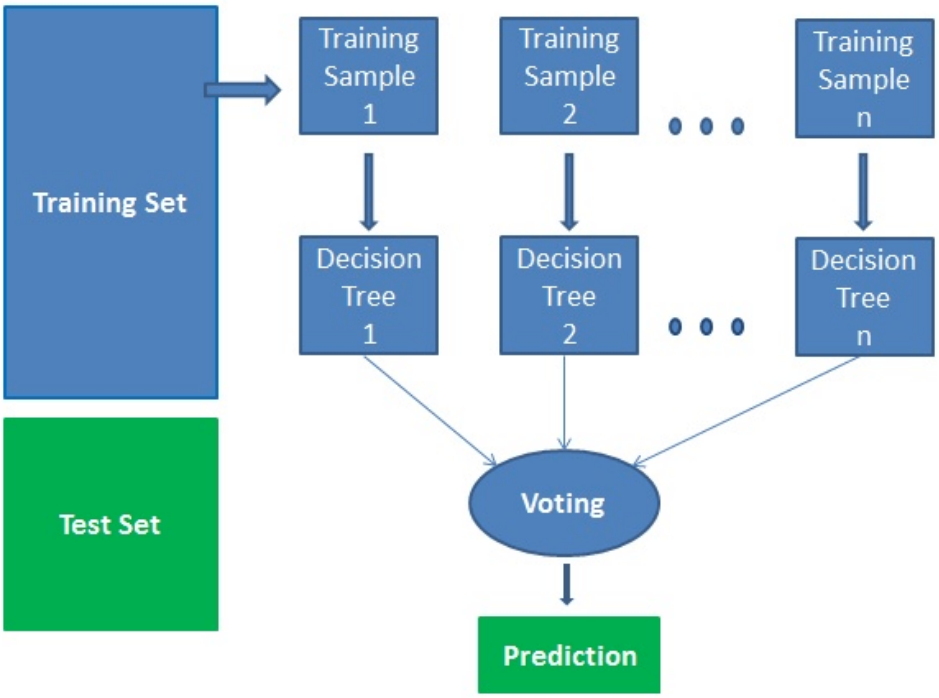
It works in four steps:

1.Select random samples from a given dataset.

2.Construct a decision tree for each sample and get a prediction result from each decision tree.

3.Perform a vote for each predicted result.

4.Select the prediction result with the most votes as the final prediction.



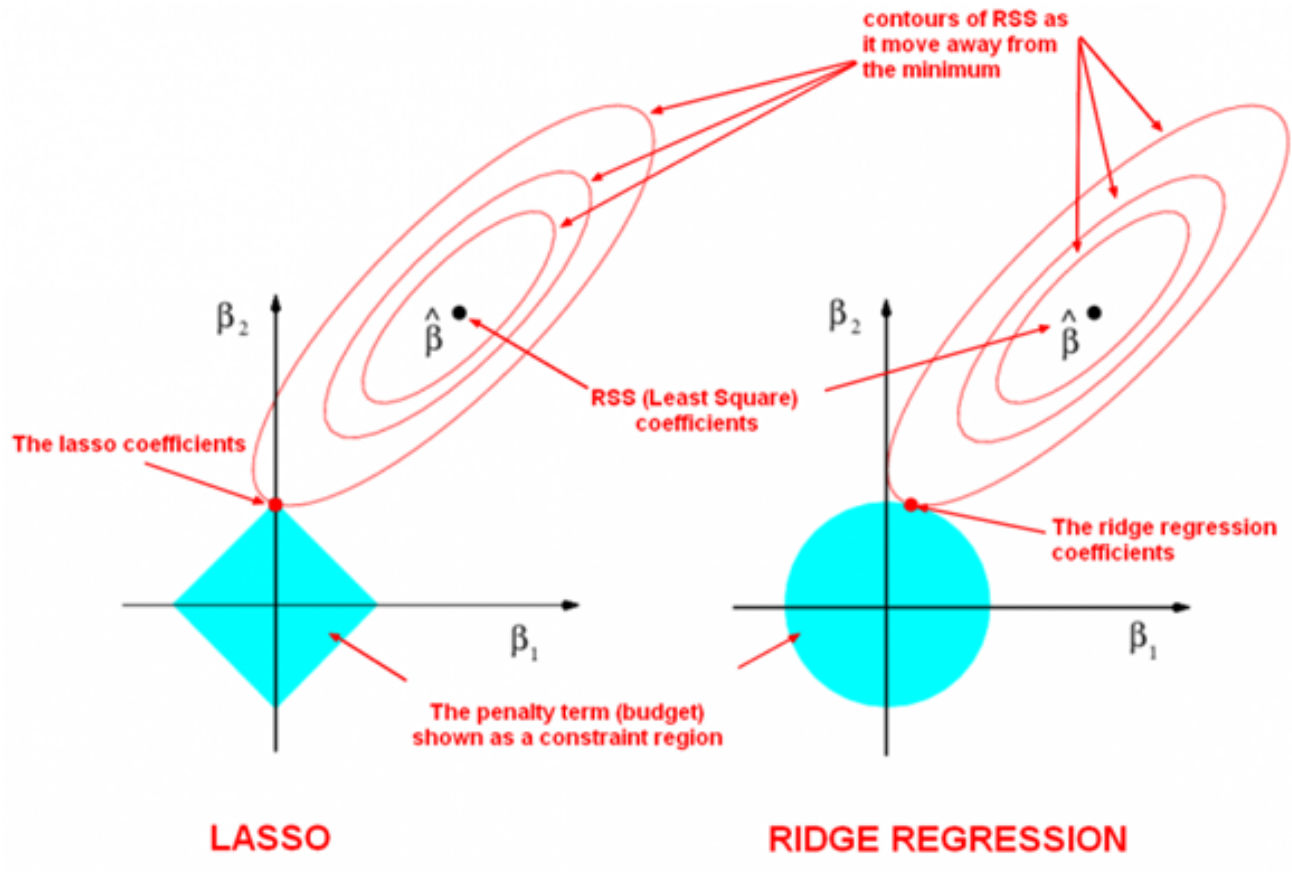
Advantages: Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases. The algorithm can be used in both classification and regression problems.Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.

1.8 Ridge Classifier

Ridge regression or Tikhonov regularization is the regularization technique that performs L2 regularization. It modifies the loss function by adding the penalty equivalent to the square of the magnitude of coefficients.

$$\sum_{j=1}^{m}\left(Y_i - W_0 - \sum_{i=1}^{n} W_i X_{ji}\right)^2 + \alpha \sum_{i=1}^{n} W_i^2 = loss\_function + \alpha \sum_{i=1}^{n} W_i^2$$

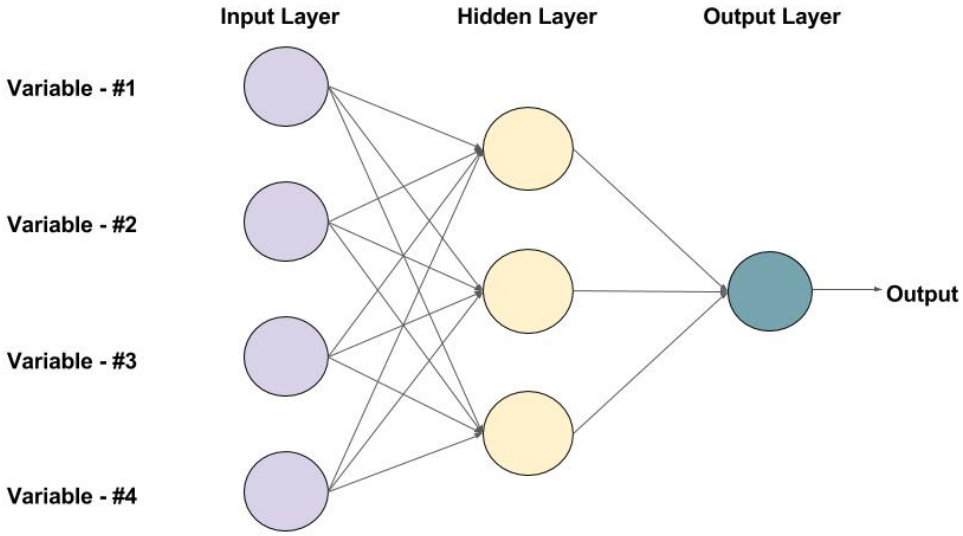
Ridge classifier is a classifier who is using Ridge regression.Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.Ridge regression tends to shrink the "group" proportionately whereas Lasso doesn't since it promotes setting individual regression coefficients to zero to reduce model size.



1.9 Multilayer Perceptron(MLP)

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons.The MLP consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly-activating nodes. Since MLPs are fully connected, each node in one layer connects with a certain weight to every node in the following layer.

The term "multilayer perceptron" does not refer to a single perceptron that has multiple layers. Rather, it contains many perceptrons that are organized into layers.The MLP consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly-activating nodes.



An example of a Feed-forward Neural Network with one hidden layer (with 3 neurons)

1.9.1 Overfitting

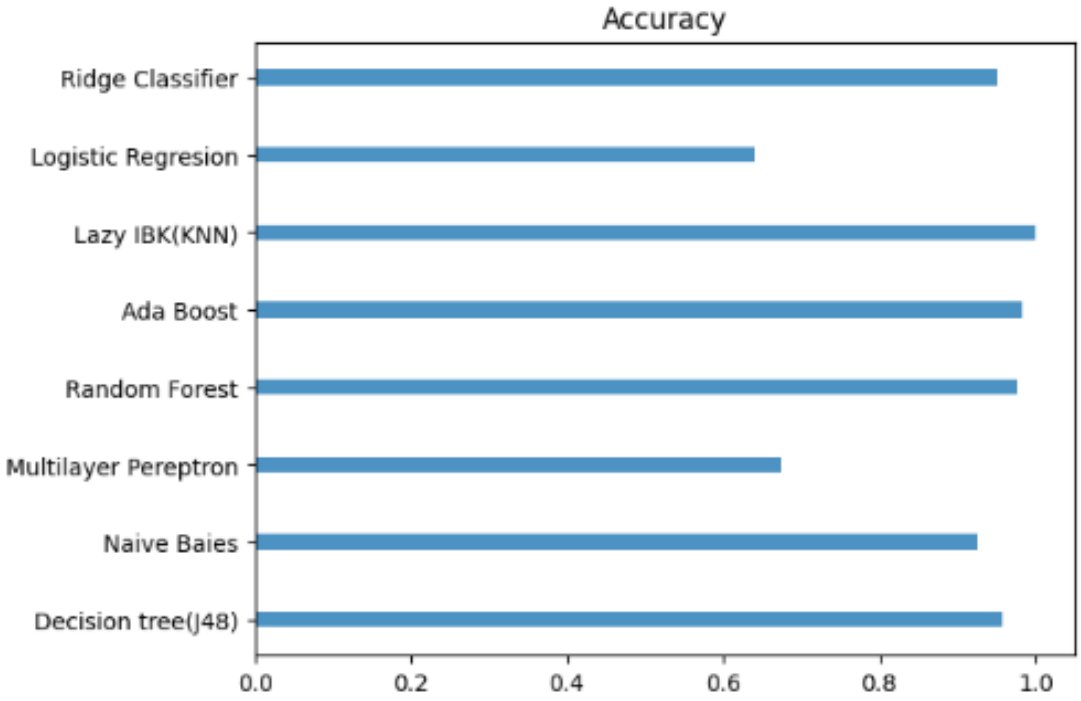
Overfitting is the case where the overall cost is really small, but the generalization of the model is unreliable. This is due to the model learning "too much" from the training data set.

1.9.2 Underfitting

Underfitting is the case where the model has " not learned enough" from the training data, resulting in low generalization and unreliable predictions.

1.10 Algorithm classification results

In the following image it can be seen the accuracy of all the algorithms.The classification accuracy percentage is achieved in two categories which more then 0.95 and less then that .Only MLP and Losgistig Regresion have a bad score but exceed 0.65.



In the next image it can be seen the Relative absolute error, mean squared error, Precision and Recall.

Classifier\Error	MAE	MSE	Precision	Recall
Naive Baise	0.14	0.29	0.91	0.93
Logistic Regression	0.72	1.44	0.32	0.5
Random Forest	0.045	0.91	0.97	0.96
KNN	0.67	1.35	0.60	0.54
MLP	0.65	1.30	0.33	0.5
DecisionTree	0.08	0.17	0.92	0.91
AdaBoost	0.034	0.068	0.98	0.97
RidgeClassifier	0.095	0.19	0.95	0.94

1.11 Conclusion and Observation

Most classifiers had a high accuracy.The MLP it didn't work like in the article even though it has a score of 60-69.The MAE and MSE error are to high and precision is low .I tried different tips like maybe is in underffitting ,but don't work and the score the score did not increase above 70. Most classifiers like AdaBoost it has a precision and accuracy as in the article ,my accuracy is 0.98 and in the article is 82 and the error are about the same.

Random Forest Classifier has has the closest score to the one in the article, even the error are very close, the difference being very very small.

Decission Tree did very well for a maxdepth value of 5 he as has the highest score , although I set the condition of the feature to be random split ,with the 'best' option for spletter the highest score is 9.3333 and with random option is 9.6666 .

For the KNN classifier I tried to improve it but the accuracy did not increase although I looked for the best value for the number of neighbors however, I failed the MAE ,MSE are still very high.

In conclusion, this article was very interesting because so far I have not thought to classify the classifiers and I did not know the different methods by which they are classified.