|  |  |
| --- | --- |
|  | 🞂Comparative analysis of Classification Techniques in Machine Learning |
|  |  |
|  | Name : Aakash Decosta  Roll : 521  Dept : MCMS 2018-2020 |
|  | Abstract Supervised learning is of two types Regression and Classification. Classification specifies the class to which data elements belongs to. It is best used when the output has finite and discrete values. It predicts a class for an input variable as well.  Classification is a learning approach in which the machine learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some classical problems of classification are speech recognition, handwriting recognition, biometric identification, document classification, etc.  In this paper we are going to study seven most commonly used classification algorithms along with the python code: Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine. Keyword Classifier, Classification model, Feature Selection, Binary Classification, Multi-class classification, Multi-label classification Introduction Machine learning techniques used to teach machines how to handle large sets of data so that it can be used for some kind of prediction or recognition purpose. While we are using classification approach for predictiong data we may find there are various classification algorithms. So the question that arise at that moment is which algorithm to choose and which one more accurate. In this paper we are going to study about few of the classification algorithm and test it on a students marks prediction dataset and compare the results of these classification algorithms to determine which algorithm is more appropriate for a particular dataset. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

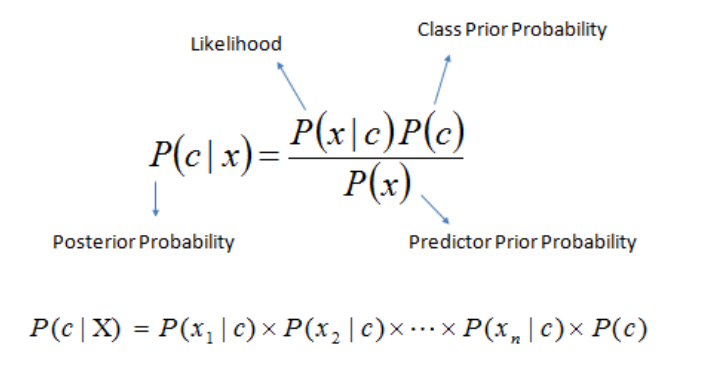
# 1.Naïve Bayes

It is a classification technique based on Bayes’ Theorem. It works with the assumption that the predictors are independence ie are not correlated among themselves. Naïve Bayes classifiers are highly scalable, requiring a few number of parameters in a learning problem. which makes it particularly useful for very large datasets. Regardless of its simplicity, the Naive Bayes model is easy to build and particularly useful for very large data sets. Naive Bayes is known to outperform even highly sophisticated classification methods

Algorithm

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

This assumption is called class conditional independence



Above,

* P(x|c) is “Probability of x given c”, the probability of x given that c happens
* P(x) is the Probability of x
* P(c|x) is “Probability of c given x”, the probability of c given that x happens
* P(c) is the Probability of c.

# 2.Decision Tree

Decision tree builds classification models in the form of a hierarchical structure. Decision tree is developed through step by step incremental process of breaking down the dataset into smaller and smaller. In the final process, it generates a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The root node in a tree which corresponds to the best predictor from given datasets. Decision trees classifier can use for both categorical and numerical data.

Algorithm

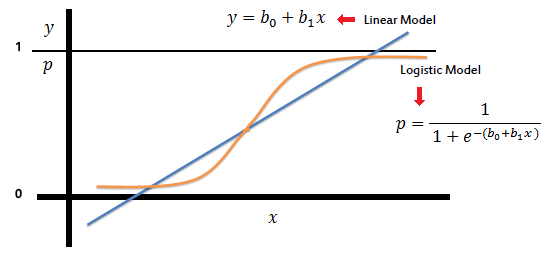
1. The root of the tree is select from the attribute of the dataset by using the concept of information gain.
2. Split the training dataset into subsets. And these subsets prepared in such a way that each subset contains data with the same value for an attribute.
3. Continue the process of step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

# 3.Logistic Regression

Logistic regression is not a regression algorithm but a probabilistic classification model. It is used to predict a binary outcome given a set of independent variables. A linear regression is not suitable to predict the value of a binary variable for two reasons.

* A linear regression cannot predicate the values within an acceptable range.
* Since the target attribute can only have one of two possible values, the residuals will not be normally distributed about the predicted line.

Logistic regressions produce a logistic curve, which is limited to values between 0 and 1.



Logistic regression is similar to linear regression but the curve is constructed using the natural logarithm of the target variable, rather than the probability. Logistic regression has a sigmoidal curve.



It ensures that the generated number is always between 0 and 1 since the numerator is always smaller than the denominator by 1.

So, to build a logistic regression model for doing a multi-class classification, the idea in logistic regression is to cast the problem in the form of a generalized linear regression model.



Multiclass classification with logistic regression can be done either through the one-vs-rest scheme in which for each class a binary classification problem of data belonging or not to that class is done, or changing the loss function to cross-entropy loss

# 4.K Nearest Neighbors

K nearest neighbors a lazy learning algorithm. It is also regarded as a non-parametric technique ie it is used to analyze data when the distributional assumptions are not possible. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). the k-NN classifier calculates the distances between the point and points in the training data set. Usually, the Euclidean distance is used as the distance metric. It also uses the distance metric like Manhattan, Minkowski, correlation, and Chi-square, etc. 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. This is the simplest case.



Algorithm

P is the point, for which label needs to predict number

1. Find the one closest point to P and then the label of the nearest point.
2. Find the k closest point to P and then classify points by majority vote of its K neighbors.
3. Each object votes for their class and the class of maximum votes of its nearest neighbor is taken as the prediction.

# 5.Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane Algorithm. In two-dimensional space, this hyperplane is a line dividing a plane into two parts wherein each class lay on either side. It can be used for both classification or regression challenges. However, it is mostly used in classification problems. SVM are inherently binary classifiers that require full labeling of the data and are directly applied to the two classes available but for the real-life problems which require multiple classes Multiclass SVM is used.

Algorithm

1. Plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate.
2. Generate different hyper-plane and then identify the right hyper-plane.
3. Optimize the hyperplane with maximize margin between the classes.
4. For high dimensional space where we reformulate problem so that data is mapped implicitly to this space.

# 6.Random forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

The fundamental concept behind random forest is A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Algorithm

The random forest algorithm can split into two stages.

1. Random forest creation.
2. Perform prediction from the created random forest classifier.

Random Forest creation:

1. Randomly select “k” features from total “m” features. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until we form the tree with a root node and having the target as the leaf node.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

Random forest prediction:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

This concept of voting is known as majority voting.

# Material and Methods

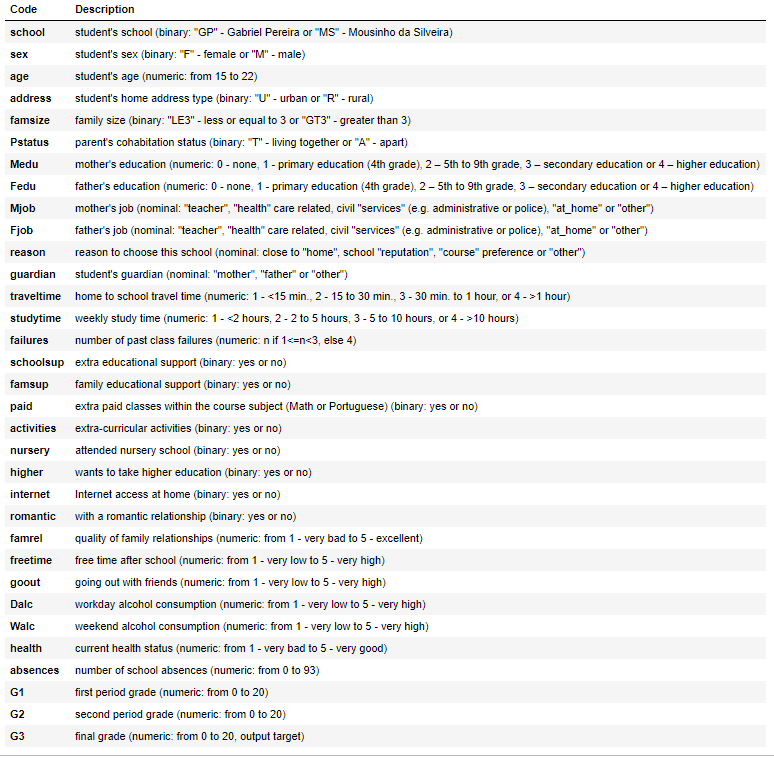
#### Dataset

In this comparative study, we will be using the student prediction dataset from the UCI Machine Learning Repository. This data was collected from the Alentejo region of Portugal from two public schools during the 2005-2006 session. The dataset has been split into the two parts one comprising the core subject of Maths and the other Portuguese.

During a year the student's marks evaluation is divided into three phases G1, G2, and G3 .G3 corresponds to the final grade which is our target attribute to be predicted. The data attributes include student grades, demographic, social and school-related features and it was collected by using school reports and questionnaires.

The maths dataset contains 365 examples and the Portuguese dataset has 649 examples. The attributes of the dataset have been describes in the table below

#### Computational Environment

All the experiments reported in this study where done using Python 3. Scikit-learn machine learning

library was mainly used to create the classification models for comparison. It was also used to split the dataset into the training and testing datasets, and also for preprocessing the data (for converting the non-numerical attributes to numerical attributes). The continuous target attribute was discretized for:

* Binary Classification-pass if G3 more than equal to 10 else fail.
* 5-level Classification-where
  + (0-9) is class V,
  + (10-11) is class IV,
  + (12-13) is class III,
  + (14-15) is class II,
  + and (16-20) is class I.

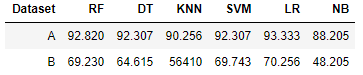
Other libraries like numpy, seaborn, matplotlib, etc were also used from time to time for dataset manipulation and plotting graphs.

# Results

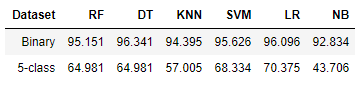
The model that we are going to compare are Decision Tree (DT), Random Forest (RF), Naïve Bayer’s (NB), K nearest neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR).In RF the number of trees was taken to be 200. The models like LR and SVM are mostly used for binary classification but are modified here using certain parameters to be able to handle multiclass classification. Before fitting the dataset into each model the dataset had to be modified in order to convert object(string) type attributes to numerical type. This was done using the sci-kit learns preprocessor known as label encoder. The target attribute of the dataset was altered from continuous values to discrete values. We the better sake of understanding we are going to represent the dataset as

* A - where there are two discrete values (pass and fail).
* B - the dataset with discrete values (5- level Classification).

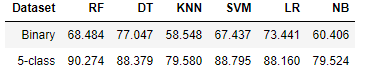
The Portuguese dataset has a larger size so we have only used that to train and test the classification models. The dataset was split using sci-kit learns model selection package ‘train\_test\_split’.The data set was randomly split into a training set and test set of 70% and 30% respectively. The models were compared using sci-kit learns metrics package ‘accuracy\_score’, ’f1\_score’ and also the ‘Pearson correlation coefficient’ from the scipy stats package. The following table shows the score of each model for both the cases.



#### Fig:- Accuracy score in %



#### Fig:- F1 score



#### Fig:- Pearson correlation coefficient in %

# Conclusion

In this paper, we have addressed the prediction of student grades. The grade was in the range of 0 -20. We have created classification models using six different algorithms and compared their accuracy, f1 score and PCC. The dataset altered fit both types classification i.e. Binary-classification and Multiclass classification. We can see the score in the above table. From the results, we can see that the final grade is very much dependent on the previous grades along with few other attributes affecting the target attribute but those other attributes are not significant in all the cases. Since no data exploration or analysis was done to prune the dataset prior to fitting the dataset into the models' redundant features and correlation among different attributes could have affected the results of a few of the models like Naïve Bayers.