**Abstract:**

***Classification is an important data mining technique. It has broad applications to classify the various kinds of data used and prediction in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This case study report is about Car evaluation dataset classification using Naïve Bayes algorithm in R statistical tool.***

Keywords**:** Classification Algorithms, Data Mining, Diabetes, statistical tool

**Introduction:**

Classification algorithms are widely used in various applications. Data classification is a two phase process in which first step is the “***training phase”*** where the classifier algorithm builds classifier with the training set of tuples and the second phase is “***classification***” phase where the model is used for classification. One of the simplest classification algorithm is **Naïve Bayes.**

**Naive Bayes Algorithm:**

Bayesian Classifiers are statistical classifiers based on ***Bayes theorem***. Bayesian classification is very simple and it shows high accuracy and speed when applied to large data sets. It works on one assumption that is the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called ***class conditional independence****.*

Bayesian classification can predict class membership probabilities, such as probability that a given tuple belongs to a particular class. The Naïve Bayesian classification predicts that the tuple X belongs to the class. Using the formula:

Where is ***maximum posteriori hypothesis*** for the class

As **P(X)** is constant for all classes, only needed to be maximized.

If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is,

Otherwise

The class prior probabilities may be estimated by

Where is the number of training tuples of classin .

Given datasets with many attributes, it would be extremely computationally expensive to compute. In order to reduce computation in evaluating**,** the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple i.e., that there are no dependence relationships among the attributes.

Thus,

**=**

Probabilities are easily estimated from the training tuples.

Here refers to the value of attribute for tuple X which may be categorical or continuous valued.

**Literature Review:**

Naive Bayes algorithm is effective in text classification and experiments. Because of its performance Naïve Bayes algorithm is used in many text classification problems. [1] Makes some modifications for Naive Bayes to improve the performance of Naïve Bayes in Spam Filter *categorization.*

In many applications ranking of examples are more desirable than just classification. Weighted Naïve Bayes is an extension of Naive Bayes, in which attributes have different weights. [2] Investigates how to learn a weighted naive Bayes with accurate ranking from data.

Performance of Naïve Bayes algorithm is poor on databases in which attributes depend on each other. This is because the conditional independence assumption is not always true in the real world. [3]used latest Hidden Naive Bayes (HNB) algorithm to increase the performance in databases which have strong correlation between attributes. In the HNB algorithm, each attribute corresponds to a hidden parent which combines the influences of all other attributes. Compared to other Bayesianalgorithms, its performance is significantly improved, but too much test time on high-dimensional datasets cost.

**Training Data:** Dataset used for training purpose is “**Car Evaluation Database**”.

**Sources**: The sample data of data set which we use for classification is Naïve Bayes algorithm is taken from link. [5] {plz add link}

* **Creator:** Marko Bohanec
* **Donors:** Marko Bohanec (marko.bohanec@ijs.si)

Blaz Zupan (blaz.zupan@ijs.si)

* **Date:** June, 1997

**Past Usage:**

The hierarchical decision model, from which this dataset is derived, was first presented in M. Bohanec and V. Rajkovic: ***“Knowledge acquisition and explanation for multi-attribute decision making”***.

* In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. Pages 59-78, 1988.
* Within machine-learning, this dataset was used for the evaluation of **HINT** (Hierarchy Induction Tool), which was proved to be able to completely reconstruct the original hierarchical model.
* This, together with a comparison with C4.5, is presented in Zupan, M. Bohanec, I. Bratko,

J. Demsar: “***Machine learning by function decomposition”***. ICML-97, Nashville, TN. 1997.

**Relevant Information:**

Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX.The model evaluates Cars according to the following concept structure:

**CAR** car acceptability

**PRICE** overall price

**buying** buying price

**maint** price of the maintenance

**TECH** technical characteristics

**COMFORT** comfort

**doors** number of doors

**persons** capacity in terms of persons to carry

**lug\_boot** the size of luggage boot

**safety**  estimated safety of the car

Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, and COMFORT.

Every concept is in the original model related to its lower level descendants by a set of examples [6].

The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug boot, safety. Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.

**Number of Instances:** Number of Instances: 1728

(Instances completely cover the attribute space)

**Number of Attributes:** 6

**For Each Attribute:**

buying v-high, high, med, low

maint v-high, high, med, low

doors 2, 3, 4, 5-more

persons 2, 4, more

lug\_boot small, med, big

safety low, med, high

**Missing Attribute Values:** None

**Solving in R:**

We have trained the classifier according to the given 6 attributes. For working on Naïve Bayes classification we need to install packages ‘caret’ and ‘klaR’. Load the both library in R.

install.packages ("klaR")

install.packages ("caret")

library(klaR)

library(caret)

For training the model, we have load the training dataset from CSV file. Attributes and classes have been passed for training. As we already know about our dataset so we will load 1700 instances for training and remaining 8 will be test after training.

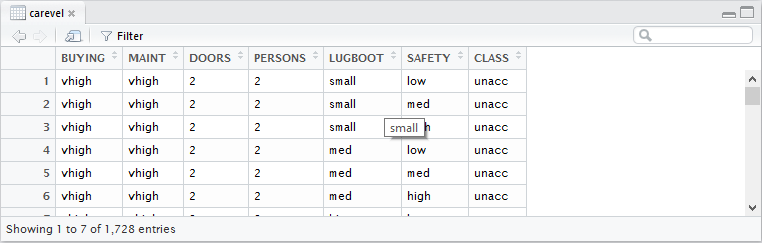
myFile <- read.csv("D://carevel.csv")

attr <- myFile[1:1700,-7]

cls <- myFile[1:1700,7]

test <- myFile[1721:1728,7]

model <- train(attr, cls, ‘nb’)



x<-predict(model,test)



As it the probability of every instance according to both classes is calculated.

summary(x)



This is summary of our predicted result, which is already according to pre-known assigned class.

**Conclusion:**

In this case study, popular Naïve Bayes algorithm is used for classification of car evaluation database using R statistical tool. Classification is done easily and quickly. Error rate in calculation is reduced by using R statistical tool. Therefore, we can say that classification can be done more efficiently with the help of Naïve Bayes algorithm and predicted results are more accurate.

# References

[1]GuoQiang. “ An Effective Algorithm for Improving the Performance of Naive Bayes for Text”.

[2]Harry Zhang ,Shengli Sheng,” Learning Weighted Naive Bayes with Accurate Ranking”.

[3]Yaguang Ji, Songnian Yu, Yafeng Zhang. “A novel Naive Bayes model: Packaged Hidden”. *Second International Conference on Computer Research and Development*.

[4] M. Bohanec , V. Rajkovic: Expert system for decision making. Sistemica , pp. 145-157, 1990.)

[5]

[6] <http://www-ai.ijs.si/BlazZupan/car.html>