Prediction of Student Alcohol Consumption Level Using Various Machine Learning Techniques

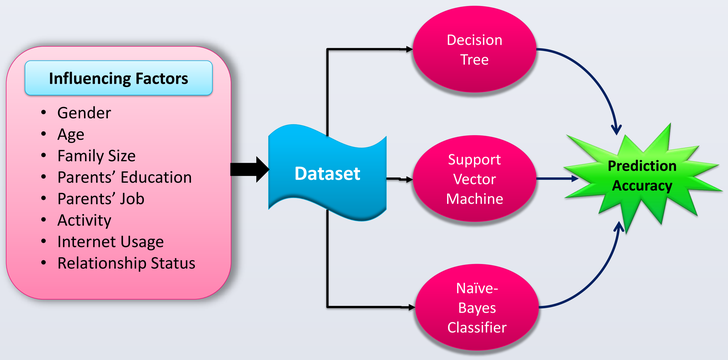
PRABAL BASU, A02049867, CS 5600 Class Project

Utah State University

ABSTRACT

This project compares the prediction accuracies of different machine learning algorithms, for alcohol consumption level among school students. In this pursuit, three machine learning models, such as Decision Tree, Support Vector Machine and Naïve-Bayes classifier are used on two real life data sets. Additionally, the importance of various features are deduced, that highly impact the prediction accuracy of an algorithm.

PROBLEM DESCRIPTION



Flow Diagram of the Project

Regular consumption of alcohol has adverse impacts on our physical and mental health. There are many factors that influence the drinking pattern of a person, for example, gender, upbringing, socio-economic state of the country etc. Accurate mining of the alcohol consumption data-set is necessary to correctly isolate the important factors that maximize the consumption. This project considers two real life data sets, from the *UCI Machine Learning Repository*, and applies three machine learning algorithms (e.g., *Decision Tree, Support Vector Machine and Naïve-Bayes Classifier*) to predict the alcohol consumption, subjected to various factors. The prediction accuracies are computed in terms of the average accuracy from 10-fold cross validation. Further, the features that strongly influence the alcohol consumption are extracted using *Information Gain* analysis from a Decision Tree model.

DATASET

Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Multivariate	**Integer**	**Classification**	**395, 649**	**27**

The original data was integrated into two data-sets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes. Contributors: *Paulo Cortez and Alice Silva*.

METHODOLOGY

➢*Data Preprocessing*:

- **Finding a binary output:** The dataset has two features, viz., *Dalc* (workday alcohol consumption) and *Walc* (weekend alcohol consumption), both in the range of 1 (very low) to 5 (very high). Derived output: ***Alc = (Walc X 2 + Dalc X 5) / 7***, again, in the range of 1 - 5. Derived binary output: ***Alc_bin = Alc < Th ? 0 : 1*** *(where 0 and 1 denote non-alcoholic and alcoholic, respectively)*. **Different values of *Th* are explored (e.g., 2, 2.5 and 3).**

- **Filtering poorly correlated features:** Using correlation co-efficient, 4 features were discarded for which the correlation values are less than a given threshold.

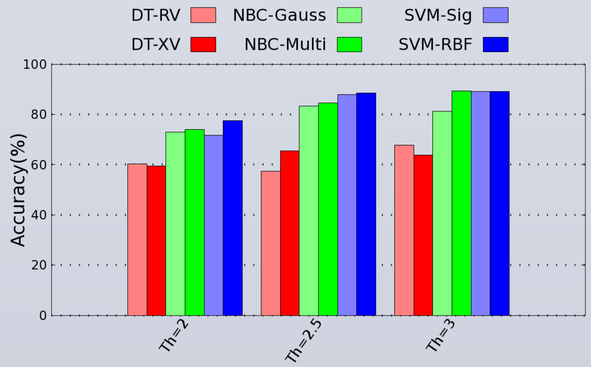
➢*Machine Learning Models:*

- **Decision Tree (DT):** *Information Gain* based analysis is used for prediction and feature extraction. Random validation (RV) and 10 -fold Cross Validation (XV) techniques are used for determining prediction accuracy.

- **Support Vector Machine (SVM):** A multi-class SVM with C-support vector classification is considered. The default parameters for the libSVM tool are used. Two kernel types (e.g., Radial Basis Function (RBF) and Sigmoid) are explored. Original categorical features are converted into discrete features in a sparse format.

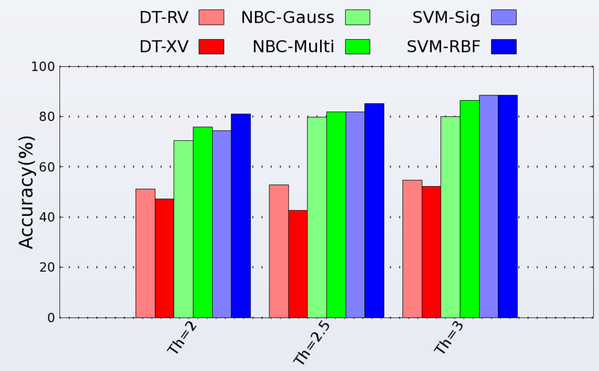
- **Naïve-Bayes Classifier (NBC):** Gaussian and Multinomial decision rules are explored. The results from Bernoulli decision rule are discarded due to extremely poor accuracy. Laplace smoothing parameter is used for the multinomial decision rule. Original categorical features are converted into discrete features in a non-sparse format.

EXPERIMENTAL RESULTS



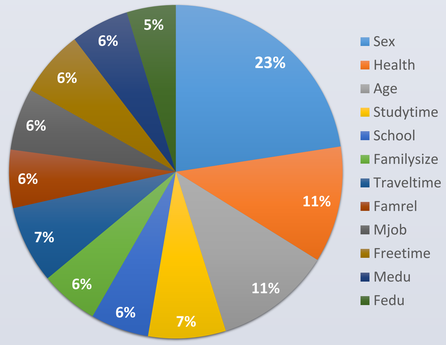
Accuracies for the students of Mathematics class

EXPERIMENTAL RESULTS (Contd.)



Accuracies for the students of Portuguese class

Considering a higher threshold (Th) for alcohol consumption to determine an alcoholic, increases the prediction accuracy for all the classifiers. For NBC, multinomial decision rule gives better accuracies than Gaussian. For SVM, RBF kernel achieves a better accuracy compared to Sigmoid kernel. The SVM, in general, offers the best accuracies amongst all the classifiers.



Important feature extraction using Information Gain. A higher percentage indicates a greater importance.

SUMMARY AND CONCLUSION

- SVM classifier gives the best result in terms of average accuracy.
- RBF kernel gives better accuracy w.r.t. sigmoid kernel for SVM.
- Multinomial rule gives better accuracy w.r.t. Gaussian for NBC.
- In general, a higher accuracy is obtained for a larger value of ***Th***.
- Sex, health and age are the top three important features.

REFERENCES

[1] http://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION#
[2] libSVM, https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3] Naïve-Bayes classifier, https://github.com/timnugent/naive-bayes
[4] http://arstechnica.com/science/2016/03/drunk-tweeting-computer-algorithm/