Drug Consumption – Decade based Classification problem

The Drug classification problem analysis is commenced with the AS-IS Dataset from the UCI website. The initial set of assumptions are the data is collected from 1885 informants. Data is collected using 5 methodologies including Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information.

Psychologists have largely agreed that the personality traits of the Five Factor Model (FFM) are the most comprehensive and adaptable system for understanding human individual differences [11]. The FFM comprises Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The data set has information about 18 psychoactive drugs.

Every respondent chose one of the 7 classes for each drug as ‘Never Used’ [C0], ‘Used over a Decade Ago’[C1], ‘Used in Last Decade’[C2], ‘Used in Last Year’[C3], ‘Used in Last Month’[C4], ‘Used in Last Week’[C5], and ‘Used in Last Day’[C7]. For the purpose classification in this Machine learning algorithm the **6 Drugs** under consideration are Alcohol, VSA, Nicotine, Legal Highs, Ecstasy, Amphet and every drug is subjected to a Binary Classification on the above-mentioned classes separated as “**Users** (1)” and “**Non-Users** (0)”. The base classes ‘Never Used’ [C0] and ‘Used over a Decade Ago’[C1] grouped as “**Users** (1)” while other classes grouped as “Non-Users”.

This analysis activity will perform the machine learning using the statical and probabilistic models for this binary classification problem. The null hypothesis [H0] we start this classification problem is there is no significant classification between Users and non-users of the respective drugs. The drugs under analysis such as **Alcohol, VSA, Nicotine, Legal Highs, Ecstasy, Amphet** will be binary classified using the models and accuracy for each drug will be calculated and analyzed for thisactivity.

**GitHub Repository Link:**

**Modelling process:**

The activity of model building to classify the drug consumption problem under consideration for this activity consists of the following steps,

* Feature Extraction
* Feature Selection
* Model Building – Decision tree [DT], Random Forest [RF], Support Vector machine [SVM], KNN [K – Nearest Neighbor]
* Evaluation
* Prediction
* Compare with the research paper

**Step 1 :Feature Extraction:**

Feature extraction is the process of converting raw unprocessed data into numerical values for the easiness of computation. The data set we have started the modelling process is already extracted and all the features are extracted and normalized. Hence, for this activity there was no extra step done for extraction.

**Step 2: Feature Selection:**

Feature selection is essential for efficient predictive models. This allows us to filter out the input variables highly dependent on the response variables. In our case we are using ANOVA (Analysis of Variance) F measure to do feature selection with "Filter Based" technique in Scikit learn. In our analysis we have 12 input features, and the selection is based on the variance of the input parameters. With 8 selected features using the SelectK best Algorithm such as Age, Gender, Ethnicity, NScore, AScore, CScore, Impulse and SS.

**Step 3: Model building:**

Four models are used to classify this binary classification problem with the selected 8 features. The classifiers used are Decision tree, Random Forest, SVM and kNN.

**Step 4: Evaluation:**

The evaluation metrics used are accuracy of the model, precision, recall. To provide a better measure of accuracy ROC and AUC is used for all the 4 classifiers. The

**Step 5: Prediction:**

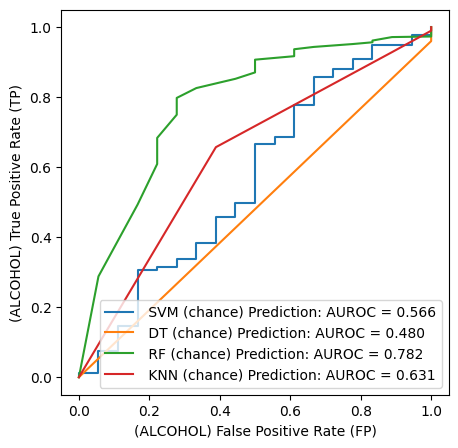
* The **Random Forest classifier is having the highest Recall (Sensitivity) rate for drugs Alcohol, Amphet, Nicotine and VSA**. (Refer Table 1.0)
* While the SVM classifier has a good prediction for Legal H and Ecstacy.

Graphical user interface, application, table, Excel

Description automatically generatedTable 1.0

The users are classified as 1 and non-users as 0. The confusion matrix highlights that for certain drugs data is skewed more towards users (Such as Alcohol) and there are less samples for non-users (Such as Semer and VSA) which is reflected in the scores. Alcohol shows the highest accuracy for RF (0.785) and a precision of 0.99 for the True Negatives (Users) while the True positives (non-users) are almost null due to unavailability of data for non-users.

While for the VSA, there are more non-users than users in the sample set the prediction classifier that worked well is RF (0.825) with a sensitivity (Recall) rate (0.99) is high for non-users in RF and Precision (Positive predictive value) is 0.87. Which makes RF a good classifier for the Alcohol and VSA.

** Chart

Description automatically generated**

**Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated**

**Lessons learnt:**

**Parallel study with Research Paper:**

1. **Feature Selection:** The Research paper has a specific feature selection for every drug to bring out the significant correlation and effect of the input features with the specific drug consumption. For example,
   1. For Drug ecstasy consumption the best classifier is DT based on features age, SS, gender and has sensitivity 76.17% and specificity 77.16% using just 3 features.
   2. For our Model we have considered 8 features selected using ANOVA F measure for the drugs
2. **Hyper Parameter Tuning:** In the research paper for the DT Gini gain, information gain and DKM gain are considered while developing the decision tree. For our activity, Gini and Entropy were used to see is the hyper parameter tuning using Grid search helped to improve the score for Alcohol. The performance was improved slightly for the prediction as denoted in the ROC curve below. (Refer Figure 1.0)

**Conclusion:**

With this binary classification it is evident that the respondents can be clearly classified as users and non-users and hence we reject the null hypothesis (H­­0).

Line chart

Description automatically generated with medium confidence

**Figure 1.0**