1) *Some of the Open source machine learning packages available are:*

TensorFlow – used by google, Scikit learn, Keras – needs TensorFlow to run, NLTK, XG – Boost, Theano, Accord.NET, Apache Spark, Py Torch – used by Facebook

The machine learning package selected for this assignment is **WEKA** "**Waikato Environment for Knowledge Analysis**"

- WEKA is preferred because it is incredibly user friendly data mining computer software, as it includes visualization tools and algorithms which makes graphical viewing of data easy.
- WEKA supports various data mining and many machine learning tasks like clustering, classification etc. Since this assignment purpose is to focus on basics of machine learning packages for a classification task. WEKA is the best choice.

a) *The main features of WEKA are:*

- It is implemented in java programming language and runs on any present-day computing platform.
- WEKA has a large number of features like classification, regression, pre-processing, clustering, data association rules, data visualization etc.
- WEKA also has large trends in Artificial Intelligence. It is platform independent, open source package and is available free under GNU General Public License.

2) *Data Preparation steps:*

- WEKA prefers to upload data in various formats like .arff, .csv, .data and many more.
- The dataset given "hazelnuts.txt" is converted to an ".arff" file which has 2 sections
- First line includes, name of the relation "@relation" followed by name of the file, next all the 12 attributes given are added with a prefix "@attribute" followed by its type and this will constitute the header information.
- The data is then added after "@data" which is the data section. Each line of data is a single instance.
- The data given in hazelnuts.txt, each row is the data for one attribute, hence converted each row into a column and added the values side by side separated by a comma to create instances for mining. And a total of 201 instances are formed. This file is now ready to be loaded into WEKA for classification.

Example of one instance:
@data
71,11.67,12.8025,8.055074738,34.65,1375.5,0.93005,19.145,4.4604,0.048667685,0.175 ,c_avellana

3) *Classification Algorithms used "J48" and "Naive Bayes Classifier":*

A. *C4.5 also known as J48 in WEKA:*

- I have used J48 algorithm which is implemented in WEKA using Java, to classify and predict "**variety**", which has 3 classes i.e. c_americana, c_cornuta, c_avellana. This algorithm can have attribute values in numeric, however the class we are trying to predict must be nominal. Hence it works perfectly fine for our current dataset. This algorithm uses **divide and conquer method**.
- The prepared data set is in .arff file, and it is loaded to weka in following steps:
- Open weka explorer, choose the hazelnuts.arff file and load it into the application. We can see all the 12 attributes specified and the graphs representing the occurrence of data.(fig attached in page 6)
- We select the attribute for which we want to generate the decision tree, so select attribute variety, go to classify choose J48 algorithm under trees, give the cross-validation value as 10, select (Nom)variety, if numeric is selected for this algorithm, start button gets disabled. So click on start after selecting appropriate attribute.
- The classifier output displays the results of the data set for attribute variety, the splitting of tree is based on highest information gain.

$$Information\ Gain(N, A) = Entropy(N) - \sum_{values(A)} \frac{|N_i|}{|N|} Entropy(N)$$

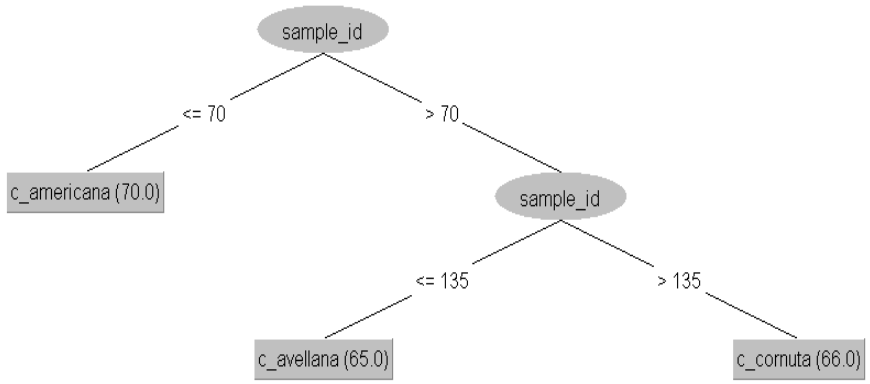
- Here, Entropy(N) is the measure of disorder of data in the dataset

Output results:

Correctly Classified Instances	199	99.005 %
Incorrectly Classified Instances	2	0.995 %

- **The algorithm made 199 correct predictions of 201 so accuracy is 99.005%**

Decision Tree for J48:



References:

1. ***DECISION TREE ANALYSIS ON J48 AND RANDOM FOREST ALGORITHM FOR DATA MINING USING BREAST CANCER MICROARRAY DATASET*** *Ajay kumar Mishra1 , Dr.Subhendu Kumar Pani2 , Dr. Bikram Keshari Ratha3 1PhD Scholar, Utkal University, Odisha, (India) 2Associate Prof., Dept. of CSE, OEC, BPUT, Odisha, (India) 3Reader, Utkal University, Odisha, (India)*
2. ***DECESION TREE ANALYSIS ON J48 Algorithm for Data Mining*** *Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava Dept. of MCA,Govt. Women Engineering College, Ajmer, India, Manish Mathuria Dept. of C.E. & I. T.*

B. Naive Bayes Classifier:

- I have used Naive Bays Classifier algorithm to classify and predict "**variety**", which has 3 classes i.e. c_americana, c_cornuta, c_avellana, as this algorithm is not covered in the class. The algorithm uses **Probability for calculating the occurrence** of the attribute variety.
- Since total number of instances are 201,

We can calculate the P(c_ americana) = 70/201
Similarly, P(c_ cornuta) = 66/201
P(c_avellana) = 65/201

- The same data set used for J48 algorithm is used here as well , and it is loaded to weka following the same steps as above:
- Once the file is loaded to weka, We select the attribute for which we want to calculate the probability , so select attribute variety, go to classify choose NaiveBayes algorithm under bayes, give the cross-validation value as 10, select (Nom)variety, if numeric is selected for this

algorithm, start button gets disabled here also. So click on start after selecting appropriate attribute

- The classifier output displays the results of the data set for attribute variety.

Output results:

Correctly Classified Instances	194	96.5174 %
Incorrectly Classified Instances	7	3.4826 %

- **The algorithm made 194 correct predictions of 201 so accuracy is 96.5174 %**

References:

1) ***An empirical study of the naive Bayes classifier*** *I. Rish T.J. Watson Research Center rish@us.ibm.co*

4) *Cross-validation and performance estimate of each classification model:*

- Cross-validation is a statistical method to compare two or more machine learning models and their efficiency.
- 10 fold cross validation is, when a particular data set is divided into 10 folds, then 9 folds are used as training data and 1 fold is used as test data**.**
- This same procedure is continued for all the folds, now suppose the 2nd fold is used as test data and the remaining 9 folds are used as training data and so on.
- If cross validation is given as 2, it means it divides the data into 2 folds, 50% training data and 50% test data
- Note that the above algorithms were both performed with 10 fold cross validation

RESULTS:

J48 ALGORITHM:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	199	**99.005 %**
Incorrectly Classified Instances	2	0.995 %
Kappa statistic	0.9851	
Mean absolute error	0.0066	

Root mean squared error 0.0814
Relative absolute error 1.493 %
Root relative squared error 17.2784 %
Total Number of Instances 201

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.000	1.000	0.986	0.993	0.989	0.993	0.991	c_americana
	1.000	0.007	0.985	1.000	0.992	0.989	0.996	0.985	c_cornuta
	0.985	0.007	0.985	0.985	0.985	0.977	0.989	0.974	c_avellana
Weighted Avg.	0.990	0.005	0.990	0.990	0.990	0.985	0.993	0.984	

=== Confusion Matrix ===

a	b	c	<-- classified as
69	0	1	a = c_americana
0	66	0	b = c_cornuta
0	1	64	c = c_avellana

****

NAÏVE BAYES ALGORITHM:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 194 **96.5174 %**
Incorrectly Classified Instances 7 3.4826 %
Kappa statistic 0.9478
Mean absolute error 0.0245
Root mean squared error 0.1474
Relative absolute error 5.523 %
Root relative squared error 31.2688 %
Total Number of Instances 201

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.957	0.008	0.985	0.957	0.971	0.956	0.999	0.998	c_americana
	0.970	0.015	0.970	0.970	0.970	0.955	0.992	0.988	c_cornuta
	0.969	0.029	0.940	0.969	0.955	0.933	0.973	0.932	c_avellana
Weighted Avg.	0.965	0.017	0.966	0.965	0.965	0.948	0.988	0.973	

=== Confusion Matrix ===

a	b	c	<-- classified as
67	1	2	a = c_americana
0	64	2	b = c_cornuta
1	1	63	c = c_avellana

Conclusion:

- *The results obtained by the algorithms revealed that J48 performs well than Naïve Bayes, as success rate of J48 algorithm nearly tends to 100 % mark while the later*

*provides up to 97%. The two models give slightly similar yet different results because, the **Relative absolute error** is slightly higher for Bayes algorithm when compared to J48. Hence, algorithms with lesser error values are always preferred.*

References:

1) ***PERFORMANCE EVALUATION OF J48 AND BAYES ALGORITHMS FOR INTRUSION DETECTION SYSTEM,*** *Uzair Bashir & Manzoor Chachoo Mewar University, Chittorgarh, Rajasthan, India University of Kashmir, Srinagar, India*

A Graph from WEKA, displaying the occurrences of all the attributes in the dataset

The algorithms were run to predict the occurrence of attribute variety which is the last graph on the bottom right.