REPORT ON ASSIGNMENT 1 (MACHINE LEARNING LAB)

❖ PROCEDURE

➢ USING WEKA

❑ ON BREAST CANCER DATASET

- The dataset was loaded in .arff format and two different classifiers were successively applied to it and the results obtained.
- Cross – validation was set to 5 folds and Naive Bayes classifier was first selected to obtain the classification results.
- Again, cross-validation was set to 5 folds and this time, 'Logistic function'(it implements Logistic regression in WEKA) was selected and the results obtained.

❑ ON HOUSE-VOTES-84 DATASET

- The dataset was loaded in .arff format and two different classifiers were successively applied to it and the results obtained.
- Same as in previous dataset, here also the two classifiers: Naive Bayes and Logistic Function were applied with 5-fold cross-validation and the results were obtained.

➢ USING SCIKIT-LEARN IN PYTHON

❑ ON BREAST CANCER DATASET

- A function was implemented to run both the classifiers on both the datasets.
- The classification was done using 5-fold cross-validation using the 'KFold' functionality.
- First, the Naive Bayes model was imported from scikit-learn and the model was trained on the data and then predictions were made on the test data; accuracy was calculated by counting the number of correct predictions.

- Similarly, the Logistic Regression model was imported from scikit-learn and after training, the model was used to predict; again, accuracy is calculated.
- The whole process is repeated 10 times and the average accuracies were taken.
- Finally, a bar graph was plotted showing the comparative performance of both the models.

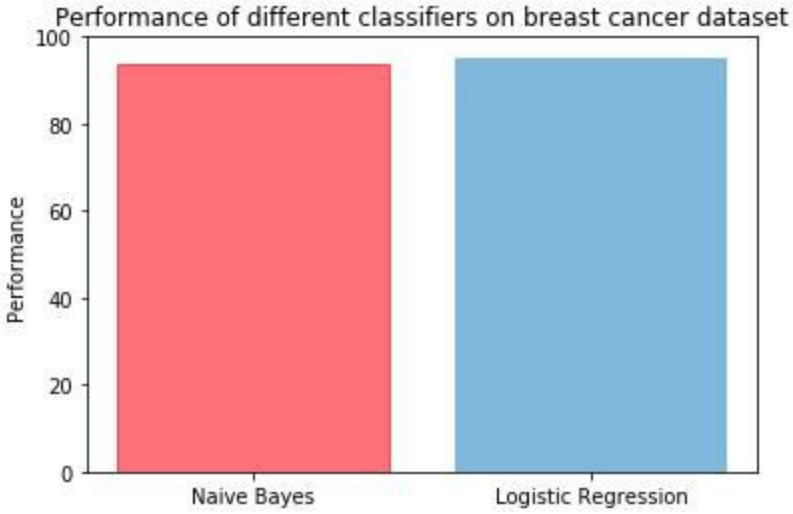
❑ ON HOUSE-VOTES-84 DATASET

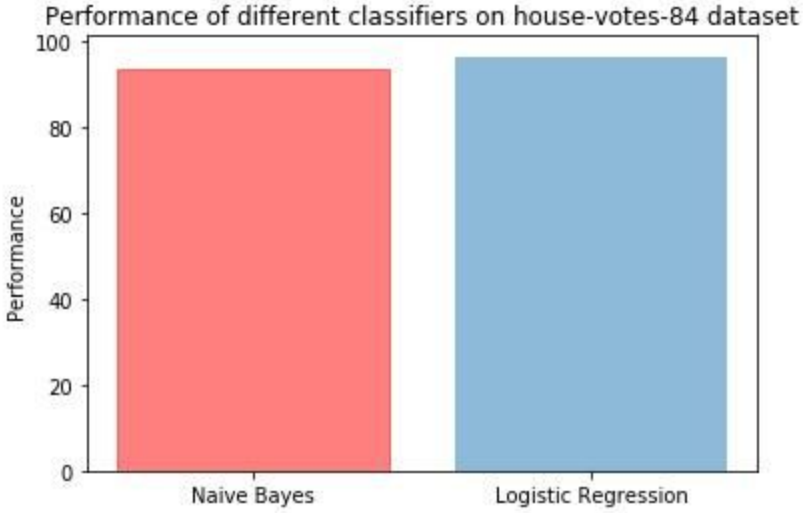
- The dataset was loaded in .csv format, and passed onto the function, which finds out the accuracies and plots the comparative bar graphs again.

❖ RESULTS

Table 1: Accuracy Comparison of different classifiers using different tools

Classifier used	**Using WEKA on Breast Cancer Dataset**	**Using WEKA on House Votes 84 Dataset**	**Using scikit-learn on Breast Cancer Dataset**	**Using scikit-learn on House Votes 84 Dataset**
Naive Bayes	96.14%	90.12%	93.86%	93.79%
Logistic Regression	96.28%	95.63%	95.11%	96.39%





❖ CONCLUSION

- For both the datasets, both the classifiers show equivalent performance and their accuracies are close to each other.
- Naive Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.
- Logistic regression splits feature space linearly, and typically works reasonably well even when some of the variables are correlated.

NAME: SANDIPAN SARMA
ENROLLMENT ID: 510515076