Report

## **AIM:**

In this assignment, we compare python Naive Bayes and Weka Naive Bayes algorithms results along with Hypothesis testing.

We use 3 datasets to complete the project. They are:

• Hayes-Roth Dataset

• Car Evaluation Dataset

• Breast Cancer Dataset

So, by using these 3 datasets we are going to Train & Test the model. We also use K fold for better/more accuracy.

## **INTRODUCTION:**

Based on the Bayes theorem, a Naive Bayes algorithm is still a direct learning technique that addresses categorization problems.

It primarily uses a huge training set for text classification.

One of the simplest straightforward and most efficient classification methods is the Naive Bayes Classifier, which aids in the development of rapid models of machine learning that can generate accurate prognostic.

Formula: Text

Description automatically generated

**K- Fold:**

K-Fold is performed k times using the data divided into k subsets, each of which is used as a test set while the remaining k-1 subsets are utilized for training. This is a validation strategy.

## **Coding:**

Now we are going to Implement the Naive Bayes algorithm in Python Language from scratch

#### **Breast- Cancer Dataset:**

Firstly, I’m using the Naive Bayes algorithm on Breast-Cancer Dataset in Python code with K-fold to see a better outcome.

Graphical user interface, text, application, email

Description automatically generated

**Analysis:** Here we can observe the Average accuracy: 0.7142292490118577 ( x100%) 🡺 71.4222% for breast-cancer dataset.

##### **Hypothesis Testing for Breast-Cancer Dataset:**

Now I applied Null Hypothesis testing to check the normality of every column in Breast-Cancer dataset to compare the accuracy like below.

Graphical user interface, application, table

Description automatically generated

###### **WEKA Result: (Breast Cancer)**

Now are using WEKA to visualize the Naïve Bayes algorithm along with confusion matrix and Accuracy for Breast Cancer Dataset.

Text

Description automatically generated with medium confidence

#### **Hayes-roth Dataset:**

Secondly, Here implementing the Naive Bayes on Hayes-roth Dataset in Python code with K-fold to see a better result.

Graphical user interface, text, application, email

Description automatically generated

**Analysis:** Here we can observe the Average accuracy: 0.6510989010989011 ( x100%) 🡺 65.109% for Hayes-roth dataset.

##### **Hypothesis Testing for Hayes Roth dataset:**

Now we are going to check the normality of every column in Hayes-Roth dataset.

Text

Description automatically generated

since the hypothesis is one sided >> use p\_value/2 >> p\_value\_one\_sided:**0.1430**

Fail to reject null hypothesis

###### **WEKA Result: (Hayes-roth)**

Now are using WEKA to visualize the Naive Bayes algorithm along with confusion matrix and Accuracy for Hayes-roth Dataset.

Graphical user interface, text

Description automatically generated

#### **Car Dataset:**

Finally, By the Naive Bayes on Car Dataset in Python code with K-fold to view accuracy.

Graphical user interface, text, application, email

Description automatically generated

**Analysis:** Here we can observe the Average accuracy: 0.62522156188093 ( x100%) 🡺 62.522% for Car dataset.

##### **Hypothesis Testing for Car Dataset:**

Now I applied Null Hypothesis testing to check the normality of every column in Car dataset to compare the accuracy like below.

Table

Description automatically generated with medium confidence

###### **WEKA Result: (Car)**

Now are using WEKA to visualize the Naive Bayes algorithm along with the confusion matrix and Accuracy for Car Dataset.

A picture containing graphical user interface

Description automatically generated

Here we are going to compare all the data and it’s accuracy result which we receive by applying Naïve bayes algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
| Datasets | Max python accuracy | Python average accuracy | Weka accuracy |
| Breast canser | 82.6086 | 71.4229 | 72.7273 |
| Hayes roth | 84.6153 | 65.1098 | 71.875 |
| Car | 71.7391 | 62.5221 | 85.5324 |

## **Conclusion:**

According to the above table, Weka Naive Bayes accuracy score is little lower than Python Naive Bayes with K-folds accuracy score. Because using k-folds gives opportunity to split data train and test one by one that model train every path of the data and test. That makes accuracy score more higher. Also, we performed Hypothesis testing for every column in all three dataset.

## **References:**

<https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>

<https://machinelearningmastery.com/k-fold-cross-validation/>

<https://archive.ics.uci.edu/ml/datasets/Hayes-Roth> (Dataset)

<https://archive.ics.uci.edu/ml/datasets/Hayes-Roth> (Dataset)

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer> (Dataset)