**Dr.D.Y.Patil Institute of Technology**

Pimpri, Pune-411018

A MINI-PROJECT REPORT

“DMW Mini Project Using Naïve Bayes”

Submitted by

BCOA42 Anubhuti Rane

BCOA43 Sampada Gaonkar

**Department of Computer Engineering**

2019-2020

**Dr.D.Y.Patil Institute of Technology**

PIMPRI, PUNE-411018

**Department of Computer Engineering**

CERTIFICATE

Certified that the mini-project work entitled **“**DMW Mini Project Using Naïve Bayes**”** is a bonafide workcarried out by

**Anubhuti Rane BCOA42**

**Sampada Gaonkar BCOA43**

The report has been approved as it satisfies the academic requirements in respect of mini-project work prescribed for the course.

……………...…………………………

**Subject In-charge**

**Content**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **I** | **Problem Statement** | **4** |
| **II** | **Introduction** | **5** |
| **III** | **Implementation** | **7** |
| **IV** | **Conclusion** | **12** |

**List of Figure**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **1** | **Rapid Miner View** | **7** |
| **2** | **Design View** | **8** |
| **3** | **Results** | **8** |
| **4** | **Accuracy Result** | **9** |
| **5** | **Performance Vector** | **9** |
| **6** | **Outlook Result** | **10** |
| **7** | **Humidity Result** | **10** |
| **8** | **Wind Result** | **11** |
| **9** | **Temperature Result** | **11** |

**I. Problem Statement**

Consider a labelled dataset belonging to an application domain. Apply suitable data pre-processing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyse the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets. For Example: Health Care Domain for predicting disease

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **II. Introduction**  **1.Classification**  Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).  For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam.  Classification belongs to the category of supervised learning where the targets also provided with the input data. There are two types of learners in classification as lazy learners and eager learners.   1. Lazy Learners. 2. Eager Learners.   **Lazy learners** simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.  Ex. k-nearest neighbor, Case-based reasoning  **Eager learners** construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.  Ex. Decision Tree, Naive Bayes, Artificial Neural Networks  **2.Naive Bayes**  **Principle:**  A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.  **Bayes Theorem:**    Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called **naive.**  **Comparison on types of Naive Bayes :**   |  |  |  | | --- | --- | --- | | **Gaussian Naive Bayes** | **Multinomial Naive Bayes** | **Bernoulli Naive Bayes** | | Because of the assumption of the normal distribution, Gaussian Naive Bayes is used in cases when all our features are continuous. For example in Iris dataset features are sepal width, petal width, sepal length, petal length. We can’t represent features in terms of their occurrences. This means data is continuous. | It is used when we have discrete data (e.g. movie ratings ranging 1 and 5 as each rating will have certain frequency to represent). In text learning we have the count of each word to predict the class or label. | Itassumes that all our features are binary such that they take only two values. Means **0s** can represent “word does not occur in the document” and **1s** as "word occurs in the document" . | | |

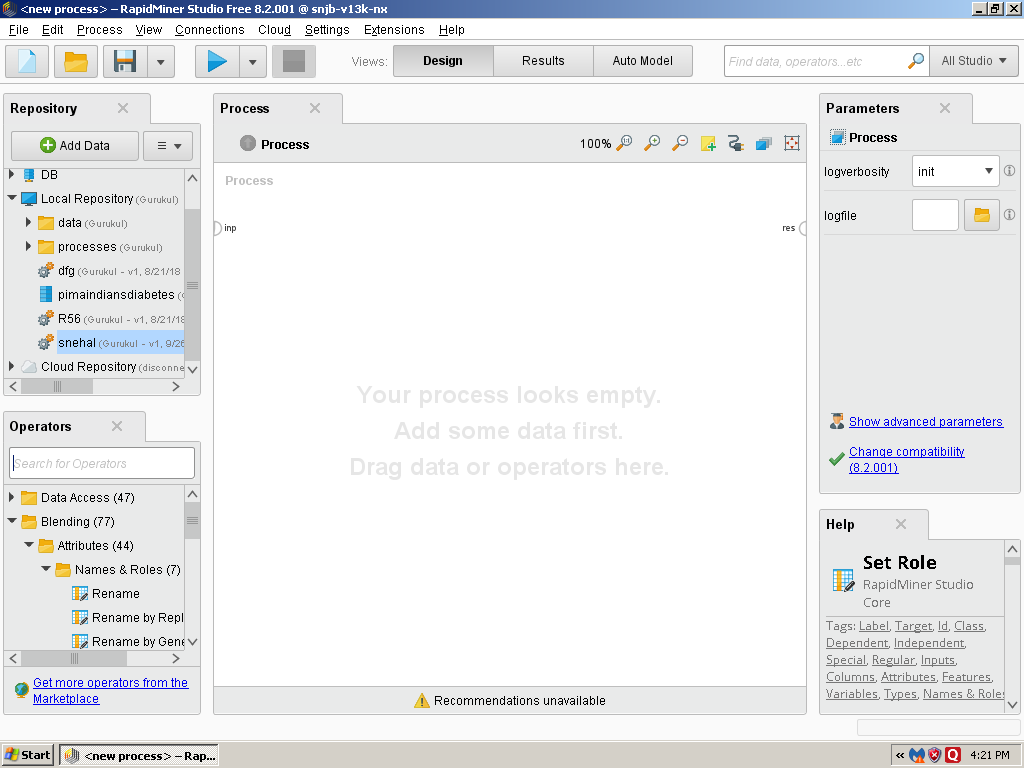
**II. Implementation**

1. **Dataset :**

* **Name:** Golf Dataset.
* **Attributes:** Outlook, Temperature, Humidity, Windy, Play Golf.
* **No of tuples:** 13
* **Source:** In-built dataset in Rapid Miner
* **Description**: We classify whether the day is suitable for playing golf, given the features of the day. The columns represent these features and the rows represent individual entries. If we take the first row of the dataset, we can observe that is not suitable for playing golf if the outlook is rainy, temperature is hot, humidity is high and it is not windy.

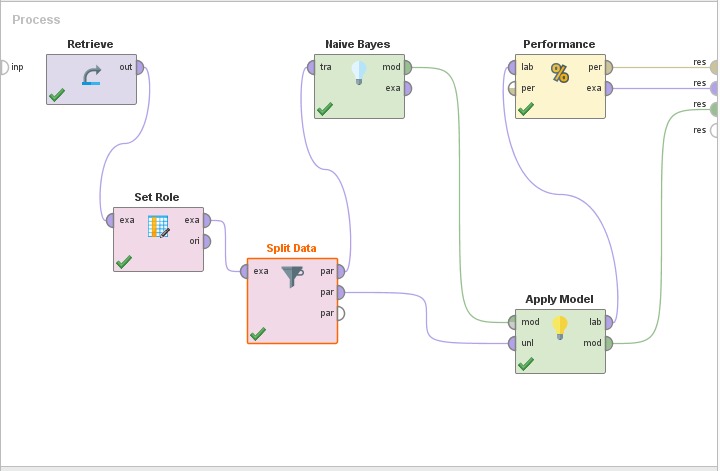
1. **Steps:**

* Open the Rapid Miner Studio



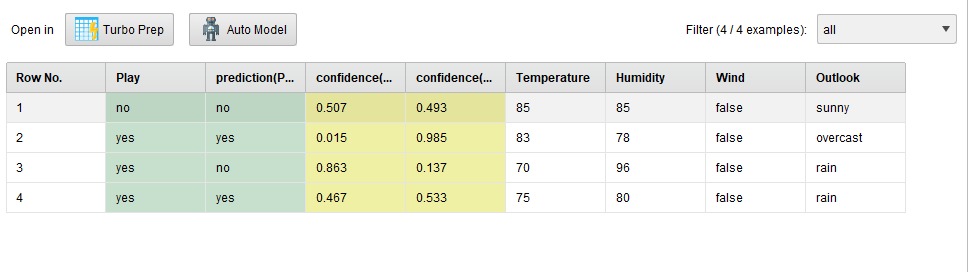
**Fiq 1 : Rapid Miner View**

* Constructing the Naive Bayes Classifier view in the design view.

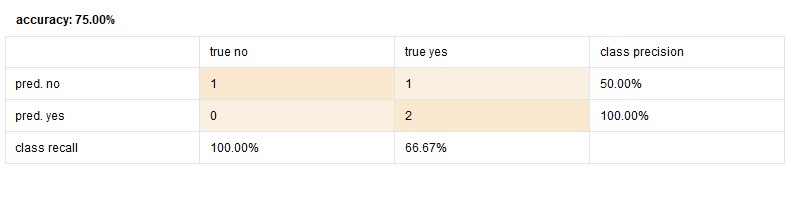
****

**Fiq 2 : Design View**

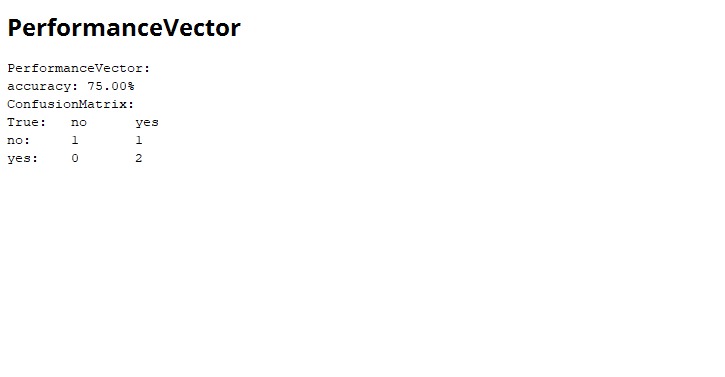
**3. Results:**

****

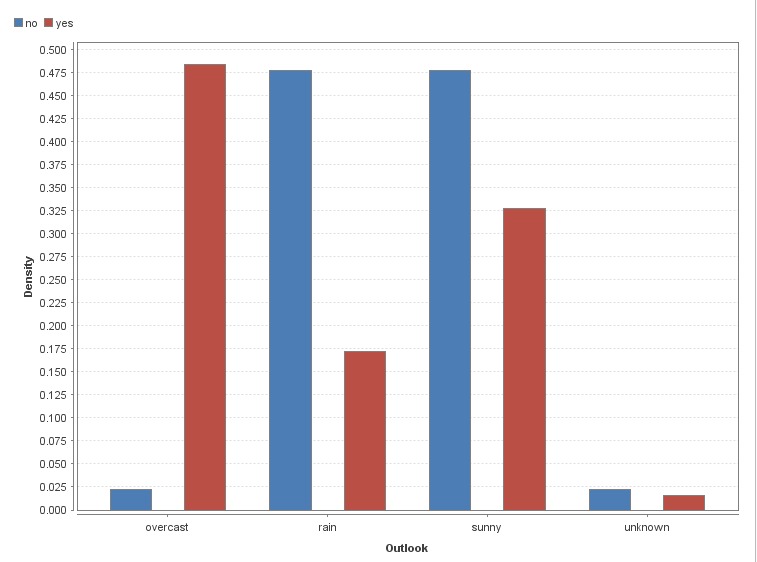
**Fiq 3 : Results**

****

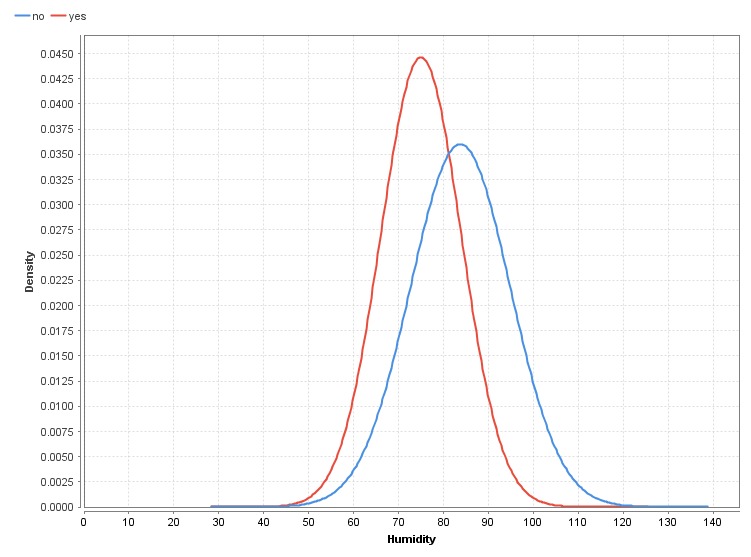
**Fiq 4 : Accuracy Result**

****

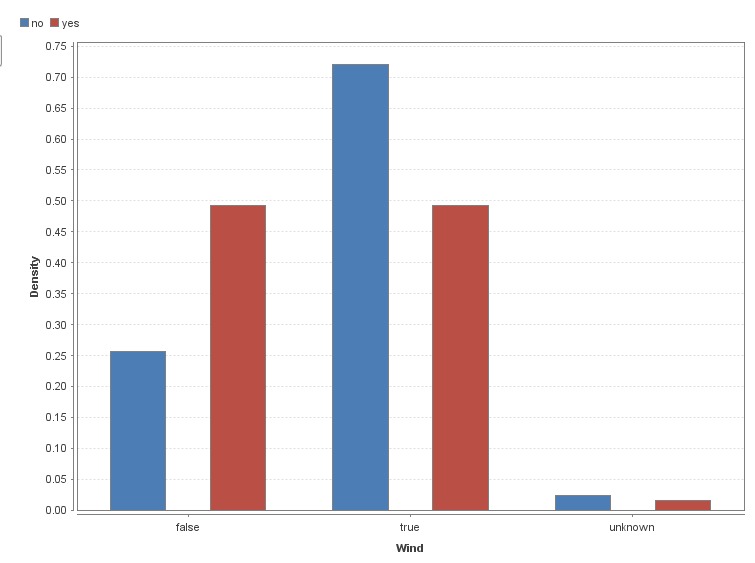
**Fiq 5 : Performance Vector**



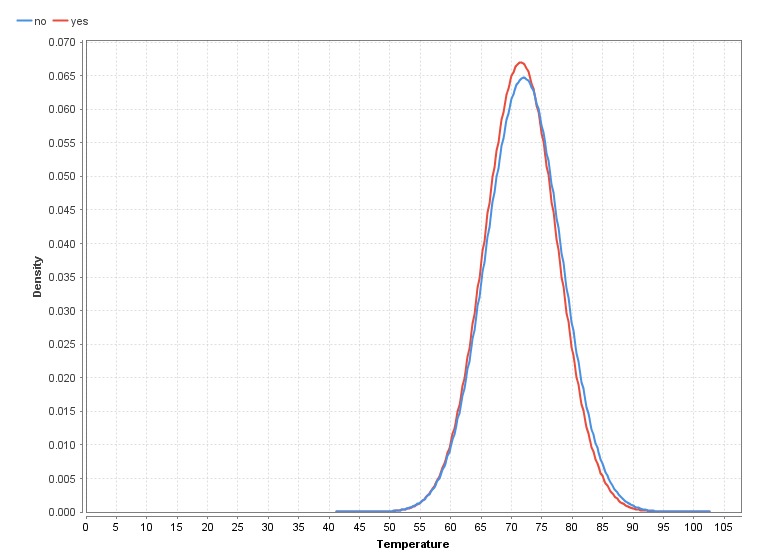
**Fiq 6 : Outlook Result**



**Fiq 7 : Humidity Result**



**Fiq 8 : Wind Result**



**Fiq 9 :Temperature Result**

**IV. Conclusion**

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent. This hinders the performance of the classifier. The golf dataset follows the Gaussian Naive Bayes classification and thus a accurate result is obtained.