**HOMEWORK 5**

**Nikhil Sanjay Thorat**

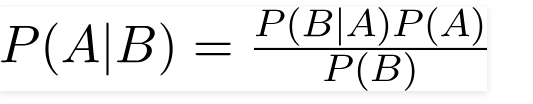
**Spring 2020 ALY 6020**

**Predictive Analytics Instructor: Prof. Marco Montes de Oca**

**Introduction to Naïve Bayes:**

Naïve Bayes classification can be termed as probabilistic classifiers which works on Bayes theorem. Naïve Bayes is also a linear classifier and it falls under a supervised machine learning algorithm. In other terms, the Naïve Bayes classifier can be called a probabilistic machine learning algorithm which is widely used for classification tasks. Naïve Bayes works on probabilistic assumptions such as predicting whether the word occurs in a text or not which is why it works on the principle of Bayes Theorem. In addition to these characteristics, this algorithm is called naïve as the predictors or features are independent which implies the presence of one feature variable does not have any impact on the other.

**Formula of Bayes Theorem:**



Where, Probability of A happening is calculated provided B is occurring. Thus, B is the evidence and A is the hypothesis.

**Types of Naïve Bayes:** Naïve Bayes classifier has three types.

* **Multinomial Naïve Bayes:** This type of classifier is mostly used for documentation problems, i.e. to predict if the given document belongs to categories such as science, sports, media, etc. Thus, the features used in this type of classification are the words present in the document. This calculates the conditional probability of a particular word in a document.
* **Gaussian Naïve Bayes:** This type of classification is used when the predictor values are not discrete and are continuous. Thus, the features present are bound to follow Gaussian distribution or normal distribution.
* **Bernoulli Naïve Bayes:** This type of classifier is similar to a multinomial naïve Bayes classifier, but the difference is that predictors are Boolean. The features that are used for prediction takes only values yes or no.

**Applications of Naïve Bayes algorithm:**

* **Text Classification**: The Naïve Bayes algorithm is widely used for text classification as it follows independent rule. Text classification applications such as email spam filtering, sentiment analysis, i.e. classifying reviews on social media into positive, negative comments.
* **Recommendation systems**: Recommendation systems are a specific type of machine learning algorithm that is used to predict if a user would like or buy a specific resource or product. Thus, Naïve Bayes is used to build recommendation systems.
* **Multi-class prediction**: This algorithm can also be used for multi-class feature prediction. This algorithm can be used for predicting the probabilities of multiple classes of labels or target variable.
* **Real-time predictions:** Since Naïve Bayes has certain advantages like time saving and as it is a fast learning classifier, it is widely used for real-time predictions.

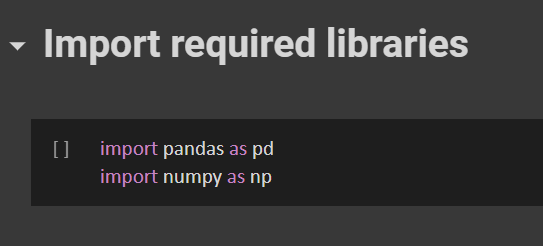
**Homework Outline:** The goal of the assignment is to build a Naïve Bayes classifier model for the classification of time series data.

**Dataset description:** The given dataset was downloaded from the link mentioned below:

<http://www.timeseriesclassification.com/description.php?Dataset=ECG200>

The dataset represents the electrical activity that was recorded during one heartbeat. The features are used for predicting two classes, i.e. normal heartbeat and Myocardial Infarction. Two separate datasets are given for train and test which is used for building the model. The target column has values 1 and -1 which classifies the heartbeat into normal and myocardial infarction respectively. This dataset is described as “Generalized feature extraction for structural pattern recognition in time series data”.

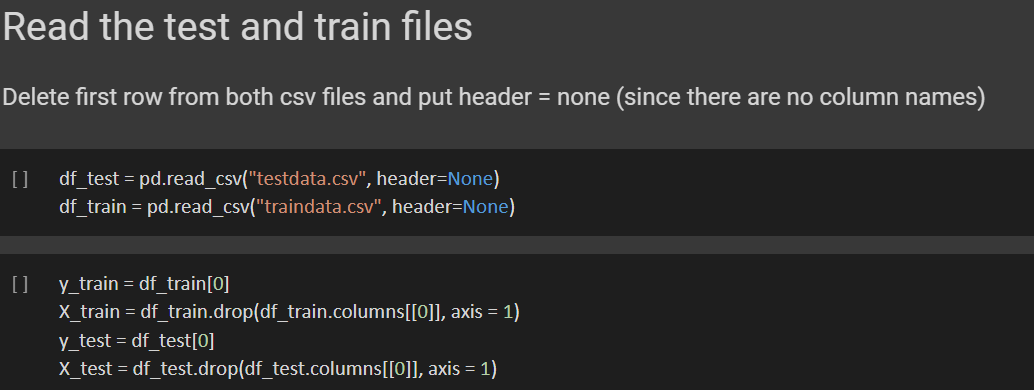
**Step1**



**Fig.1**

From fig.1 it is seen that necessary packages required to start the process of building a naïve Bayes model are imported. Pandas is called using variable pd and NumPy using np.

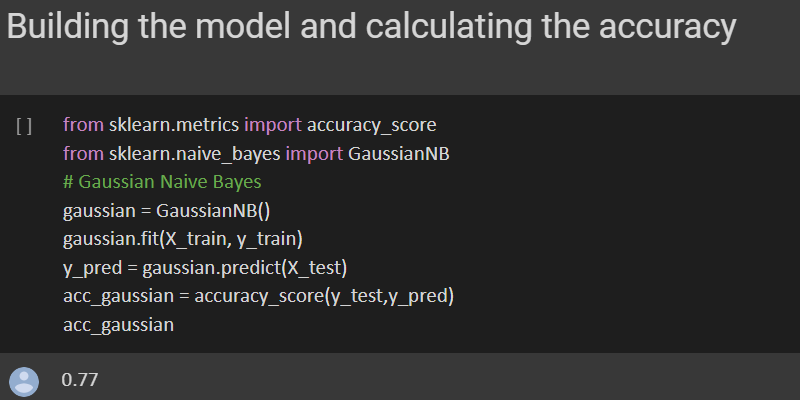
**Step 2:**



**Fig.2**

In Fig.2, the train and test datasets are loaded in df\_test and df\_train variables respectively. X\_Train and X\_test contain feature variables and Y\_train and Y\_test contains the target variables.

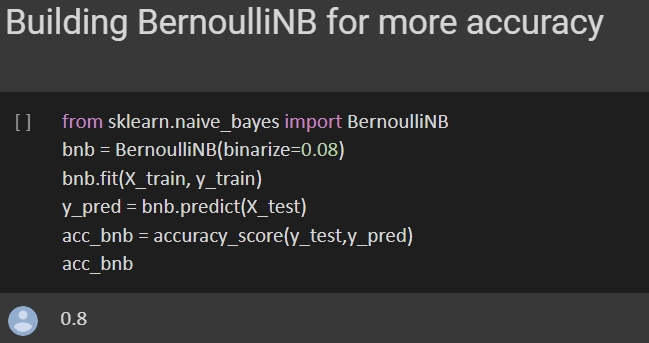
**Step 3:**



**Fig.4**

In fig.4 from sklearn.metrics, accuracy\_score is imported and from sklearn.naive\_bayes GaussianNB was imported. GaussianNB is efficient when the target variables are continuous. Using gaussian.fit the model is fitted with x\_train and y\_train. For prediction y\_pred, x\_test is used. And later on, the accuracy score is calculated. The accuracy score obtained is 77%.

**Step 4:**



**Fig.5**

In fig.5 instead of GaussianNB, BernoulliNB is used. BernoulliNB is more efficient when the target variables are categorical. In the current dataset, the target variables are categorical. A model is built and accuracy is calculated using accuracy\_score function. The accuracy obtained is 80%.

**Results:**

The accuracy is calculated using two models which are GaussianNB and BernoullinNB. The accuracy obtained is 77% and 80% respectively. This is because GaussianNB works efficiently when the target variable is continuous. BernoullinNB works efficiently when the target variables are having boolean values. In the given dataset the target values are having only two types because of which using BernoullinNB we got high accuracy and the accuracy was low when GaussianNB was used.

**References:**

[1] Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. (2020, April 01). Learn Naive Bayes Algorithm: Naive Bayes Classifier Examples. Retrieved June 26, 2020, from <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

[2] Gandhi, R. (2018, May 17). Naive Bayes Classifier. Retrieved June 26, 2020, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

**Dataset:**

<http://www.timeseriesclassification.com/description.php?Dataset=ECG200>