# Naive Bayes

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Prajay B. Urkude Batch ID: 16092021**

**Topic: Naïve Bayes**

1.) Prepare a classification model using the Naive Bayes algorithm for the salary dataset. Train and test datasets are given separately. Use both for model building.



Ans:-

Business Objective :-

To categories the salary of the employees based on the inputs such as age, education, work class, capital gain etc.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Feature** | **Description** | **Type** | **Relevance** |
| Age | Age of the employee | Quantitative, Ratio | Relevance, It gives the useful information |
| Work class | Class of work, Private or Self employee. | Qualitative, Nominal | Relevance, It gives the useful information |
| education | Education of the employee | Qualitative, ordinal | Relevance, It gives the useful information |
| education | Grads based on the education | Quantitative, ordinal | Relevance, It gives the useful information |
| marital status | Marital status of the employee | Qualitative, Nominal | Irrelevance, It does not gives the useful information |
| occupation | Occupation of the employee | Qualitative, ordinal | Relevance, It gives the useful information |
| relationship | Family member of the employee | Qualitative, Nominal | Irrelevance, It does not gives the useful information |
| race | Race of the employee | Qualitative, Nominal | Irrelevance, It does not gives the useful information |
| sex | Gender of the employee | Qualitative, Nominal | Irrelevance, It does not gives the useful information |
| capital gain | Capital gain by the employee | Quantitative, Ratio | Relevance, It gives the useful information |
| capital loss | Capital loss by the employee | Quantitative, Ratio | Relevance, It gives the useful information |
| Hours per week | Total hours per week work done | Quantitative, Ratio | Relevance, It gives the useful information |
| Native | Nation of the employee | Qualitative, Nominal | Irrelevance, It does not gives the useful information |
| Salary | Salary label of the employee whether it is >50k or <50K | Quantitative, Nominal | Relevance, It gives the useful information |

Step1:

* Import the libraries like pandas, numpy, sklearn.
* Import the CountVectorizer and TfidfTransformer from Sklearn.feature\_extraction.text library.

CountVectorizer = It is used to convert the collection of text documents to a matrix of tokens counts.

TfidfTransformer = It is used to convert the collection of raw documents to a matrix of TF-IDF features.

* Load the datasets
* Data preprocessing and Exploratory data analysis like checking the null values, dropping the unwanted columns from the datasets , conversion of non numeric data to numeric data etc.
* Then Normalization of the data is to be done to bring the data in the scale between 0 to 1.

Step2:

Preparation of the data into naïve\_baysclassifier model

* Then splitting the data into train and test. As in the given datasets train data and test data is given separately so we don’t need this step.
* Preparing a naïve\_bayes model on training datasets . Import the sklearn.naive\_bayes library and from that import Multinomoal function.
* Initialising the function and fit for the training datasets by considering the smoothing process to solve the emergence of the zero probablility case by taking alpha = 2
* Evaluation of the model on the test datasets and finding the accuracy of the model. Here we get the 76% of the accuracy of the model .
* Draw the crosstab table and check it manually
* Check the accuracy of the Training model and it is found out upto 76%.
* Model is good but the accuracy of the model id only 76% .

**Problem Statement: -**

This dataset contains information of users in a social network. This social network has several business clients which can post ads on it. One of the clients has a car company which has just launched a luxury SUV for a ridiculous price. Build a Bernoulli Naïve Bayes model using this dataset and classify which of the users of the social network are going to purchase this luxury SUV. 1 implies that there was a purchase and 0 implies there wasn’t a purchase.

A screenshot of a cell phone

Description automatically generated

Ans:- Business objective:-

To classify which of the users of the social network are going to purchase luxury SUV

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Feature** | **Description** | **Type** | **Relevance** |
| User ID | ID of the person | Quantitative, Nominal | Irelevant, it does not give useful information |
| Gender | Gender of the person | Qualitative, Nominal | Relevance , It gives useful information |
| Age | Age of the person | Quantitative, ratio | Relevant, It gives useful Information |
| EstimatedSalary | Salary of the person | Quantitative, ratio |  |
| Purchased | The person purchased the Suv or not | Qualitative, Nominal | Relevant, It gives useful information |

Step1:

* Import the libraries like pandas, numpy, sklearn.
* Import the CountVectorizer and TfidfTransformer from Sklearn.feature\_extraction.text library.

CountVectorizer = It is used to convert the collection of text documents to a matrix of tokens counts.

TfidfTransformer = It is used to convert the collection of raw documents to a matrix of TF-IDF features.

* Load the datasets
* Data preprocessing and Exploratory data analysis like checking the null values, dropping the unwanted columns from the datasets , conversion of non numeric data to numeric data etc.
* Then Normalization of the data is to be done to bring the data in the scale between 0 to 1.

Step2:

Preparation of the data into naïve\_bays classifier model

* Then splitting the data into train and test. As in the given datasets train data and test data is given separately so we don’t need this step.
* Preparing a naïve\_bayes model on training datasets. Import the sklearn.naive\_bayes library and from that import **GaussianNB** function.
* Initialising the function and fit for the training datasets.
* Evaluation of the model on the test datasets and finding the accuracy of the model. Here we get the 88% of the accuracy of the model.
* Check the accuracy of the Training model and it is found out upto 88% by using cross validation score function.

**Problem Statement: -**

In this case study, you have been given Twitter data collected from an anonymous twitter handle. With the help of a Naïve Bayes model, predict if a given tweet about a real disaster is real or fake.

1 = real tweet and 0 = fake tweet

A screenshot of a cell phone

Description automatically generated

Ans:- Business objective:

To identify the tweet real or fake

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Feature** | **Description** | **Type** | **Relevance** |
| Id | Id of the member | Quantitative, Nominal | Irelevant, It does not give useful information |
| keyword | Keywords from the tweet | Qualitative, Nominal | It gives the useful information |
| location | Location of the person who tweeted | Qualitative, Nominal | It gives the useful information |
| target | Its shows the tweet real of fake | Quantitative , Nominal | It gives the useful information |

Step1:

* Import the libraries like pandas, numpy, sklearn.
* Import the CountVectorizer and TfidfTransformer from Sklearn.feature\_extraction.text library.

CountVectorizer = It is used to convert the collection of text documents to a matrix of tokens counts.

TfidfTransformer = It is used to convert the collection of raw documents to a matrix of TF-IDF features.

* Load the datasets
* Data preprocessing and Exploratory data analysis like checking the null values, dropping the unwanted columns from the datasets, conversion of nonnumeric data to numeric data etc and cleaning of the text data to be done by removing the stopwords, special symbols, numbers etc.

Step2:

* converting the collection of all the documents in the matrix of all the documents and preparation of the bag of words for all messages and
* learning term weighing and normalizing on entire text data by using the TfidfTransformer.

Step2:

Preparation of the data into naïve\_baysclassifier model

* Then splitting the data into train and test. As in the given datasets train data and test data is given separately so we don’t need this step.
* Preparing a naïve\_bayes model on training datasets. Import the sklearn.naive\_bayes library and from that import Multinomoal function.
* Initialising the function and fit for the training datasets by considering the smoothing process to solve the emergence of the zero probablility case by taking alpha = 2
* Evaluation of the model on the test datasets and finding the accuracy of the model. Here we get the 76% of the accuracy of the model.
* Draw the crosstab table and check it manually
* Check the accuracy of the Training model and it is found out upto 84%.
* Model is good but the accuracy of the model id only 84%.