# 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: DHEERAJ MISHRA Batch ID:** DS\_01072021

**Topic: Naïve Bayes**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Build a Naïve Bayes model.**

**5.3 Validate the model with test data and obtain a confusion matrix, get precision, recall, and accuracy from it.**

**5.4 Tune the model and improve the accuracy**

**6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statement:**

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. 

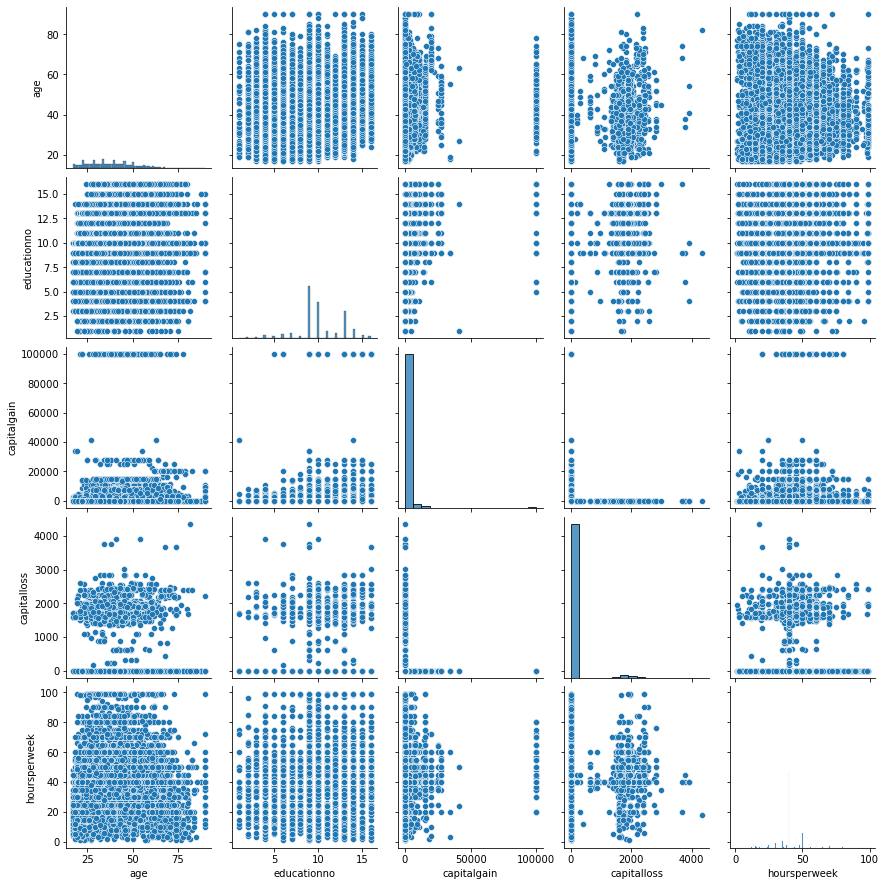
1. BUSINESS OBJECTIVE:-

Maximize mapping of salary

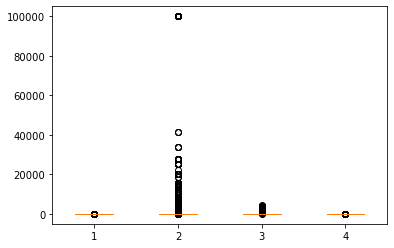
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| age | Age of employers | Discrete | Relevant |
| workclass | Work class of employers | Char | Relevant |
| education | Employers education | Char | Relevant |
| educationno | Employers education code | Discrete | Not relevant |
| maritalstatus | Employers married or not | Char | Relevant |
| occupation | Employers occupation | Char | Relevant |
| relationship | Employers relationship | Char | Relevant |
| race | Employers face | Char | Relevant |
| sex | Employers sex | Char | Relevant |
| capitalgain | Employers capital gain | Discrete | Relevant |
| capitalloss | Employers capital loss | Discrete | Relevant |
| hoursperweek | Working hours per week | Discrete | Relevant |
| native | Employers native place | Char | Relevant |
| Salary | Employers salary class | Discrete , char | Relevant |

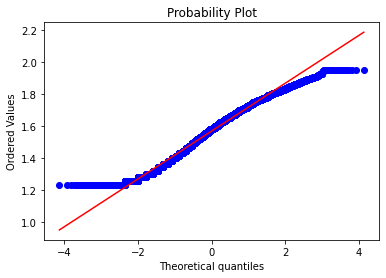
1. DATA CLEANSING :-
2. From test and train datasets merged to single datasets
3. Dataset consists of 14 colums and 39239 rows
4. Dropping nominal column education no
5. Obtained duplicates rows and removed
6. All data types are of form int64 and object
7. No null values found in each column
8. From describe function mean , median and standard deviation obtained
9. Outliers detected but retained
10. Log transformation used for normal distribution
11. Dummy column obtained for object columns
12. Scaling is done through normalization techniques
13. EDA:-
14. From pair plot analysis

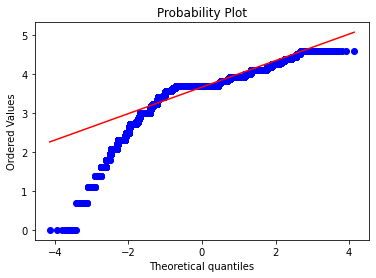


1. Box plot for outliers



1. **QQ plot for normal distribution**





1. MODEL BUILDING:-
2. Splitting data to train part for 75% and test part for 25%
3. Model builded by Multinomial nave bayes
4. Test accuracy = 0.8150866462793068
5. Train accuracy = 0.8080464847599307
6. Laplace transformation for alpha = 2 for zero probablity
7. Laplace test accuracy = 0.8148827726809378
8. Laplace train accuracy = 0.8084542458119542
9. BENEFITS :-

From above information we can assign the salary class on basis of new dataset .

**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.

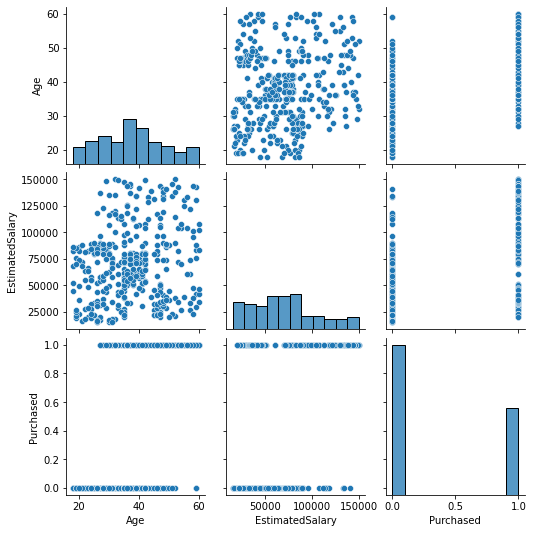
1. BUSINESS OBJECTIVE:-

Maximize mapping of purchasing

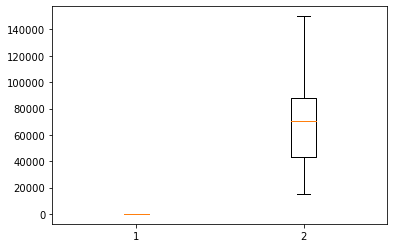
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| User ID | Id of customer | Discrete | Not relevant |
| Gender | Gender of customer | Char | Relevant |
| Age | Age of customer | Discrete | Relevant |
| EstimatedSalary | Estimated salary of customer | Discrete | Relevant |
| Purchased | Customer purchased car or not | Binary | Relevant |

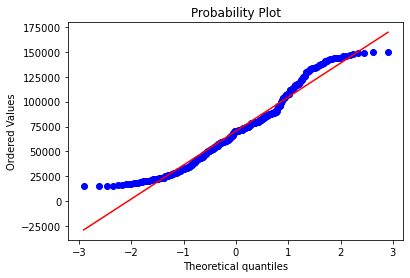
1. DATA CLEANSING :-
2. Dataset consists of 5 colums and 400 rows
3. Dropping nominal column user id
4. Obtained duplicates rows and removed
5. All data types are of form int64 and object
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers not detected
9. Log transformation used for normal distribution
10. Dummy column obtained for object columns
11. Scaling is done through normalization techniques
12. EDA:-
13. From pair plot analysis

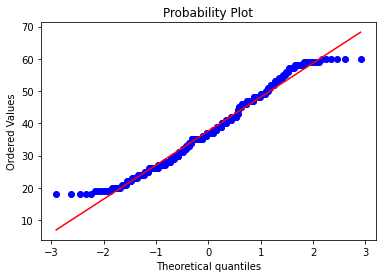


1. Box plot for outliers



1. **QQ plot for normal distribution**





1. MODEL BUILDING:-
2. Splitting data to train part for 75% and test part for 25%
3. Model builded by Guassian nave bayes
4. Test accuracy = 0.8421052631578947
5. Train accuracy = 0.8491228070175438
6. BENEFITS :-

From above information we can predict whether car is purchased or not on basis of new dataset .

**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
2. BUSINESS OBJECTIVE:-

Maximize mapping tweet

1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| ID | Id of customer | Discrete | Not relevant |
| keyword | Keyword in a tweet | Char | Not relevant |
| location | Location of tweet | Char | Not relevant |
| text | Text of tweet | Char , strings | Relevant |
| target | Dimension of output | Binary | Relevant |

1. DATA CLEANSING AND EDA:-
2. Dataset consists of 5 colums and 7613 rows
3. Dropping nominal column id , keyword and location
4. Duplicate rows found and removed
5. All data types are of form int64 and object
6. No null values found
7. Stop words removed
8. From regular expression text column cleaned
9. All punctuations removed by selecting words greater than three
10. Empty rows removed
11. Count vectorizer applied for bag of words
12. Tokenization for splitting in to words
13. TFIDF applied for weightage
14. MODEL BUILDING:-
15. Splitting data to train part for 75% and test part for 25%
16. Model builded by Multinomial naive bayes
17. Test accuracy = 0.77
18. Train accuracy = 0.89
19. Laplace test accuracy = 0.73
20. Laplace train accuracy = 0.82
21. BENEFITS :-

From above information we can predict dimension of output on basis of new dataset .