Complete Flowchart - Gaussian Naive Bayes

- Start
  - Impdt necessary libraries
  - Reading the dataset
  - Explorotory data Analysis
    - check whether dataset is balanced of imbalanced
    - check information about dataframe
    - check description of dataframe
    - find shape of dataframe
    - checking first and last 5 rows
    - Checking data types
  - Data preprocessing
    - Missing values and outlier detection
    - Outlier treatment (using IQR)
    - Missing value treatment (mode, median)
    - Dropping unnecessary columns
    - Conversion of categorical to numerical
    - Conversion of data types
  - Data Visualization.
  - Split the training and testing sets
  - Define the Gaussian Naive Bayes function
    - Calculate prid probability of each class
    - Calculate mean and variance of each feature for each class
    - Define function to calculate Gaussian probability density function
    - Define function to calculate class posterior probability for each test instance
  - Training Gaussian Naive Bayes classifier using training set
  - Make predictions on testing set
  - Evaluate performance of model
- End

Calculations

→ Let us consider a sample dataset as follows:

Age	Income	Buy product
20	20000	yes
25	30000	yes
30	25000	yes
35	40000	yes
40	50000	yes
45	60000	NO
50	70000	NO
55	80000	NO
60	90000	NO
65	100000	NO

Now, let's use Gaussian Naive Bayes to classify whether a person is likely to buy product or not based on age and income.

* Step-1: Calculate prid probability

P(Yes) = No. of yes / Total no. of records = $\frac{5}{10}$ = 0.5

P(NO) = No. of NO / Total no. of records = 0.5

* Step-2: Calculate Mean and standard deviation

For Age:

Class	Mean	standard deviation
Yes	30	3.95
NO	55	3.95

For Income:

class	Mean	Standard deviation
Yes	33000	169.55
No	80000	250

* Step-3: Calculate likelihood probability

for example, let's consider first record:

Age	Income	By product
20	20000	yes

We need to calculate likelihood probability of this record belonging to 'yes' and 'No' class.

for "Yes" class;

$$P(Age = 20 | Yes) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} = 0.002$$

$P(Income = 20000 | Yes) = 2.153$

Now;

$P(Yes | Age = 20, Income = 20000) = P(Yes) * P(Age = 20 | Yes) * P(Income = 20000 | Yes)$

$= 0.5 \times 0.002 \times 2.153$

$= 0.0021$

for "No" class;

$P(Age = 20 | No) = 5.075 \times 10^{-19}$

$P(Income = 20000 | No) = 0.0001$

$P(No | Age = 20, Income = 20000) = P(No) * P(Age = 20 | No) * P(Income = 20000 | No)$

$= 0.5 \times 0.0001 \times 5.075 \times 10^{-19}$

$= 0.0002 \times 10^{-19}$

* Step-4: Make predictions

for instance, for the first record;

$P(Yes | Age = 20, Income = 20000) = 0.0021$

$P(No | Age = 20, Income = 20000) = 0.0002 \times 10^{-19}$

Since; $P(Yes | Age = 20, Income = 20000) > P(No | Age = 20, Income = 20000)$

We predict that this person is likely to buy the predict.

Similarly, we can predict for other records and can see that Gaussian Naive Bayes algorithm is correctly predicting the outcome or not.