***RANDOM REPORT M3 [NETW504]*** *by Khaled Emad t-3 58-4604, Omar El Henawy 58-4941 t-3 , Mostafa El Shinawi 58-1407 t-2*

This project implements two approaches for anomaly detection using Naïve Bayes classifiers on numerical and categorical data from our dataset. The two main tasks are:

Custom Naïve Bayes Classifier:

This classifier calculates probability for anomalies using fitted probability density functions (PDFs) for numerical features and probability mass functions (PMFs) for discrete features.

Scikit-Learn Naïve Bayes Classifiers:

Three standard Naïve Bayes classifiers (GaussianNB, MultinomialNB, BernoulliNB) are applied using pipelines that preprocess categorical data through one-hot encoding.  
  
Custom Naïve Bayes Approach

PDF Fitting for Numerical Data:

For each numerical column, the best distribution is fitted using scipy.stats distributions (norm, lognorm, gamma, beta, expon) conditioned on the class ('anomaly' or 'normal').

PMF Calculation for Discrete Data:

For discrete columns (with unique values ≤ 10), probability mass functions are computed.

Prediction Function:

The classifier calculates probabilities for each test row based on the fitted PDFs/PMFs and assigns the label ('anomaly' or 'normal') with the highest probability.

Metrics:

Accuracy, Precision, and Recall are calculated based on predictions.

We applied three types of standard Naïve Bayes classifiers from Scikit-Learn:

GaussianNB: Suitable for continuous numerical data.

MultinomialNB: Designed for multinomially distributed data, typically used for categorical features.

BernoulliNB: Best suited for binary/boolean features.  
  
Results:

Naïve Bayes classifier:

Accuracy: 0.75

Precision: 0.66

Recall: 0.97  
The Naïve Bayes classifier achieves a high recall, indicating it successfully identifies most anomalies, but the lower precision shows that some normal instances are incorrectly classified as anomalies.

Scikit-Learn Classifiers

GaussianNB:

Accuracy: 0.56

Precision: 0.77

Recall: 0.07

GaussianNB performs poorly in recall, suggesting it misses most anomalies, though the precision is reasonably high.

MultinomialNB :

Accuracy: 0.53

Precision: 0.44

Recall: 0.04

MultinomialNB shows low accuracy, precision, and recall, indicating it is not well-suited for this dataset.

BernoulliNB :

Accuracy: 0.91

Precision: 0.94

Recall: 0.85

BernoulliNB performs the best overall, achieving high accuracy, precision, and recall. This suggests it is effective at identifying anomalies while minimizing false positives.

Conclusion:

The code begins by loading and preprocessing the dataset, specifically splitting it into numerical and categorical features. The categorical features are encoded using one-hot encoding to convert them into numerical form, allowing compatibility with Naïve Bayes models. The core part of the code implements a Naïve Bayes classifier to perform anomaly detection. For numerical features, the classifier calculates probabilities using Probability Density Functions (PDFs) under the assumption that the data follows a Gaussian (normal) distribution. For categorical features, the classifier uses Probability Mass Functions (PMFs) to calculate probability for each possible category.

The Naïve Bayes classifier achieves a good balance with high recall (0.97), making it useful when catching all anomalies is critical.

BernoulliNB from Scikit-Learn outperformed the other standard classifiers with an accuracy of 0.91 and a precision of 0.94, making it a reliable choice for this dataset.

GaussianNB and MultinomialNB were less effective, particularly in recall, indicating limitations with the dataset's structure for these models.