**Predicting Brain Tumor Presence using Gaussian Naïve Bayes**



**Student Name:** Tharun Dulam

**Batch:** AI Elite 16



**Project Name:** Brain Tumor Prediction using Gaussian Naïve Bayes

**Domain:** Medical

**Type of ML:** Supervised ML

**Type of Problem:** Classification

**Project Methodology:** CRISP-ML (Q)



**Phase I: Business and Data Understanding**

* **Business Understanding:  
    
  Business Objective:** The objective of this project is to develop a machine learning model that can accurately predict the presence of brain tumors based on certain medical features. This will aid in early detection and treatment planning for patients with brain tumors.

**Constraints:**

* Availability of labeled medical data for training the model.
* Model should provide accurate predictions with minimal false positives and false negatives.

**Success Criteria:**

* **ML success criteria**: Achieve an average accuracy of 85% in predicting brain tumors using machine learning algorithms.
* **Business Success criteria**: Ensure timely and accurate diagnosis of brain tumors to improve patient outcomes and satisfaction.
* **Economy Success criteria**: Minimize healthcare costs associated with brain tumor diagnosis and treatment through efficient and effective prediction methods.
* **Data Understanding:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Feature Name** | **Data Type** |
| 1 | Area | Numerical |
| 2 | Perimeter | Numerical |
| 3 | Convex Area | Numerical |
| 4 | Solidity | Numerical |
| 5 | Equivalent Diameter | Numerical |
| 6 | Major Axis | Numerical |
| 7 | Minor Axis | Numerical |
| 8 | Eccentricity | Numerical |
| 9 | Class | Categorical |
|  |  |  |

**Phase 2: Data Preparation  
  
a) Exploratory Data Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type** | **Feature Names** | **Observation** |
| 1 | Missing Values | NA | NA |
| 2 | Duplicates | All | 51 |
| 3 | Outliers | All Except Class | 18 for each Feature |
| 4 | Distributions | All Except Class | Right skewed |
| 5 | Noisy data | NA | NA |
|  |  |  |  |

**b) Data Cleaning/wrangling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Type of Cleaning** | **Technique** | **Feature Name** | **Reason** |
| 1 | Missing value | None | None | No outliers |
| 2 | Encoding | Label | Class | nominal |
| 3 | Scaling | Standard Scaling | All Features except Class | Normalization |
| 4 | Outliers | Winsorization | All Features except Class | Robustness |
| 5 | Duplicate Removal | drop\_duplicates | All | Bias |
|  |  |  |  |  |

**Phase 3: Model Building:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type of Problem** | **Approach** | **Algorithm Name** |
| 1 | Classification | Probability Based | Gaussian Naïve Bayes |

**Phase 4: Model Evaluation:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Metric Name** | **Metric Score** | **Hyper Parameters** |
| 1 | Accuracy | 75 | Alpha = 4 |
| 2 | Precision | 88 | Alpha = 4 |
| 3 | Recall | 82 | Alpha = 4 |
| 4 | F1-Score | 85 | Alpha = 4 |

**Phase 5: Model Deployment:**

**Deployment Platform:**  Stream lit

**Link/URL:** provide the link

**A Comprehensive Analysis**

Data Preparation

During this phase, the dataset underwent thorough cleaning and preprocessing to ensure its suitability for model training.

Exploratory Data Analysis (EDA):

* The dataset was subjected to EDA to understand its structure and identify any anomalies.
* No missing values were found, indicating a complete dataset.
* 51 duplicate rows were identified and subsequently removed to prevent bias in model training.
* Outliers were detected in numerical features and treated using winsorization, maintaining data integrity while handling extreme values.
* Distributions of numerical features were examined, revealing right-skewed distributions.
* No significant noisy data or anomalies were observed, contributing to the dataset's reliability.

Data Cleaning/Wrangling:

* No missing values required imputation, and duplicates were successfully removed.
* Winsorization effectively addressed outliers, ensuring their impact on model performance was minimized.
* Categorical encoding was applied to the 'Class' feature for interpretability.
* Standard scaling was performed on numerical features to standardize their ranges, facilitating effective model training.

Feature Selection:

All features were retained for subsequent model training and evaluation, as no explicit feature selection was performed.

Model Building

In this phase, the cleaned and preprocessed dataset was used to train a machine learning model.

Model Building Approach:

* The problem was identified as a classification task.
* The Gaussian Naïve Bayes algorithm was selected for its simplicity, efficiency, and effectiveness in handling classification tasks.

Model Evaluation

The trained model underwent evaluation to assess its performance and effectiveness.

Model Performance:

* The Gaussian Naïve Bayes model demonstrated strong performance, achieving a recall score of 82% on the test dataset.
* This high recall indicates that the model can effectively predict the presence or absence of brain tumors based on the provided medical features.
* Additionally, precision and F1-score metrics were computed to provide a more comprehensive understanding of the model's performance across different evaluation criteria.

Conclusion

* Meticulous data preparation, model building, and evaluation processes culminated in the development of a robust machine learning model for brain tumor prediction.
* The model exhibited high recall and reliability in detecting brain tumors using medical features, indicating its potential to aid medical professionals in early detection and treatment planning.
* Implementation of this model holds promise for improving patient outcomes and reducing healthcare costs by facilitating timely intervention and personalized treatment strategies for individuals with brain tumors.