# Predicting The Grade OF Glioma

# On Basis Of Mutations

Submitted in partial fulfillment of the requirements for
the award of degree of

Bachelor of Technology
Computer Science and Engineering

**LOVELY PROFESSIONAL UNIVERSITY**
**PHAGWARA, PUNJAB**


**SUBMITTED BY**

Name: Kanishk Rao
Registration Number: 12015255

# Predicting The Grade OF Glioma On Basis Of Mutations


## Abstract:-

Glial cells, which nourish and insulate neurons in the brain, give birth to gliomas, which are extremely aggressive primary brain tumors. These tumors are among the most difficult to identify and manage, and how they are categorized and graded is essential to choosing the best treatment options and forecasting patient outcomes.

The standard for categorizing gliomas is the World Health Organization (WHO) classification system, with grade I denoting the least aggressive tumors and grade IV denoting the most malignant. This technique is based mostly on histological characteristics that are used to assess the level of malignancy, such as cellularity, mitotic activity, and necrosis.

Current developments in molecular biology have resulted in the identification of certain mutations that are connected to various glioma subtypes and have been added to the WHO categorization scheme. Isocitrate dehydrogenase (IDH) gene mutations and the 1p/19q co-deletion have been discovered as crucial predictors of prognosis and therapeutic responsiveness in gliomas, respectively.

We offer a comprehensive dataset of clinical and mutation aspects of gliomas, containing information on over 1000 individuals, to better clarify the importance of molecular markers in glioma grading. Age, sex, tumor site, histology, WHO grade,

IDH status, 1p/19q co-deletion, and other genetic changes are all included in our dataset.

We looked at the relationship between clinical and molecular characteristics and glioma grade using this dataset. Our findings show that the 1p/19q co-deletion and IDH gene alterations are both significantly related with higher overall survival and lower-grade cancers. Also, our research showed considerable variations in the prevalence of certain mutations among various glioma subtypes, emphasizing the use of genetic profiling in the diagnosis and treatment of gliomas.

This extensive dataset highlights the crucial importance of molecular characteristics in glioma diagnosis and therapy and offers a useful resource for academics and medical professionals interested in glioma grading. The development of individualized treatment plans that take into consideration the unique molecular characteristics of each glioma is significantly impacted by these discoveries.

## Introduction:-

Tumor grading, or the capacity to assess a glioma's biological aggressiveness, is crucial for effective management, including therapy planning and monitoring as well as patient survival rates. The most prevalent primary brain tumor that arises from glial cells is a glioma, which progresses quickly, is neurologically fatal, and is a glial cell tumor.

The standard of care for treating gliomas depends on the tumor grade, but it typically entails complete surgical removal of the tumor, followed by radiation therapy (RT), with concurrent or sequential administration of the chemotherapy drug

temozolomide (TMZ), or PCV or PC (Procarbazine, CCNU with or without vincristine) as an alternative. Gliomas can currently be broadly categorized into two groups, low-grade gliomas (LGG), and high-grade gliomas (HGG), with glioblastoma multiforme (GBM), a high-grade glioma, based on histological and molecular parameters, according to the World Health Organization (WHO) classification of Central Nervous System (CNS) tumors. The most prevalent, harmful, invasive, and main kind of tumor is a GBM. Grading is crucial in this neuro-oncology situation since glioma accounts for about 80% of all primary malignant tumors of the brain and more than 60% of all adult brain tumors.

The categorization of CNS malignancies has become more complicated in recent years as molecular changes have become more significant, and the desire for value-added treatment has complicated the debate. The vast and diverse range of molecular parameters that are available emphasize the significance of identifying and choosing the necessary molecular alterations, and two aspects emerged: the desire to lower the cost of molecular testing to enable its more widespread use and the reduction of health disparities. For instance, the isocitrate dehydrogenase (IDH) mutation is a crucial molecular indicator for differentiating between low-grade and high-grade gliomas. IDH testing has raised questions about price and turnaround time. IDH1 immunohistochemistry for p. R132H costs $135, single gene sequencing is $420, and next-generation sequencing is $1800.

Depending on the method used, turnaround time might range from two days for immunohistochemistry to fourteen days for next-generation sequencing. The process of grading tumors takes into account clinical characteristics like age and gender, but there is a dearth of higher-level robust clinical annotation in publicly available datasets, restricting the connections between relevant molecular features and clinical information that could advance value-added care as more widespread molecular testing may eventually benefit from the increase in reimbursement. Therefore,

choosing the best discriminative molecular and clinical markers not only lowers the cost to healthcare systems and patients while assisting in the reduction of growing health disparities in access to testing, but also improves tumor grading performance, which can enable the selection of pertinent molecular features for subsequent analyses and bench to bedside work as well as the testing of novel targeted agents. We hypothesize that feature selection plays a substantial role in this domain since the pattern recognition required to properly use fragmented molecular information is probably not attainable without computer analysis.

In feature selection, the optimal feature subset is chosen from all features or patterns using procedures that exclude irrelevant, redundant, and unrelated characteristics. In particular, this technique ensures the greatest performance for class prediction and lowers computational demand and expense, boosting efficiency and offering more cost-effective features, raising the classification accuracy rate, and enhancing the clarity of the findings. Nowadays, a lot of data analysis applications, pattern recognition, and mining jobs require feature selection.

# Proposed Methodology:-

(1) Data Collection:

The process of acquiring raw data from numerous sources, including databases, surveys, experiments, observations, sensors, and more, is known as data collection or acquisition. The objective is to gather information that may be utilized for modeling, analysis, and decision-making. The initial phase in the data science pipeline, data collecting, is essential for assuring the accuracy and applicability of the analysis. In this case Glioma Grading Clinical and Mutation Features Dataset is used

.

(2) Data Preprocessing:

Preparing and cleaning data to make it ready for analysis is referred to as data preparation. It is a crucial stage in the pipeline for data analysis and can have a big impact on the precision and efficacy of any subsequent study.

The following are typical steps in data preprocessing:

Data cleaning is eliminating or fixing any mistakes or discrepancies in the data, including missing values, duplicate entries, or inaccurate data types.

Data transformation entails converting the data into a format that is more suited for analysis. This might entail lowering the dimensionality of the data using methods like feature selection or extraction, scaling or normalizing the data, or turning categorical variables into numerical representations.

Data integration is the process of gathering information from several sources and merging it into one dataset for analysis. This might entail dealing with errors or duplication in the data, as well as reconciling discrepancies in data type or structure.

Data reduction entails shrinking the dataset while keeping as much information as you can. Techniques like sampling or data summarization may be used in this.

Data discretization is the process of dividing continuously changing data into distinct groups or bins. For some analyses or models, such decision trees or rule-based systems, this could be helpful.

Data preparation is, in general, a crucial stage in any data analysis project and may have a big impact on the precision and efficacy of the outcomes

(3) Method Suggestion :

Our proposed model is based on a RandomForestClassifier, AdaBoostClassifier, DecisionTreeClassifier, KNeighborsClassifier, GradientBoostingClassifier, SVC

(4) Model performance:

The examination of a machine learning model's capacity to produce precise predictions on unobserved data is known as model performance. It is a crucial step in the machine learning process that aids in determining a model's strengths and weaknesses and can direct changes to the model's architecture.In this case we will compare the results of RandomForestClassifier, AdaBoostClassifier, DecisionTreeClassifier, KNeighborsClassifier, GradientBoostingClassifier, SVC to see which one of the classifier is best suitable and accurate for the given dataset

## (1) Data Collection:-

Researchers and clinicians interested in the genetic and clinical characteristics of gliomas will find the Glioma Grading Clinical and Mutation Features Dataset to be an invaluable resource. This data set includes clinical and genetic information for 262 glioblastoma patients, including patient age, gender, tumor location, histological grade, and genetic mutations.

The magnitude and variety of this dataset are among its virtues. This dataset provides a comprehensive view of the clinical and genetic characteristics of gliomas, with information on over 260 patients. In addition, the dataset contains patients with both high-grade and low-grade gliomas, allowing researchers to examine the distinctions between these tumor types.

The level of detail provided on the genetic mutations present in each patient's tumor is another asset of this dataset. The dataset includes information on mutations in TP53, IDH1, and ATRX, which are among the most frequently mutated genes in gliomas. This data can be utilized to investigate the association between specific mutations and the development and progression of gliomas.

Nonetheless, there are limitations to this dataset. For instance, the dataset only contains patients from a single institution, limiting its applicability to other populations. In addition, while the dataset contains valuable

information on the genetic and clinical characteristics of gliomas, it lacks information on treatment outcomes and survival rates.

The Glioma Grading Clinical and Mutation Features Dataset is an invaluable resource for researchers and clinicians interested in glioma genetic and clinical characteristics. Its magnitude and degree of genetic mutation detail make it an invaluable resource for glioma development and progression research. However, researchers should be aware of the dataset's limitations and consider including additional information on treatment outcomes and survival rates.

## (2) Data Pre-processing:-

Reading CSV files: Two CSV files are read using the pandas read_csv() function. The first file TCGA_GBM_LGG_Mutations_all.csv contains mutation data and the second file TCGA_InfoWithGrade.csv contains patient information including the tumor grade, age, and gender.

Checking for missing values: isna().sum() method is used to count the number of missing values in each column of the two dataframes. In the first dataframe (mutation data), the number of missing values is counted and printed.

Dropping rows with missing values: dropna() method is used to drop rows with missing values from the first dataframe (mutation data).

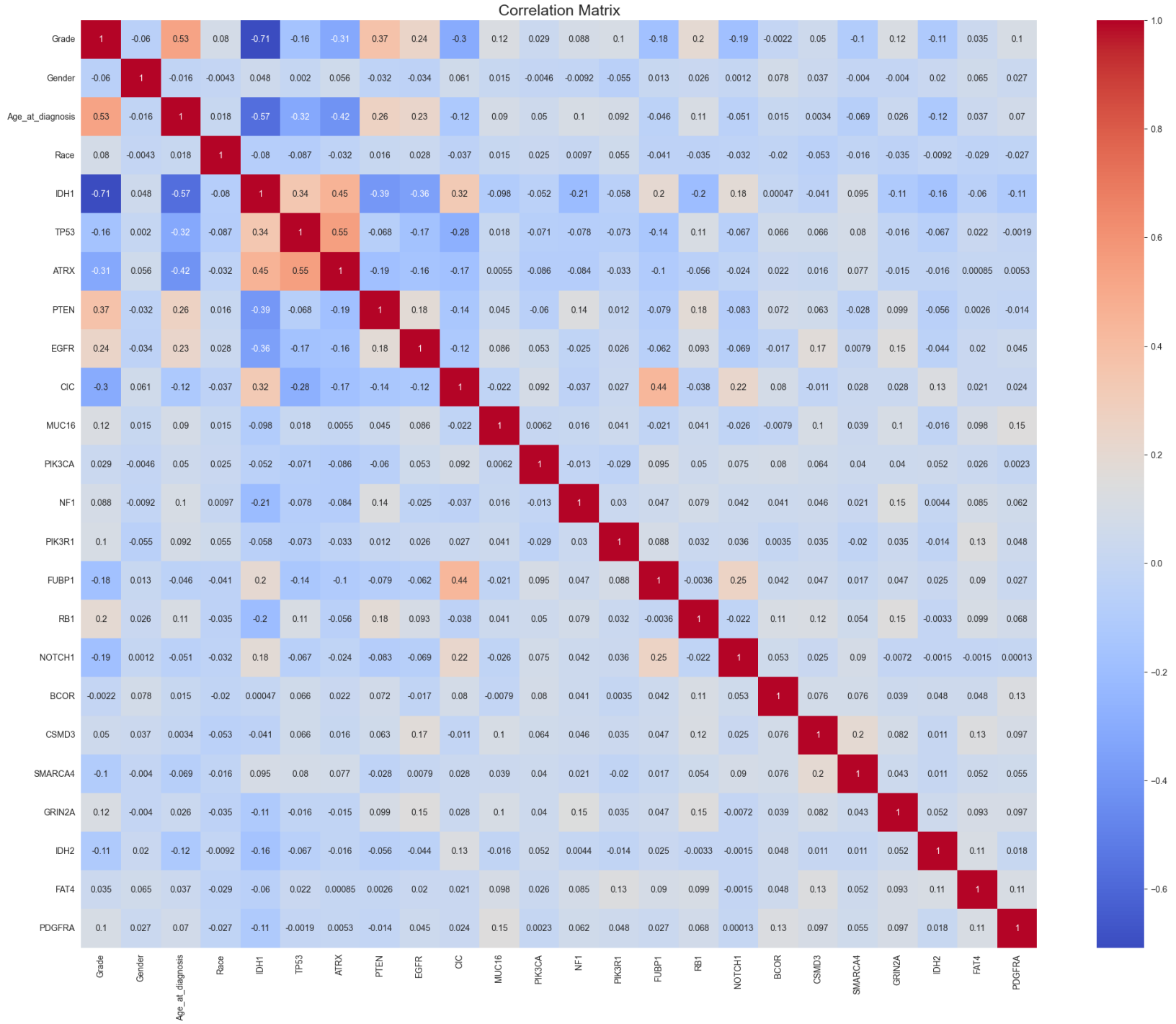
Feature scaling: MaxAbsScaler() is used to scale the values of the features in the second dataframe (patient information). The scaled dataframe is stored in df1_scaled_df. The purpose of feature scaling is to bring all the features to the same level of magnitude, which is required by some machine learning algorithms.

Inverse scaling function: The inverse scaling function is extracted using the inverse_transform() method of the scaler object. This function can be used to convert the scaled features back to their original values.

Overall, the data preprocessing steps are intended to clean and prepare the data for machine learning analysis by dealing with missing values and scaling the features.

## 2.1 Dataset Observations:-

## Correlation Matrix:-



The association between several attributes connected to a medical dataset is represented by this correlation matrix. The matrix has 20 rows and 20 columns and is square in form. The correlation coefficient between the two qualities corresponding to the row and column indices is represented by the values in each matrix cell. The correlation coefficient ranges from -1 to 1, where -1 denotes a perfect negative connection, 0 denotes a complete lack of association, and 1 denotes a complete positive correlation.

The grades are represented in the first row of the matrix, and their relationship to the other qualities is shown in the first column. The correlation between grade and gender is 0.06, which indicates a somewhat negative association. The correlation between the grade and at diagnosis is 0.53, which denotes a moderately positive association. The correlation between grade and race is 0.08, which indicates a somewhat favorable association. The correlation between the grade and IDH1 is -0.71, indicating a very adverse connection. The correlation between the grade and TP53 is 0.16, which indicates a negligible association. The correlation between the grade and ATRX is -0.31, which suggests a somewhat negative connection. The correlation between the grade and PTEN is 0.37, which denotes a moderately positive association. The correlation between grade and EGFR is 0.24, showing a somewhat favorable association. The correlation between the grade and CIC is -0.3, which denotes a sluggishly negative association. The correlation between the grade and MUC16 is 0.12, showing a somewhat favorable link.

The gender property indicates the gender of the patient, with 1 denoting a male and 2 denoting a female. The first column of the matrix displays the gender's relationship to the other characteristics. The correlation between gender and grade is 0.06, which indicates a somewhat negative association. At diagnosis has a correlation with gender of -0.016, meaning there is no relationship. The correlation between gender and race is -0.0043, which means there is no relationship. The correlation between gender and IDH1 is 0.048, which shows a somewhat favorable connection. The correlation between gender and TP53 is 0.002, which means there is no association. The correlation between gender and ATRX is 0.056, which shows a somewhat favorable association. The correlation between gender and PTEN is -0.032, which shows a somewhat negative connection. The correlation between gender and EGFR is -0.034, which suggests a somewhat negative connection. While grades and IDH1, TP53, ATRX, and CIC have negative correlations, the correlation between gender and CIC is 0.061, showing a weakly positive link. IDH1 exhibited the largest negative connection with grade, which was predicted given that IDH1 mutations are known to be linked to improved prognosis and lower grade cancers (more specifically, IDH1 mutations

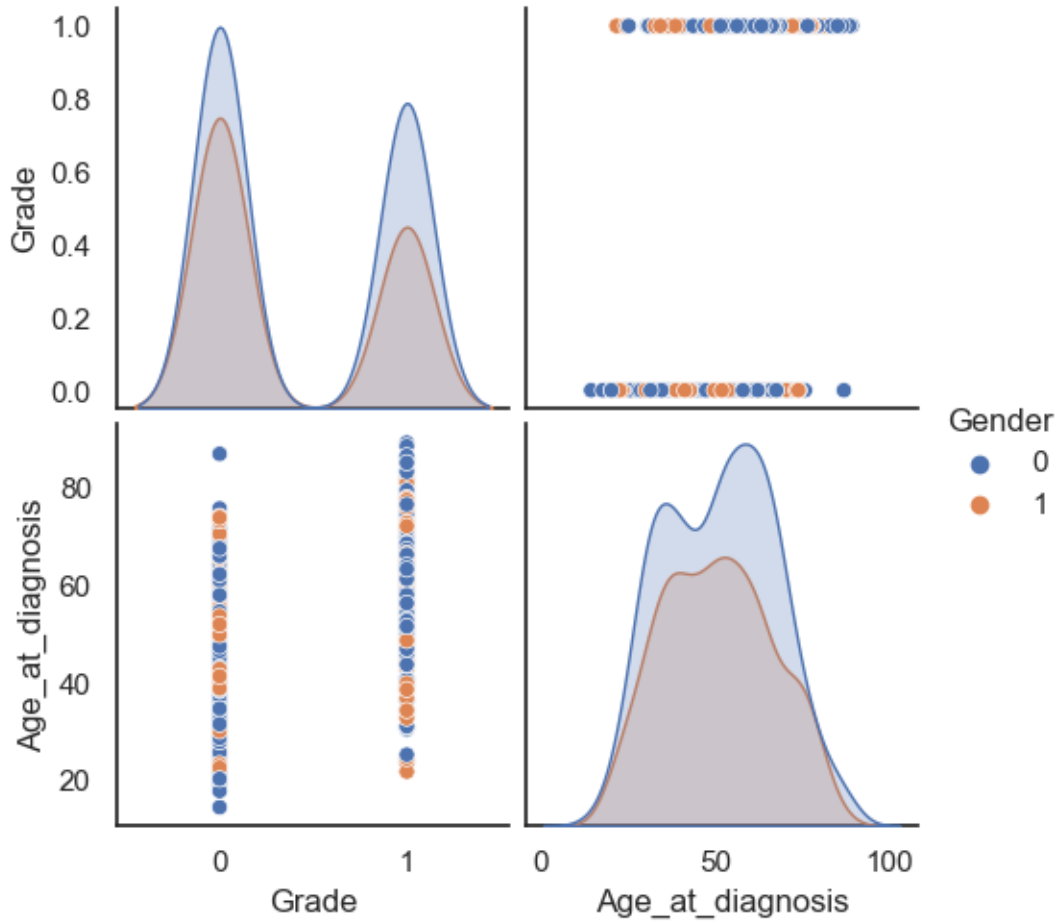
TP53 also showed a substantial negative connection with grade, which is consistent with other research demonstrating that TP53 mutations are linked to higher-grade malignancies and a poorer prognosis Moreover, there was a negative connection between grade and ATRX and CIC, suggesting that these gene alterations may be linked to higher-grade malignancies.

Given that EGFR mutations are often linked to lower-grade cancers, it is interesting that EGFR exhibited a positive connection with grade (Louis et al., 2016). The correlation coefficient between EGFR and grade was, however, somewhat low, and more research would be required to ascertain the nature of this association.

Race exhibited relatively modest connections with a few factors, and there were no substantial relationships between gender and any of the other variables. This shows that in this specific dataset, gender and ethnicity may not be reliable indicators of tumor grade or molecular characteristics.

Overall, the correlation matrix offers useful information on the connections between different variables in the dataset. A greater understanding of the nature and scope of these associations would require more research because correlation does not always indicate causation. While analyzing the findings of the correlation analysis, it is also important to consider the sample size and unique features of the dataset.

Pair plot between Grade, Gender and Age at diagnosis:-




An illustration of pairwise relationships between variables in a dataset is called a pair plot. Each scatterplot in the resulting grid illustrates the association between two different variables. The distribution of each variable is displayed in the diagonal graphs.

The figure is produced using the df1 dataframe as input. The hue option is set to "Gender," which denotes that the plot's colors should be determined by the

dataframe's gender field. When the vars argument is set to ["Grade","Age at diagnosis"], the plot will only show the variables Grade and Age at diagnosis.

As a consequence, a scatterplot of the Grade and Age at diagnosis variables is displayed, with the dots colored in accordance with the Gender variable. The distribution of each variable is displayed in the diagonal graphs.

We can visually examine the relationships between the variables thanks to the plot. For instance, the fact that the points tend to cluster in the plot's upper-right quadrant indicates that Age at diagnosis and Grade have a positive association. Also, the figure shows that there are more blue than orange dots, indicating that the dataset has more males than females. In the Age at diagnosis vs. Grade plot, we can also see that there is considerable overlap between the male and female points, which shows that gender may not be a major predictor of either variable.

## (3) Method Suggestion:-

RandomForestClassifier:-

Based on Decision Trees, RandomForestClassifier is an ensemble learning method. Many decision trees are produced in this technique, and the ultimate result is chosen based on the average of all the trees' forecasts. Each split in the tree is based on the best split available in a random subset of the features, and each tree is trained on a different subset of the data.

Since it can handle complicated classification issues and high-dimensional data, RandomForestClassifier is a well-liked classification technique.

AdaBoostClassifier:-

Another ensemble learning technique built on Decision Trees is AdaBoostClassifier. This technique employs an iterative process to generate several decision trees, with each new tree aiming to fix the flaws of the prior trees. In order to lower the overall classification error, AdaBoostClassifier gives instances that were incorrectly categorized larger weights and utilizes those weights to train the following tree. The weighted aggregate of each tree forecast makes up the final prediction. Another well-liked classification system is AdaBoostClassifier, which is very useful for issues with uneven data.

DecisionTreeClassifier:-

This classification technique builds a model of decisions and their outcomes that resembles a tree. Recursively dividing the data into smaller groups depending on the characteristic that yields the greatest information gain results in the model. In order for the tree to generalize well to new data, it must have high accuracy and be simple to construct. Little to medium-sized datasets can benefit from using DecisionTreeClassifier, which performs best when the data is well-structured and has few characteristics.

KNeighborsClassifier:-

It is a classification technique that bases its decisions on the separation between a data point's input and its k nearest neighbors. Based on the majority class of the input data point's closest neighbors, the algorithm classifies the input data point. K's value is a hyperparameter that may be adjusted to enhance the model's accuracy. For small to medium-sized datasets, KNeighborsClassifier can be useful and easy to use, but big datasets can be computationally costly.

GradientBoostingClassifier:-

This ensemble learning technique also uses Decision Trees as its foundation. This method builds a decision tree after each iteration in an effort to fix the flaws in the previous tree. In order to minimize the total classification error, the technique optimizes the loss function using a gradient descent method. The weighted aggregate of each tree forecast makes up the final prediction. High-dimensional data may be handled with GradientBoostingClassifier, which is effective for complicated classification issues but can be computationally costly.

SVC:-

A classification technique known as SVC (Support Vector Machine Classifier) determines the hyperplane that best divides the input data into distinct groups. In order to improve the model's generalization to new data, the approach optimizes the distance between the hyperplane and the closest data points. kernel functions to solve non-linear classification issues. With big datasets, SVC can be computationally costly.

# Observations and Conclusion:-

| Classifier | Best Parameters | Accuracy score |
|---|---|---|
| Random Forest Classifier | {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300} | 0.932 |
| SVM | {'C': 1, 'gamma': 1, 'kernel': 'rbf'} | 0.9154 |
| AdaBoost Classifier | {'learning_rate': 0.5, 'n_estimators': 50} | 0.8962 |
| Decision Tree Classifier | {'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2} | 0.8508 |
| K-Nearest Neighbor Classifier | {'n_neighbors': 11} | 0.8372 |
| Logistic Regression Classifier | {'C': 0.1, 'penalty': 'l1'} | 0.8713 |

According on the accuracy ratings, the Random Forest Classifier, which has an accuracy rating of 0.932, is the best classifier. AdaBoost Classifier comes in second with an accuracy score of 89.62%, followed by SVM, which performs well with a score of 91.54%. Lower accuracy values are achieved by the Logistic Regression

Classifier, Decision Tree Classifier, and K-Nearest Neighbor Classifier. The optimal classifier, it is crucial to note, may differ based on the particular dataset and the issue at hand. While choosing a classifier, it's crucial to keep things like interpretability, computational cost, and overfitting in mind.

GITHUB LINK OF PROJECT:- https://github.com/fusemate/Glioma-Grading-Clinical-and-Mutation