A PROJECT REPORT

On

"PERSONALITY PREDICTION BY USING ML MODELS"

Submitted to
KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

Bachelor's Degree in
Computer Science

BY

PRATIK PANDEY	**2029065**
SUKRITI SHUKLA	**2029075**
ALOK RANJAN	**2029198**
CHIRAG JAYDEEP BEURA	**2029199**
AAYUSH RAGHUVANSHI	**2029201**
PURNAPRAJNA SATAPATHY	**2029211**

Under the Guidance of
Dr. HITESH MAHAPATRA

School of Computer Engineering
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR ODISHA - 751024

ACKNOWLEDGEMENT

We are profoundly grateful to **DR. Hitesh Mahapatra, ASSOCIATE PROFESSOR** for his expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion. .....

Sukriti Shukla
Pratik Pandey
Alok Ranjan
Chirag Jaydeep Beura
Aayush Raghuvanshi
Purnaprajna Satapathy

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, Odisha - 741024

Certificate

This is to certify the entitled project **"PERSONALITY PREDICTION BY USING ML MODELS"** submitted by **Sukriti Shukla, Pratik Pandey, Alok Ranjan, Chirag Jaydeep Beura, Aayush Raghuvanshiand Purnaprajna Satapathy** is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2023-2024, under our guidance.

Date: / /

(Dr. Hitesh Mahapratra)
Project Guide

ABSTRACT

This project aims to develop an innovative approach for personality prediction utilizing machine learning (ML) models. Personality traits play a crucial role in understanding human behavior, influencing various aspects of life, from career choices to interpersonal relationships. Traditional methods of personality assessment, such as surveys and questionnaires, are time-consuming and often subjective. In this study, we propose a more efficient and objective solution by leveraging ML techniques.

The project will employ a diverse dataset containing behavioral and psychological features, collected through various mediums such as social media, online activities, and self-reported data. Through extensive preprocessing and feature engineering, we aim to create a robust input for ML models. Our approach involves training and fine-tuning a variety of ML algorithms, including but not limited to neural networks, support vector machines, and ensemble methods. The models will be evaluated based on their accuracy in predicting key personality traits, such as the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).

Keywords : ML Algorithms, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism

Content

Chapter 1

Introduction

- Traditional methods of personality assessment, such as surveys and questionnaires, are time-consuming and often subjective.

- To solve the problems of the current system, an automated personality categorization system is developed, which employs data mining techniques and machine learning algorithms to categorize the personalities of various users.

- Also, techniques such as the Big Five Personality Model, Logistic Regression, Decision Tree, Support Vector Machine, KNN, Naïve bais are used.

- By detecting historical data and patterns, it is simple to identify a person's personality using new techniques, hence defeating the old system.

- Each candidate must complete the test. It has several questions, and the user must complete it to determine the Big five personality traits.

- After completing the survey, the user will be able to know his or her personality.

- This is useful in a variety of fields such as interviews, recruiting processes, government sectors etc if a user's results are acceptable.

- He or she can work in any organization that is based on personality type occupations.

- In this system, we determine each user's personality. The personality type is predicted based on the answers given by the user in the personality test.

- Users who have their personality type predicted can simply apply for jobs and learn about their personality type.

- Students can also learn about their personalities and compete in competitive exams in the same way.

The Big Five Factors on which basis the personality is being predicted are:

1. **Openness to Experience.**
2. **Agreeableness**
3. **Extraversion**
4. **Neuroticism**
5. **Conscientiousness**

Chapter 2

2.1 DATA

The dataset for personality prediction encompasses a rich array of data collected from diverse sources, providing a comprehensive understanding of individual behavior. Behavioral and psychological features are amalgamated from various mediums, including social media platforms, online activities, and self-reported data. Social media contributions encompass text-based posts, comments, and user interactions, allowing for the extraction of linguistic patterns and communication styles.

2.2 INFORMATION

The dataset consists of information about both males and females. The whole personality prediction is based on the five factors. The information spans a spectrum of the Big Five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—allowing for a nuanced understanding of individual personalities. Furthermore, demographic details may be included to explore potential correlations between personality traits and factors such as age, gender, or location. The dataset undergoes meticulous preprocessing and feature engineering to ensure its quality, relevance, and compatibility with machine learning models. This amalgamation of diverse information sources empowers the models to capture the intricacies of human behavior, enabling accurate and insightful predictions of various personality traits.

Chapter 3

Problem Statement/ Requirement Specification

Traditional methods of personality assessment, such as surveys and questionnaires, are time-consuming and often subjective.To solve the problems of the current system, an automated personality categorization system is developed, which employs data mining techniques and machine learning algorithms to categorize the personalities of various users.

3.1 Project Planning

For the project the requirement is a dataset of personality traits which help the machine to train and test on the test dataset. After getting an appropriate dataset, first data sanity and cleaning is needed to remove any absurd information or to add any missing value. Then all the models need to be trained and tested in order to find the best model to predict the personality more accurately or have the best accuracy that has been found during the testing on the test data.

3.2 System Design

We use Python as our go to language and Machine learning concepts to develop this project

Chapter 4

Implementation

4.1 Visualization

Data visualization plays a pivotal role in enhancing the understanding and interpretation of the dataset in the personality prediction project. Here are key aspects of its role:

- Exploratory Data Analysis (EDA): Data visualization is crucial during the initial stages of Exploratory Data Analysis. Visual representations such as histograms, box plots, and scatter plots help in identifying patterns, trends, and potential outliers within the dataset.
- Feature Distribution: Visualizations assist in comprehending the distribution of features related to behavioral and psychological traits. Understanding the spread and variability of these features aids in making informed decisions about feature engineering and preprocessing.
- Correlation Analysis: Visualization tools, like correlation matrices and heatmaps, facilitate the exploration of relationships between different features. This is vital for understanding which features might be strongly correlated and could influence the prediction models.
- Temporal Trends: For data collected over time, visualizations such as time-series plots can reveal temporal trends and patterns in behavior, contributing to a more nuanced understanding of personality traits as they evolve.

In summary, data visualization serves as a powerful tool throughout the project lifecycle, aiding in exploration, preprocessing, model development, and communication of results. It transforms raw data into actionable insights, facilitating more informed decision-making in the context of personality prediction using machine learning.

4.2 Normality Test

A normality test is often conducted to assess whether a given dataset follows a normal distribution. Normality is a key assumption in many statistical

analyses, and deviations from normality might influence the reliability of certain statistical tests.

It's important to note that normality tests are sensitive to sample size. With large sample sizes, even small departures from normality may lead to statistically significant results. Therefore, the interpretation of normality tests should be considered along with the context of the data and the specific requirements of the analysis.

When conducting a normality test on the dataset, a combination of visual inspection and statistical tests is often recommended for a more comprehensive assessment of the distribution characteristics. If the data deviates significantly from normality, appropriate transformations or non-parametric tests may be considered for subsequent analyses.

4.3 Conversion

Label encoding is a technique used to convert categorical data into numerical format, particularly when the categories have an ordinal relationship. It assigns a unique integer to each category, essentially encoding them with numerical values.

In this dataset, there are 2 categorical columns, Gender and Personality.
For Gender after encoding,

- Female == 0
- Male ==1

For the Personality after encoding,

- serious 4
- extraverted 3
- responsible 2
- lively 1
- dependable 0

4.4 Splitting of Data

In machine learning, it's common to split the dataset into training and testing sets to evaluate the model's performance on unseen data. The scikit-learn library provides a convenient function for this purpose.

Here, features represent your input data (including both numerical and label-encoded columns), and labels represent the target variable you want to predict. The train_test_split function randomly shuffles and splits the data into training and testing sets. The test_size parameter determines the proportion of the data allocated for testing (here set to 20%), and random_state ensures reproducibility.

After splitting the data, you can use X_train and y_train for training your machine learning model and X_test and y_test for evaluating its performance on unseen data. Adjust the test_size parameter based on your specific requirements.

4.5 Training and Testing of Models

4.5.1 OLS

The Ordinary Least Squares (OLS) model serves as a foundational technique for regression analysis. OLS is applied to test the dataset's predictive capabilities, aiming to establish a relationship between the input features and the target personality traits. The OLS model assumes that the relationship between variables is linear and seeks to minimize the sum of squared differences between the predicted and observed values. In the context of personality prediction, this involves fitting the model to the training dataset, where features extracted from diverse sources serve as predictors for the Big Five personality traits. The model's performance is then evaluated on a separate testing dataset, assessing its ability to generalize to new, unseen data.

4.5.2 Linear Regression

The Linear Regression model assumes a linear connection between the input features and the output variable, making it particularly relevant for predicting continuous variables, such as scores on personality traits. During the testing phase, the model is trained on a subset of the dataset, typically referred to as the training set, and then evaluated on a separate portion known as the testing set. The testing dataset contains unseen instances,

allowing for an assessment of the model's generalization performance. The model's effectiveness is gauged by metrics such as mean squared error or R-squared, which quantify the accuracy of its predictions.

4.5.3 PCR

Principal Component Regression (PCR) emerges as a powerful modeling technique, particularly effective in handling datasets with high dimensionality. PCR combines the dimensionality reduction capabilities of Principal Component Analysis (PCA) with the predictive prowess of regression. During the testing phase, the PCR model is employed on the personality prediction dataset, which often comprises a myriad of behavioral and psychological features. PCA is initially applied to identify the principal components that capture the most significant variability in the data. Subsequently, these components are used as predictors in a regression framework to model the relationship with the target personality traits. The testing dataset, distinct from the training set, allows for an assessment of the model's performance on unseen instances.

4.5.4 PLS

The Partial Least Squares (PLS) model takes center stage as an advanced technique capable of handling multivariate relationships. PLS, a latent variable approach, seeks to maximize the covariance between the predictor variables (features) and the response variables (personality traits), making it particularly well-suited for datasets with high dimensionality and potential multicollinearity. During the testing phase, the PLS model is trained on a subset of the dataset and evaluated on a separate, previously unseen testing set. This evaluation involves assessing how well the model captures the underlying patterns and relationships within the data to predict personality traits accurately. Given its ability to handle complex relationships and noisy data, the results of testing the PLS model on our personality prediction dataset contribute valuable insights into the nuanced interplay between various behavioral and psychological features and the Big Five personality traits.

4.5.5 Ridge Regression

The Ridge Regression model emerges as a powerful tool for addressing potential multicollinearity issues and refining the predictive accuracy of our models. Ridge Regression is an extension of Linear Regression that introduces regularization to the cost function, preventing overfitting and stabilizing the model. During the testing phase, the Ridge model is trained on a subset of the dataset, incorporating diverse features associated with behavioral and psychological traits. This model is then evaluated on a distinct testing set, allowing us to assess its performance on unseen data. The introduction of a regularization term in Ridge Regression not only enables a robust fit to the training data but also encourages simplicity in the model by penalizing large coefficients. The testing outcomes of the Ridge model contribute valuable insights into its ability to generalize and predict personality traits accurately.

4.5.6 Lasso Regression

The Lasso Regression model plays a pivotal role in testing and refining predictive capabilities. Lasso Regression is a regularization technique that not only aims to predict personality traits based on input features but also incorporates a penalty term to shrink certain feature coefficients towards zero. This promotes feature selection, allowing the model to identify and emphasize the most influential predictors. During the testing phase, the Lasso model is trained on a subset of the dataset, typically the training set, and subsequently evaluated on a distinct testing set to assess its generalization performance.

4.5.7 ENeT Model

The Elastic Net (Enet) model emerges as a versatile tool for evaluating the dataset's complexity and relationships among features. The Elastic Net model combines the L1 and L2 regularization techniques, effectively incorporating both feature selection and regularization to prevent overfitting. During the testing phase, the Enet model is trained on a subset of the personality prediction dataset, leveraging diverse behavioral and psychological features as predictors for personality traits. The testing dataset, comprising previously unseen instances, provides a rigorous evaluation of the model's performance and its ability to generalize to new data.

4.5.8 KNN

The K-Nearest Neighbors (KNN) model introduces a distinctive approach. Unlike traditional linear models, KNN relies on the proximity of instances in feature space. During the testing phase, the KNN model is trained on a subset of the dataset, and predictions are made for unseen instances based on the personality traits of their nearest neighbors in the training data. The evaluation of the model's performance involves assessing its ability to correctly predict personality traits for instances not encountered during training. Key metrics, such as accuracy and precision, shed light on the effectiveness of the KNN model in capturing patterns within the dataset. The outcomes of testing the KNN model on the personality prediction dataset contribute nuanced insights into the impact of local relationships in feature space on personality trait predictions.

4.5.9 XgBoost

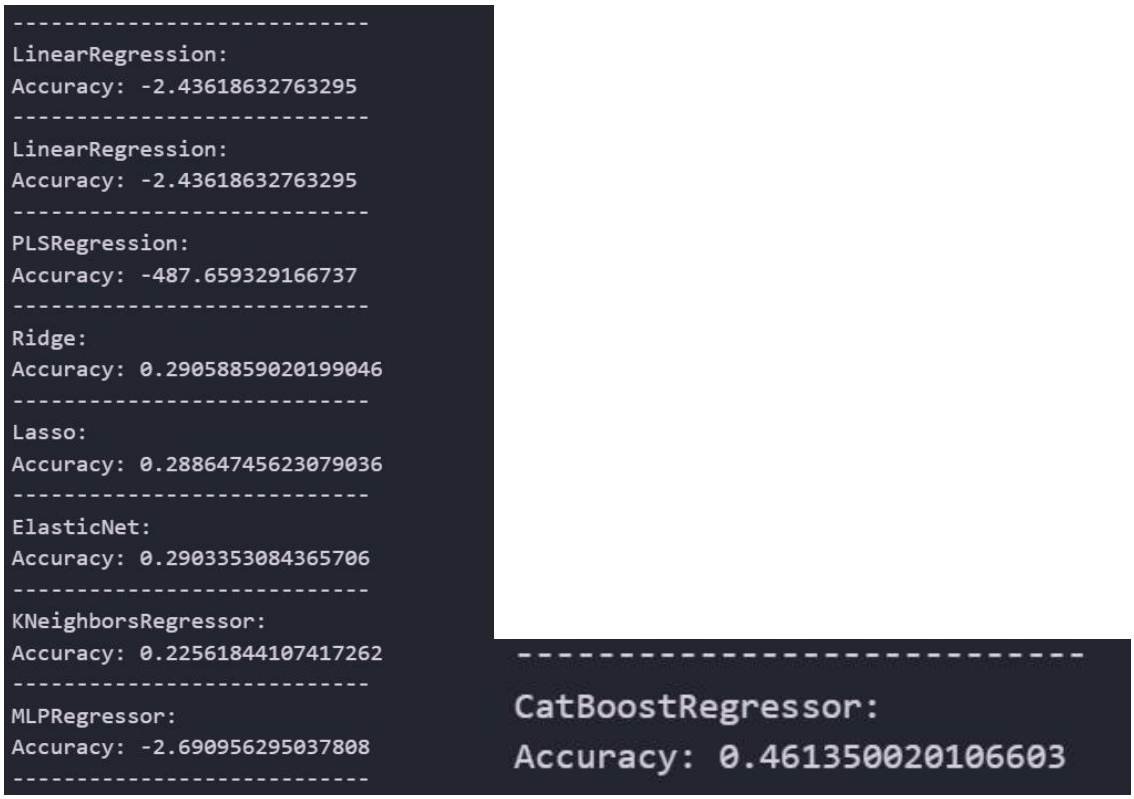
the XGBoost model emerges as a powerful tool for its ability to handle complex relationships and capture intricate patterns within the data. XGBoost, an implementation of gradient-boosted decision trees, excels in predictive tasks, providing high accuracy and robustness. During the testing phase of the personality prediction project, the XGBoost model is trained on a carefully curated dataset, leveraging features derived from social media, online activities, and self-reported data. The model's performance is then assessed on an independent testing set, enabling an evaluation of its predictive capacity on unseen instances. Key evaluation metrics, such as accuracy, precision, recall, and F1-score, are employed to quantify the model's effectiveness in predicting personality traits, specifically the Big Five. The XGBoost model's testing outcomes contribute crucial insights into its ability to discern nuanced relationships within the dataset, offering valuable information for refining the model's hyperparameters and optimizing its predictive capabilities.

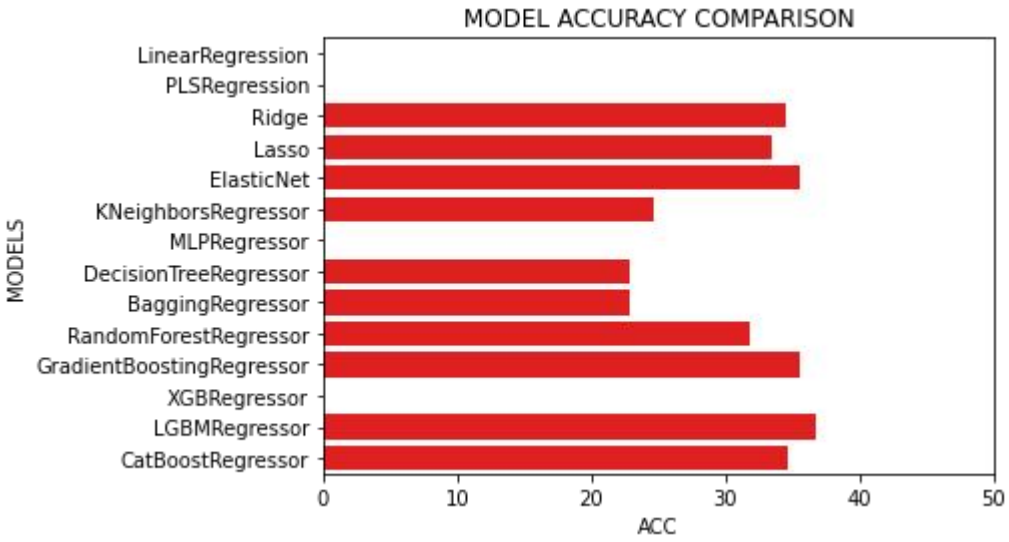
4.5.10 CatBoost

The CatBoost model emerges as a powerful and adaptive tool for machine learning. CatBoost, a gradient boosting algorithm, excels in handling categorical features without the need for extensive preprocessing. During the testing phase, the CatBoost model is trained on a subset of the dataset,

leveraging its capacity to handle both numerical and label-encoded categorical variables. The model's predictive performance is then rigorously evaluated on a distinct testing set, allowing us to gauge its ability to generalize to unseen data. The CatBoost model offers advantages such as robustness to overfitting, efficient handling of missing data, and automatic feature scaling, making it particularly well-suited for datasets with diverse and complex features, as often found in personality prediction studies. The testing results from the CatBoost model provide valuable insights into its predictive accuracy and generalization capabilities, contributing to the assessment of its suitability for real-world applications in understanding and forecasting individual personality traits based on a rich array of behavioral and psychological features.

Accuracy and comparison of regressor models





4.5.11 Logistic Regression

The Logistic Regression model plays a pivotal role in discerning the probability of categorical outcomes, such as the likelihood of specific personality traits manifesting in individuals. The testing phase involves training the Logistic Regression model on a designated subset of the dataset and subsequently evaluating its performance on an independent testing set. The dataset, enriched with diverse behavioral and psychological features, serves as the foundation for predicting binary outcomes related to personality characteristics. During testing, the model's ability to discriminate between different personality categories is assessed using metrics like accuracy, precision, recall, and the receiver operating characteristic (ROC) curve. By scrutinizing the model's predictions against the true personality labels in the testing set, insights are gained into its classification performance.

4.5.12 Gaussian Naive Bayes

The application of the Gaussian Naive Bayes (GNB) model offers a probabilistic approach to classification tasks. Unlike linear regression, GNB is particularly suited for categorical prediction, making it a valuable tool for assessing personality traits that can be discretized into distinct classes. The testing of the GNB model involves training it on a portion of the dataset and evaluating its performance on a separate, previously unseen testing set. Leveraging the assumption of conditional independence among features given the class, GNB calculates probabilities based on the Gaussian

distribution. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive understanding of its predictive capabilities.

4.5.13 KNN Classification

The k-Nearest Neighbors (KNN) classification model offers a distinctive approach by relying on the similarity of instances. During the testing phase, the KNN model is employed to assess the dataset's predictive capabilities for personality traits. The model is trained on a subset of the dataset, utilizing the features extracted from diverse sources to categorize individuals based on their personality profiles. The KNN algorithm classifies a test instance by considering the personalities of its k-nearest neighbors in the feature space. The model's performance is then evaluated on a separate testing dataset, allowing for an examination of its ability to accurately classify individuals into specific personality categories.

4.5.14 ANN Classification

An Artificial Neural Network (ANN) classification model emerges as a potent tool for capturing intricate patterns and non-linear relationships. The ANN, inspired by the human brain's neural structure, is particularly adept at handling complex datasets with diverse features. During the testing phase, the ANN model is trained on a portion of the dataset, utilizing a backpropagation algorithm to adjust weights and biases iteratively. Subsequently, the model's efficacy is evaluated on a distinct testing subset, gauging its ability to generalize to unseen instances. Classification metrics such as accuracy, precision, recall, and F1-score are employed to assess the model's performance in predicting personality traits. The ANN's strength lies in its capacity to learn hierarchical representations, allowing it to discern subtle nuances within the data.

4.5.15 Regression Tree Classification

The application of a Regression Tree Classification model represents a dynamic and interpretable approach to personality prediction within the dataset. Unlike traditional linear models, Regression Trees capture complex, non-linear relationships between input features and personality traits. During

model testing, the dataset is divided recursively into subsets based on feature conditions, forming a tree structure where each leaf node corresponds to a predicted personality outcome. The testing phase involves assessing the model's predictive accuracy on unseen data, evaluating metrics such as Mean Squared Error or R-squared for regression tasks. The interpretability of the tree structure provides valuable insights into the hierarchical importance of different features in predicting personality traits.

4.5.16 Random Forest Classification

The application of a Random Forest classification model introduces a robust and versatile approach to analyze the complex interplay of features influencing individual traits. The Random Forest algorithm excels in capturing non-linear relationships and interactions within the dataset, making it particularly suitable for discerning the nuanced nature of personality traits. During the testing phase, the Random Forest model is trained on a designated portion of the dataset, leveraging an ensemble of decision trees to collectively make predictions. The model's performance is then assessed on a separate testing set, enabling the evaluation of its predictive accuracy and generalization capabilities. Metrics such as accuracy, precision, recall, and F1-score provide a comprehensive understanding of the model's effectiveness in classifying personality traits.

4.5.17 XgBoost Classifier

The XGBoost classification model emerges as a powerful tool for discerning patterns within the dataset. Unlike linear regression, XGBoost specializes in classification tasks, making it well-suited for predicting categorical variables, such as personality traits. During the testing phase of the project, the XGBoost model is trained on a designated portion of the dataset, known as the training set, and subsequently evaluated on a separate set, the testing set. The model's ability to categorize personality traits, often based on the widely studied Big Five factors, is assessed using performance metrics such as accuracy, precision, recall, and F1-score. XGBoost's ensemble learning approach, combining the strengths of multiple decision trees, enables it to capture intricate relationships and nonlinearities in the data.

4.5.18 LightGBM Classifier

The LightGBM classification model stands out as a powerful tool for capturing complex patterns and relationships within the dataset. Employing a gradient boosting framework, LightGBM excels in handling large datasets with high dimensionality, making it well-suited for the intricate nature of personality prediction tasks. During the testing phase, the LightGBM model is trained on a meticulously prepared training dataset, leveraging its ability to build decision trees and ensemble them for superior predictive performance. The model is subsequently evaluated on a distinct testing dataset to assess its generalization capabilities and classification accuracy in predicting personality traits. Metrics such as precision, recall, and F1-score provide a comprehensive understanding of the model's performance.

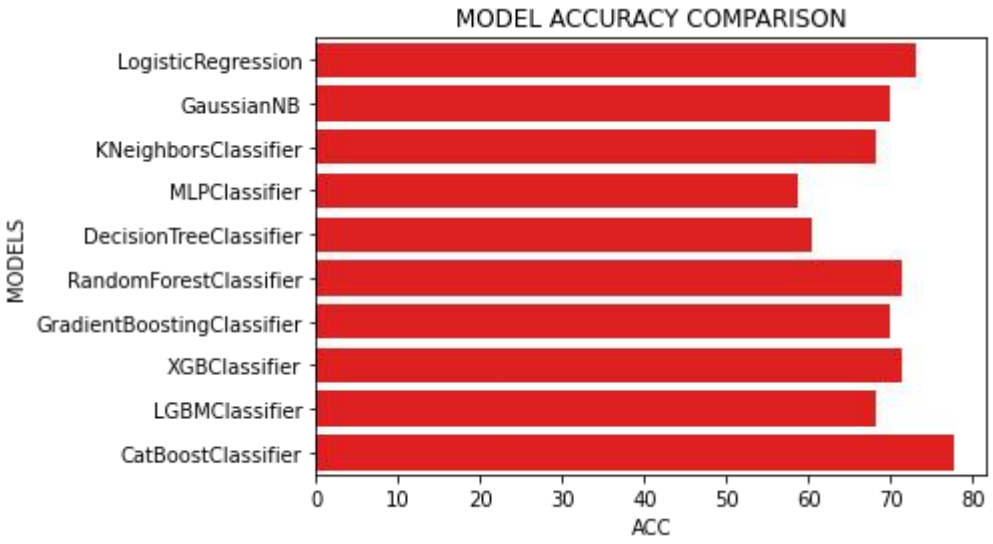
4.5.19 CatBoost Classifier

The CatBoost Classification model emerges as a powerful tool for handling categorical features and intricate relationships within the dataset. CatBoost, renowned for its capability to efficiently manage categorical data, is applied to assess and predict personality traits in this study. During the testing phase, the model undergoes training on a designated portion of the dataset, known as the training set, leveraging its gradient boosting algorithm to capture complex patterns and interactions among diverse features. Following this training, the model is rigorously evaluated on an independent testing set, comprised of previously unseen instances. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to gauge the model's performance in classifying individuals into distinct personality categories. The utilization of CatBoost in this context is particularly advantageous due to its inherent ability to handle categorical variables without the need for extensive pre-processing.

Accuracy and Comparison of classifier model

```
LogisticRegression:
Accuracy: 0.7301587301587301
---------------------------
GaussianNB:
Accuracy: 0.6984126984126984
---------------------------
KNeighborsClassifier:
Accuracy: 0.6825396825396826
---------------------------
MLPClassifier:
Accuracy: 0.5873015873015873
---------------------------
LGBMClassifier:
Accuracy: 0.6825396825396826
---------------------------
CatBoostClassifier:
Accuracy: 0.7777777777777778
```

```
DecisionTreeClassifier:
Accuracy: 0.6031746031746031
---------------------------
RandomForestClassifier:
Accuracy: 0.7142857142857143
---------------------------
GradientBoostingClassifier:
Accuracy: 0.6984126984126984
---------------------------
XGBClassifier:
Accuracy: 0.7142857142857143
```



Chapter 5

Discussion and Conclusion

Evaluation of machine learning algorithms for project is critical. To forecast the personality system, we employed machine learning algorithms such as Linear Regression, KNN, MLP, Random Forest, Naïve Bais, Logistic Regression Decision Tree and many other methods. For predicting behaviour from test data, we have 7 qualities and one attribute labelled as personality. Gender, age, openness, agreeableness, extraversion, neuroticism or emotional stability, and conscientiousness are some of the attributes employed in this method. In our system, we used Five Personality Traits and produced performance measures for each trait. Accuracy were calculated for each attribute.

And from the calculated accuracies it is concluded that for the given dataset Categorical Boosting Classifier is best for predicting personalities. So it can be used for further future use to predict personalities.

References

1. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In NeurIPS.

2. McCrae, R. R., & Costa, P. T. Jr. (1987). Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 52(1), 81–90.

3. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis (7th ed.). Prentice Hall. (For a comprehensive overview of normality testing methods in the context of multivariate data analysis.)

4. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785 (The original XGBoost paper introducing the algorithm.)

5. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

6. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

7. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Li, Q. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (NeurIPS), 30.

Repository Link for the project:

https://github.com/ialokranj/Major-Project