**AUTISM PREDICTION**

**Project Report**

Submitted to the Faculty of Engineering of

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA**

In partial fulfillment of the requirements for the award of the Degree of

## BACHELOR OF TECHNOLOGY

## In

## CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Submitted by

**NAGANABOYINA KALYANI (22481A4274)**

**RAVULAPALLI KOMALI(22481A4293)**

**NADAKUDHITI GOWTHAM MANI VARMA(22481A4273)**

**SHAIK AMAR(22481A4299)**

Under the Enviable and Esteemed Guidance of

# Dr. Y. ADILASKHMI, M. Tech, Ph.D

Professor & Head of the Department,

CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)



**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**

## SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

**(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada) SESHADRIRAO KNOWLEDGE VILLAGE GUDLAVALLERU – 521356**

**ANDHRA PRADESH**

**2024-25**

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

**(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)**

**SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU**

**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**



**CERTIFICATE**

This is to certify that the project report entitled **“Autism Prediction”**is a bonafide record of work carried out,Naganaboyina Kalyani (22481A4274), Ravulapalli Komali(22481A4293), Nadakudhiti Gowtham Mani Varma(22481A4273), Shaik Amar(22481A4299) under the guidance and supervision of **Dr. Y. ADILASKHMI, Professor and Head of the Department,** CSE (Artificial Intelligence & Machine Learning), in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

**Project Guide Head of the Department**

**(Dr. Y. ADILAKSHMI) (Dr. Y. ADILAKSHMI)**

**ACKNOWLEDGEMENT**

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr. Y. ADILAKSHMI, Professor & Head of the Department,** CSE (Artificial Intelligence & Machine Learning) for her constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. Y. ADILAKSHMI**, **Head of the Department**, CSE (Artificial Intelligence & Machine Learning) for her encouragements all the way during analysis of the project. Her annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B. KARUNA KUMAR** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project intime.

Team Members

**NAGANABOYINA KALYANI (22481A4274)**

**RAVULAPALLI KOMALI (22481A4293)**

**NADAKUDHITI GOWTHAM MANI VARMA (22481A4273)**

**SHAIK AMAR (22481A4299)**

## 

## INDEX

**TITLE PAGENO**

LIST OF TABLES i

LIST OF FIGURES ii

[ABSTRACT iii](#_bookmark0)

CHAPTER 3: INTRODUCTION 1

* 1. [Introduction](#_TOC_250000)
  2. Problem definition 3

[CHAPTER 2: PROPOSED METHOD 4](#_bookmark1)

* 1. Methodology
     1. Block Diagram
     2. Algorithm and Explanation
  2. Data Preparation 6
     1. Data Description
     2. Data Pre-processing

CHAPTER 3: RESULTS 11

3.1 Machine learning Model

3.2 Result of the model

3.3 Visualization

CHAPTER 4: CONCLUSION AND FUTURE SCOPE 18

References

List of Program Outcomes and Program Specific Outcomes Mapping of Program Outcomes with graduated POs and PSOs

## ABSTRACT

The increasing prevalence of Autism Spectrum Disorder (ASD) highlights the need for efficient and early prediction techniques to facilitate timely interventions. This project aims to develop a machine learning-based approach to predict the likelihood of autism in individuals using questionnaire-based datasets. By leveraging machine learning algorithms, we can identify patterns and relationships in the data that may not be apparent through traditional methods.

The project's methodology involves data preprocessing, feature encoding, and class imbalance handling via SMOTE (Synthetic Minority Over-sampling Technique). We train and evaluate machine learning models using **Decision Tree**, **Random Forest**, and **XGBoost classifiers**. The study's findings will provide insights into the effectiveness of these models in predicting autism likelihood, as well as identify relevant features contributing to the autism prediction in this project here. The ultimate goal is to enable early intervention and support services for individuals with autism, improving their quality of life and outcomes.

## CHAPTER 1

## INTRODUCTION

* 1. **Introduction**

Autism Spectrum Disorder (ASD) is a complex developmental condition marked by challenges in social interaction, communication, and repetitive behaviors, impacting individuals to varying degrees. Machine learning (ML) models offer a promising solution by automating and enhancing the prediction of autism tendencies from questionnaire data. ML models can facilitate rapid assessments in both clinical and remote environments, potentially streamlining the diagnostic process and enabling earlier access to targeted therapies and support. This integration of technology and healthcare has the potential to positively impact individuals with ASD and their families.

The **Random Forest Classifier** is a powerful ensemble learning method that constructs multiple decision trees during training and merges their predictions to enhance accuracy and reduce overfitting. It excels in handling high-dimensional datasets, such as those containing behavioral traits, genetic markers, and developmental records, while robustly managing noise and missing values common in medical data. Key advantages include its ability to identify critical features through built-in feature importance analysis, its resilience to imbalanced data via class weighting or sampling techniques, and its capacity to model non-linear relationships without manual feature engineering.

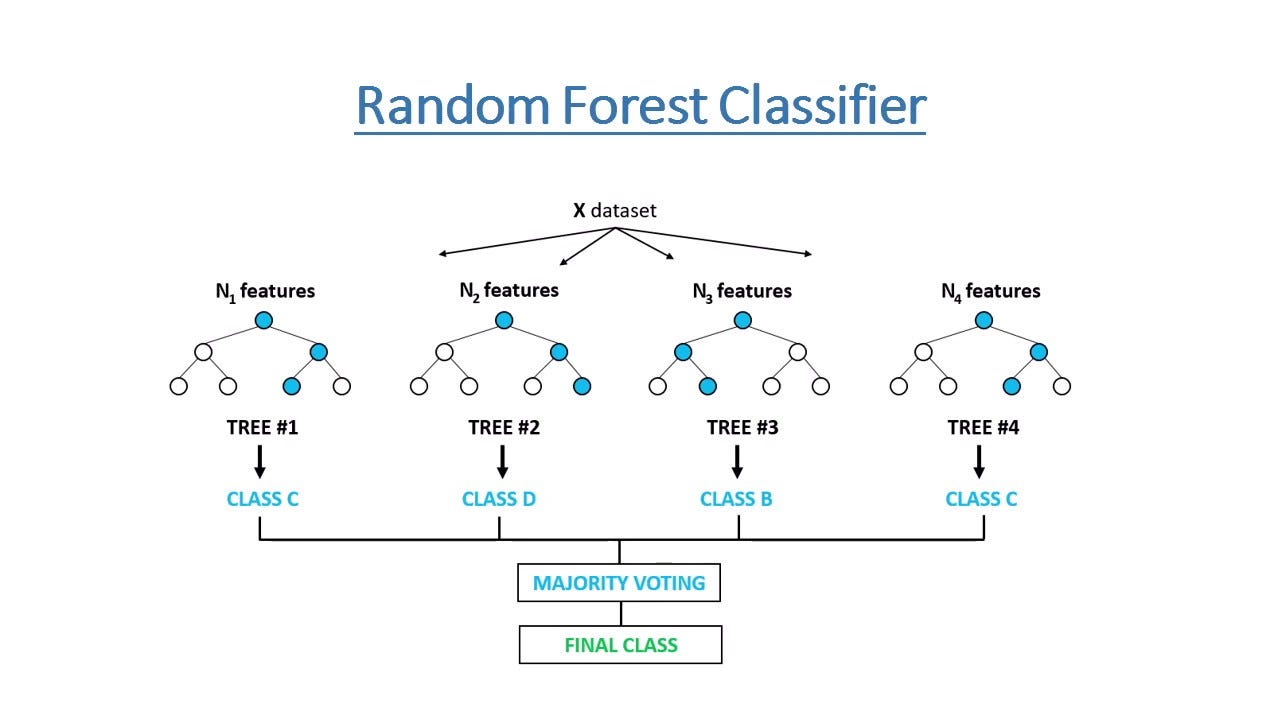


Fig:1.1.1 Random forest classifier

**XGBoost** is an advanced gradient-boosting algorithm that excels in autism prediction by sequentially building decision trees to correct errors from previous ones, enhancing accuracy. It handles missing values natively, prevents overfitting through regularization, and processes data efficiently with parallel computing With hyperparameter tuning (e.g., learning rate, tree depth), it outperforms traditional models like Random Forest. XGBoost's speed, precision, and interpretability make it a top choice for early ASD detection systems.

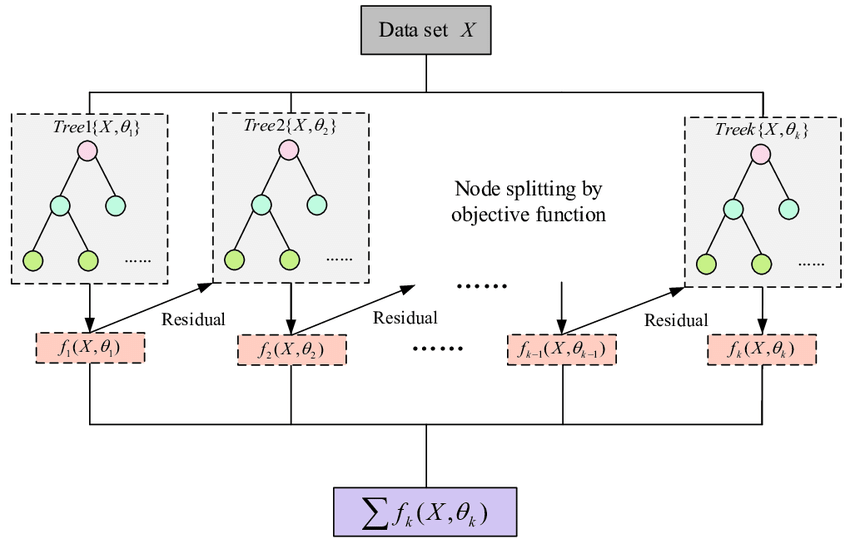


Fig 1.1.2 XGBoost Algorithm

A **Decision Tree Classifier** is a simple yet powerful algorithm that makes predictions by learning hierarchical decision rules from features like behavioral traits, genetic markers, or developmental history. It splits data into branches based on feature thresholds (e.g., "Does the child maintain eye contact?") until reaching a classification (ASD or non-ASD). While intuitive and interpretable, it risks overfitting without proper depth control (max\_depth)

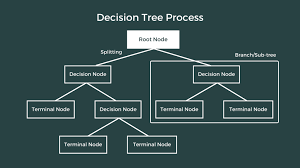


Fig 1.1.3 Decision tree classifier

## 1.2. Problem statement

Traditional Autism Spectrum Disorder (ASD) diagnosis involves in-person clinical evaluations, which can be time-consuming and inaccessible in remote or underserved areas. To address this challenge, this work aims to develop a machine learning model that can accurately predict autism based on questionnaire data. This approach offers a scalable and faster alternative for preliminary screening, enabling early identification and intervention. By leveraging machine learning, this solution can increase accessibility to ASD diagnosis, particularly in areas where traditional diagnostic methods are limited. The goal is to create a reliable and efficient model that can support timely interventions and improve outcomes for individuals with autism.

**CHAPTER 2**

**PROPOSED METHOD**

**2.1. Methodology**

The methodology involves collecting and preprocessing questionnaire-based datasets related to Autism Spectrum Disorder (ASD), followed by feature encoding and handling class imbalance using techniques like SMOTE. Machine learning models, including Decision Tree, Random Forest, and XGBoost classifiers, are trained and evaluated on the dataset to predict autism tendencies. Model performance is assessed using metrics such as accuracy, precision, recall, and F1-score. The goal is to identify the most effective model for predicting autism likelihood and provide insights into the most relevant features contributing to the prediction, ultimately supporting early detection and intervention for individuals with ASD.

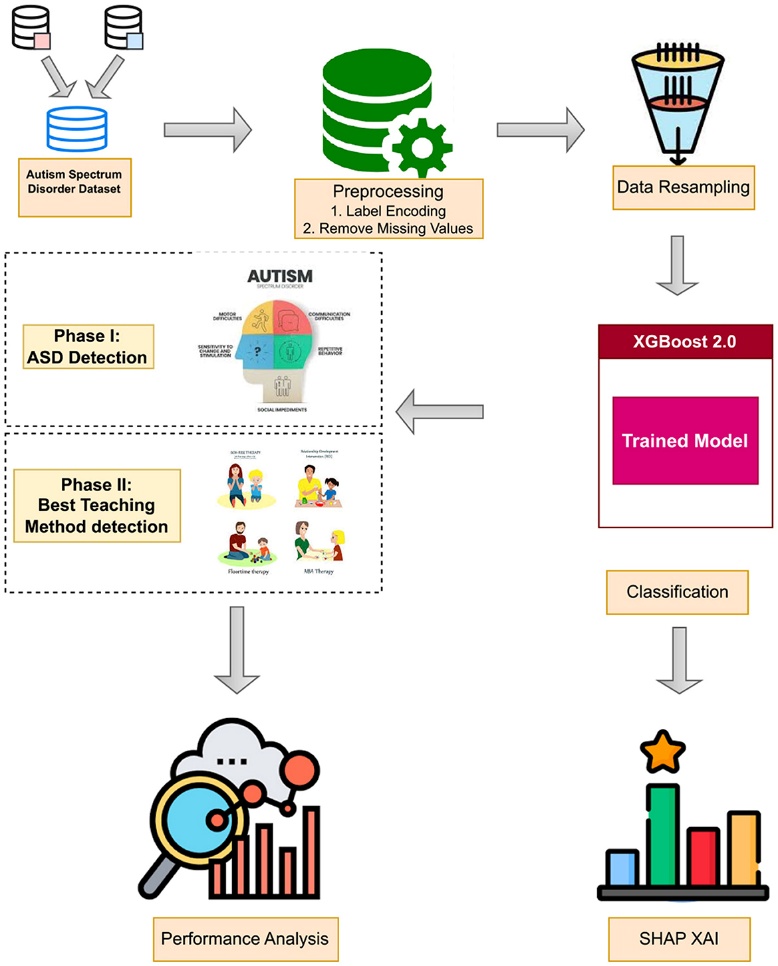


Fig 2.1.1 Methodology

* + 1. **Block Diagram**

The Random Forest Classifier block involves training an ensemble learning model that combines multiple decision trees to predict autism tendencies. The Random Forest Classifier block takes the preprocessed data as input and generates a predicted likelihood of autism tendencies. By combining the strengths of multiple decision trees, this block provides a robust and accurate prediction model.

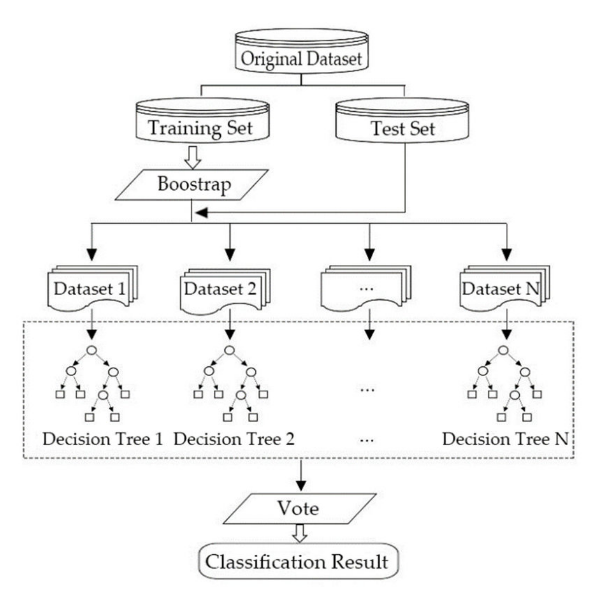


Fig. 2.1.2. Random Forest

* + 1. **Algorithm and Explanation:**

**Random Forest Classifier:**

Random Forest is an ensemble machine learning technique that builds multiple decision trees and merges their results to improve the accuracy and robustness of predictions.

* **Input**:
* Training dataset D={(x1,y1),(x2,y2),...,(xn,yn)}
* Number of trees: T
* Number of features to consider when splitting: mmm
* **For each tree t=1 to TTT**:
* Draw a bootstrap sample Dt from the training data D
* Train a decision tree ht​ on Dt by:
  + At each node:
    - Randomly select mmm features from the full set of features
    - Pick the best split among the mmm features
  + Fully grow the tree (no pruning)
* **Prediction (for new input xxx)**:
* Each of the TTT trees makes a prediction : h1(x),h2(x),...,hT(x)
* The final prediction is the **majority vote**:

**y^​=mode(h1​(x),h2​(x),...,hT​(x))**

* **Important Parameters:**
* n\_estimators: Number of trees in the forest.
* max\_features: Number of features to consider when looking for the best split.
* max\_depth: Maximum depth of the tree.
* min\_samples\_split: Minimum number of samples required to split a node.
* bootstrap: Whether bootstrap samples are used.

**2.2. Data Preparation**

**2.2.1 Data Description**

The dataset used for this project is related to **Autism Spectrum Disorder (ASD) screening** and consists of behavioral and demographic information collected through an autism screening questionnaire. It is designed to help predict whether an individual is likely to have ASD based on their responses.

This dataset is publicly available and has been widely used for research in the domain of medical AI and mental health analytics Source of dataset:

* The dataset is typically sourced from platforms like **Kaggle** or the **UCI Machine Learning Repository**.
* It includes both **categorical** and **numerical** features.

| **Feature Name** | **Description** |
| --- | --- |
| A1\_Score-to- A10\_Score | Answers to 10 screening questions (binary: yes/no) |
| age | Age of the individual (numerical) |
| gender | Gender of the individual |
| ethnicity | Ethnic background |
| jaundice | Whether the individual had jaundice at birth |
| austim | Whether any immediate family member has autism |
| contry\_of\_res | Country of residence |
| used\_app\_before | Whether the individual used the screening app before |
| result | Score from the screening tool (numerical) |
| age\_desc | Description of age group (e.g., "18 and more") |
| relation | Relation of the respondent to the child (e.g., Parent, Health Professional) |
| Class/ASD | Target variable (Yes/No – indicates ASD positive or negative) |
| ID | Unique identifier for each record (not useful for training) |

**2.2.2 Data Preprocessing**

Preprocessing is a critical step in building any machine learning model. The quality of the data directly impacts the performance and reliability of the final model. In this project, we performed several preprocessing operations to clean, standardize, and prepare the Autism dataset for model training.

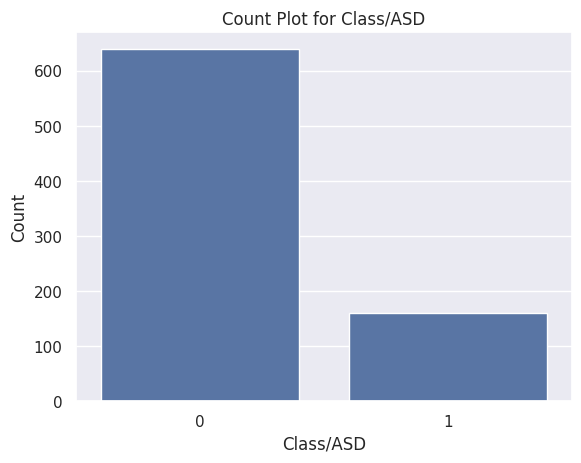


Fig. 2.1.3. count plot of classes

**1.Data Type Conversion**

**Purpose**: Ensure that numerical columns are correctly treated as numbers and not as strings. This is important for plotting, statistical analysis, and algorithm training.



Fig.2.1.4.Datatype conversion code

**2.Country Name Standardization**

**Purpose**: Correct inconsistencies in categorical values. For example, “Viet Nam” and “Vietnam” should be treated as the same country.

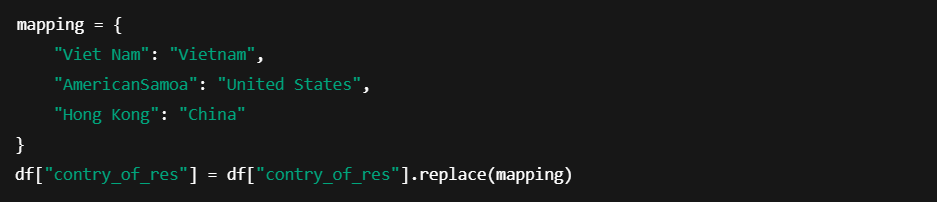


Fig.2.1.5.Standardization code

**3.Dropping Irrelevant Columns**

**Purpose**: Remove features that do not contribute to prediction, such as identifiers or redundant data.



Fig.2.1.6.Dropping columns code

**4.Handling Missing and Inconsistent Values**

**Purpose**: Address incomplete or ambiguous entries in categorical columns to improve model robustness.

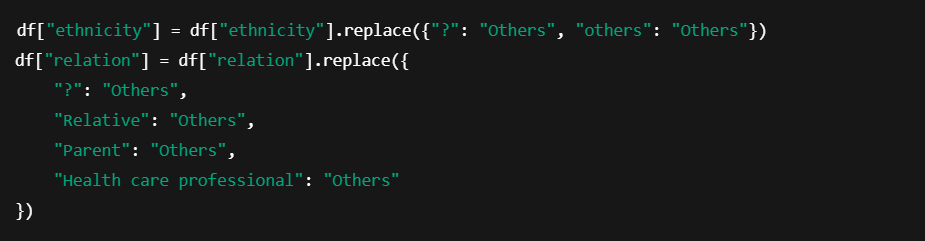


Fig.2.1.7.Handling missing values code

**5.Label Encoding for Categorical Features**

**Purpose**: Convert non-numeric categorical variables into integers so that machine learning algorithms can process them.

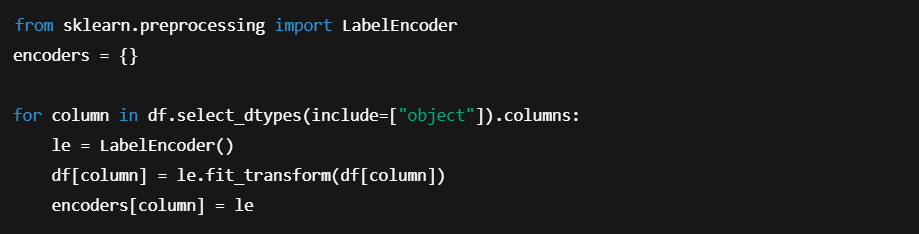
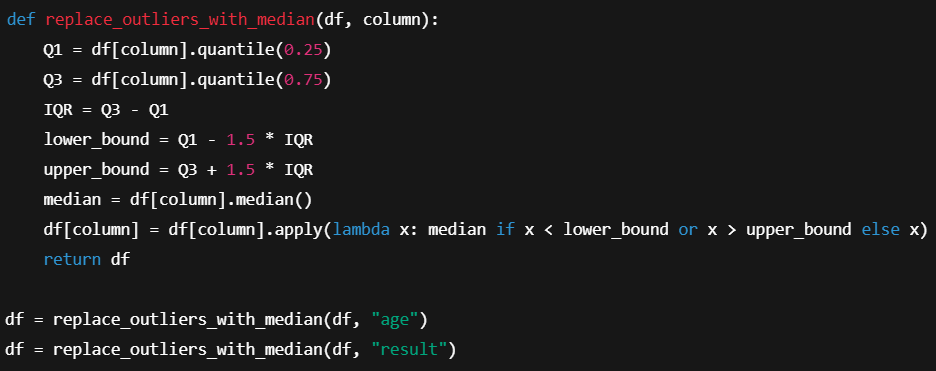


Fig.2.1.8.Encoding For Categorical Features code

**6.Outlier Detection and Treatment**

**Purpose**: Remove or replace extreme values that may skew results or reduce model performance. We used the Interquartile Range (IQR) method to detect outliers and replaced them with the median value. This avoids discarding data while still reducing noise.



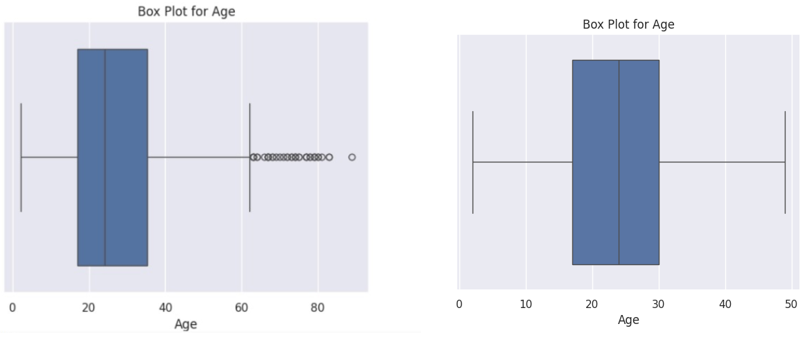


Fig. 2.1.9. before and after removal of outliers

**7.Handling Class Imbalance with SMOTE**

**Purpose**: Improve model performance on the minority class by generating synthetic data points, avoiding bias toward the majority class.

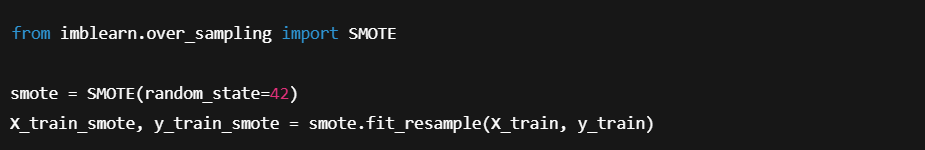
**Top of FormBottom of Form**

Fig.2.1.10.handling class imbalance code

**CHAPTER 3**

**RESULTS**

**3.1 Machine Learning Model :**

* **Data Splitting :**

Separate the dataset into two parts—training and testing. This allows us to evaluate model performance on unseen data. An 80/20 split was used to ensure enough data for both training and evaluation. A fixed random seed ensures reproducibility.

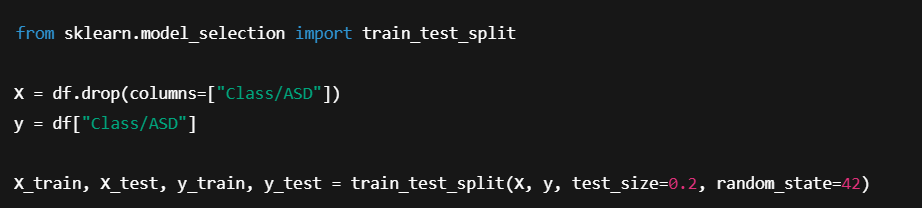


Fig.3.1.1.data splitting code

* **Model Definition and Compilation :**
  + **Model Definition:** In this project, we aimed to predict whether an individual is likely to be diagnosed with Autism Spectrum Disorder (ASD) based on a set of behavioral and demographic features. To accomplish this, we defined and trained three powerful tree-based classification models: Decision Tree, Random Forest, and XGBoost.

**models = {**

**"Decision Tree": decision\_tree,**

**"Random Forest": random\_forest,**

**"XGBoost": xgboost\_classifier**

**}**

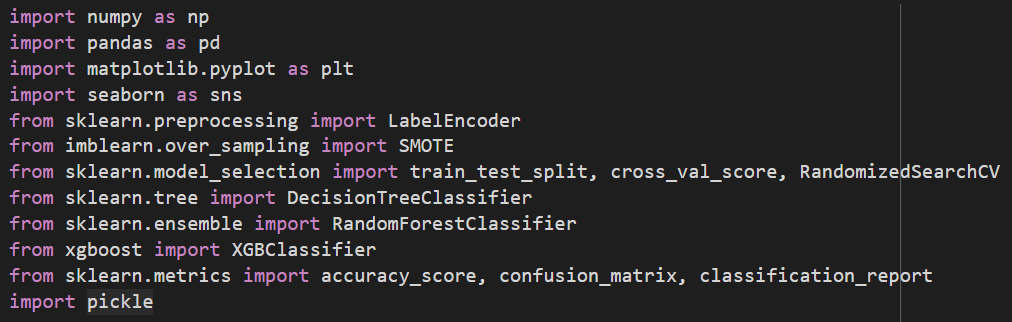
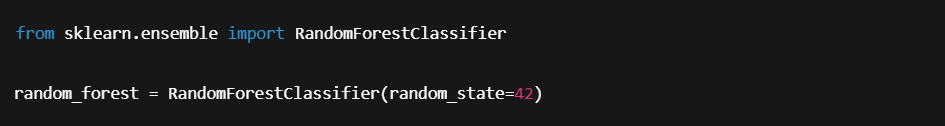
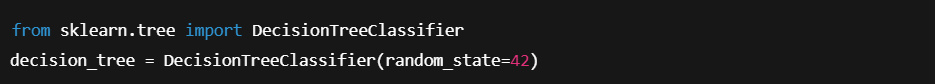


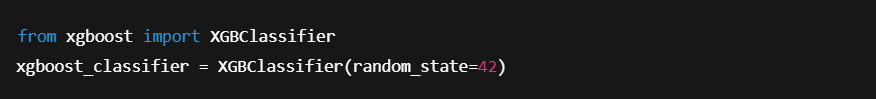
Fig.3.1.2.importing dependencies code

* + **Model Initialization:** The Random Forest Classifier was initialized using the RandomForestClassifier class from the sklearn.ensemble module. During initialization, a random\_state parameter was set to ensure reproducibility of results across different runs. By default, the model creates 100 decision trees, each trained on a different subset of the data using bootstrap sampling. At this stage, no additional hyperparameters were specified, allowing us to evaluate the baseline performance of the model before applying further tuning through techniques such as RandomizedSearchCV.

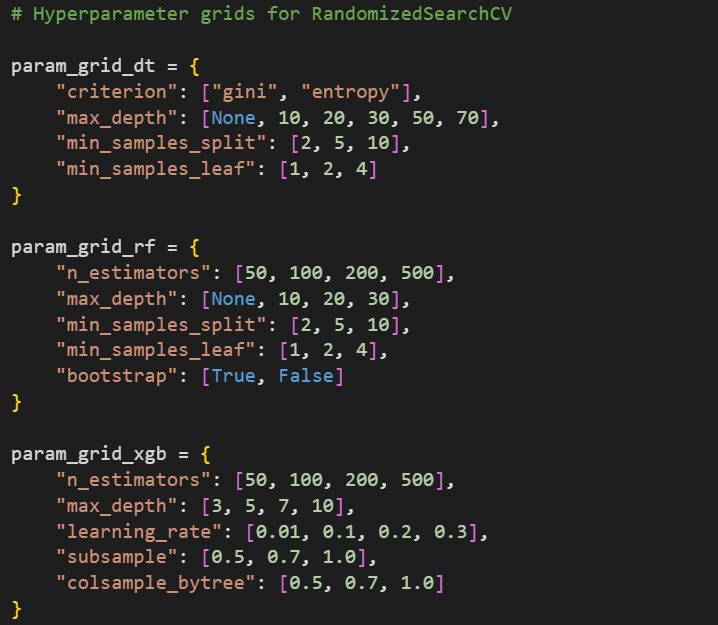


Other model:

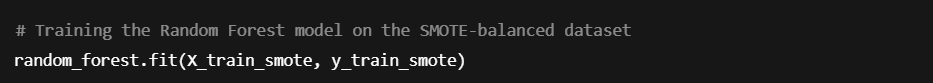
****

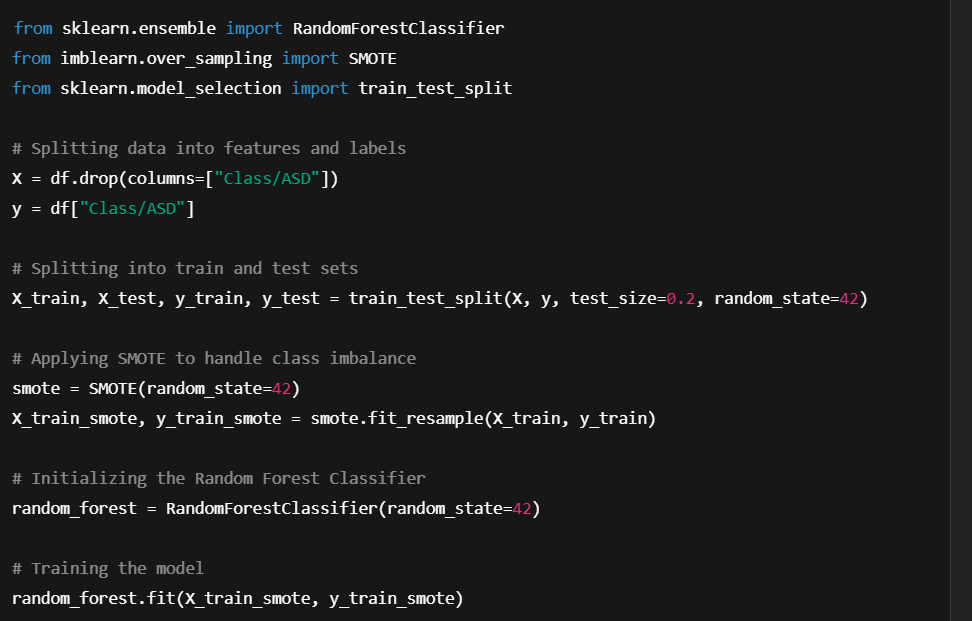
****

**Hyperparameter tuning:** It was performed using RandomizedSearchCV, a powerful technique that searches through a defined set of hyperparameter combinations. For each model—Decision Tree, Random Forest, and XGBoost—customized hyperparameter grids were created based on their core configurable settings.



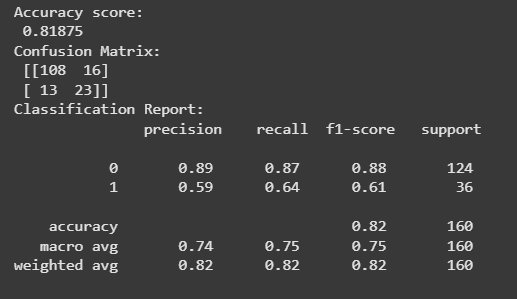
* **Model Training :** Once the Random Forest Classifier was initialized, the model training process was carried out using the .fit() method from the scikit-learn library. The model was trained on the resampled dataset obtained through SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in the target variable (Class/ASD). During training, the Random Forest algorithm constructed an ensemble of decision trees, each trained on different bootstrapped subsets of the training data. At each node in these trees, a random subset of features was considered for splitting, which promotes diversity among the trees and reduces overfitting**.**

****

****

**3.2 Results of the Model**

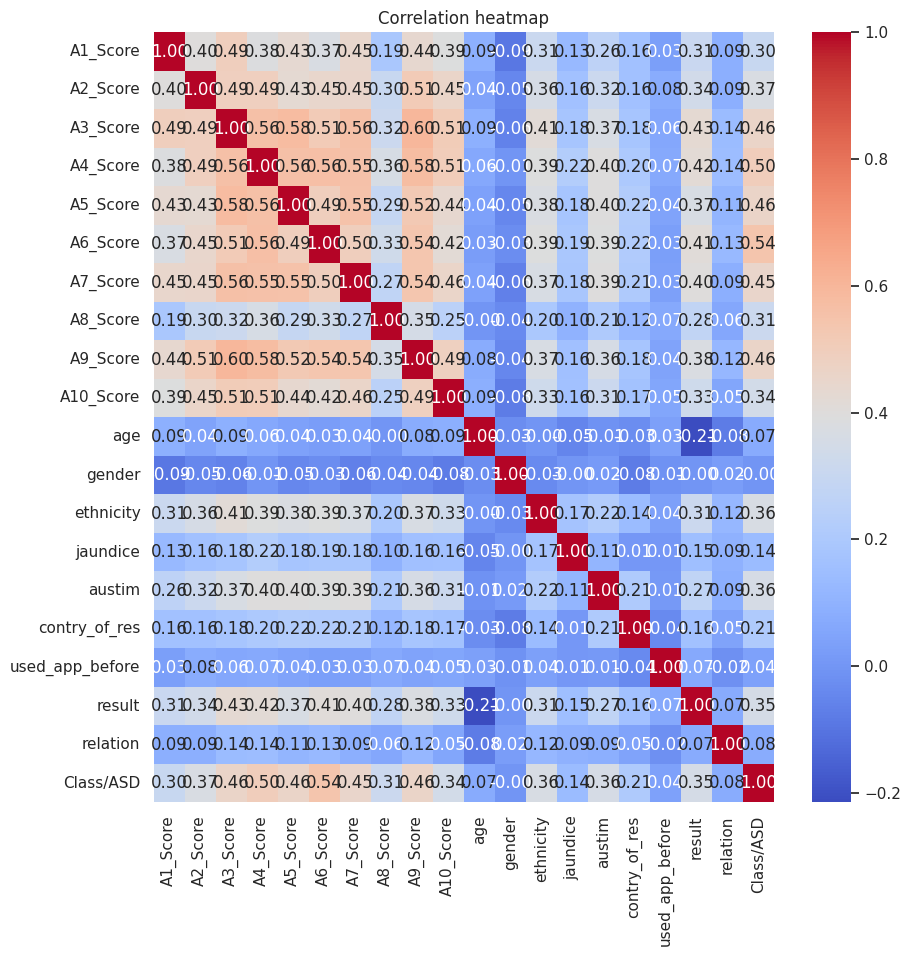
After training and evaluating the three machine learning models—Decision Tree, Random Forest, and XGBoost—on the test dataset, it was observed that the Random Forest Classifier consistently delivered the highest performance across all key evaluation metrics. While the Decision Tree classifier performed adequately, it showed signs of overfitting and comparatively lower generalization. The XGBoost model offered a balanced performance and better precision than the Decision Tree, but it was the Random Forest classifier that achieved the highest accuracy, as well as superior precision, recall, and F1-score. Its ensemble nature and robustness to noise and overfitting made it the most reliable model for the classification of Autism Spectrum Disorder (ASD) cases in this project.



**3.3 Visualization**

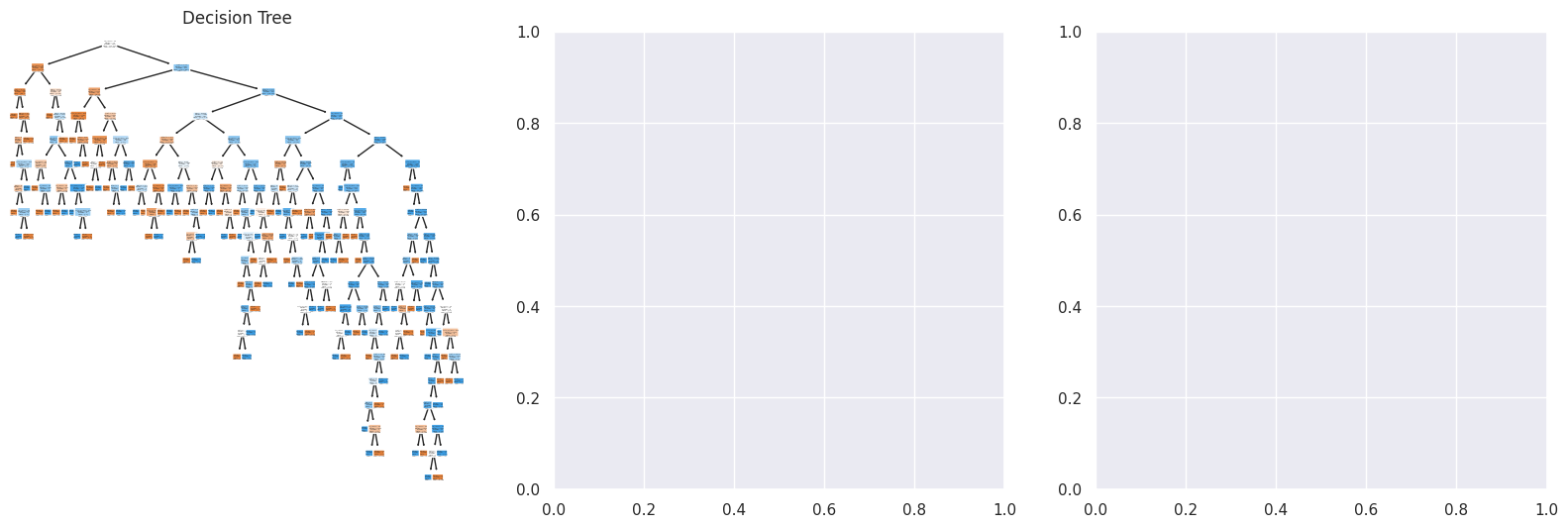
**3.3.1 Bivariate Analysis:**

Bivariate analysis involves exploring the relationships between two variables. In this project, it was used to understand how different categorical features relate to the target variable, Class/ASD. This analysis helps in identifying patterns and potential predictors for Autism Spectrum Disorder.

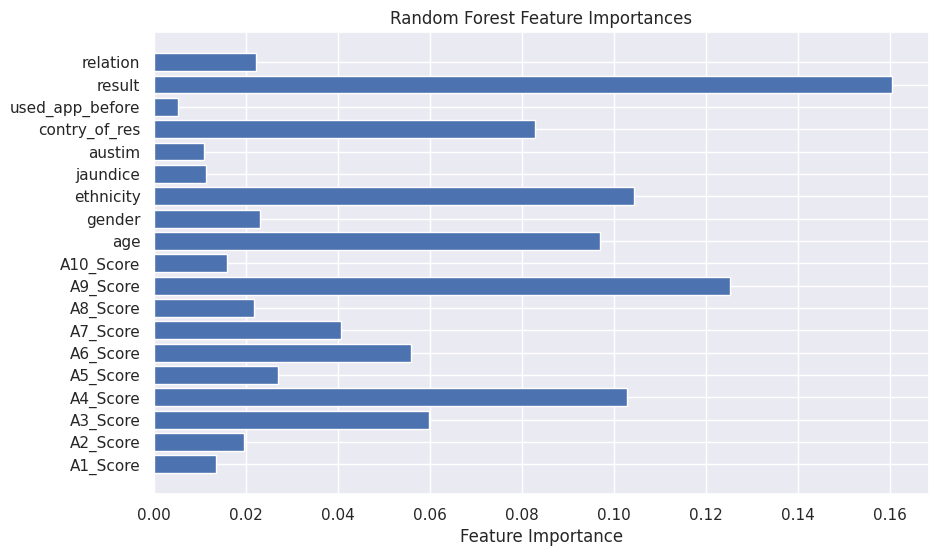


**3.3.2 Decision Tree Visualization:**

To enhance the interpretability of the Decision Tree model, we visualized its structure using a tree diagram. This visualization provides a clear understanding of how the model makes predictions based on different feature splits.

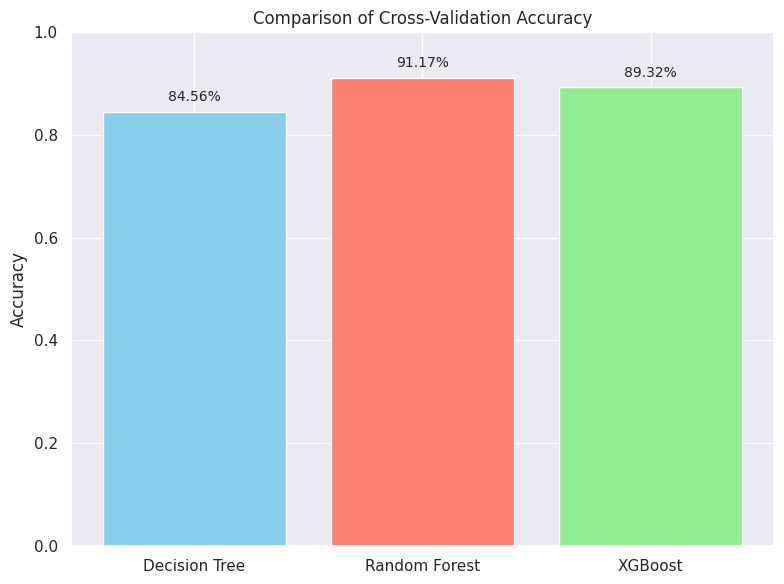


**3.3.3 Random Forest and XGBoost - Box Plots of CV Scores:**



**3.3.4 COMPARING THE DIFFERENT MODELS:**

Comparing different machine learning models is crucial to select the best-performing model for your specific task. Each model has its strengths and weaknesses, and the best choice depends on the dataset, the desired performance metrics, and the overall goals of the project.



**CHAPTER 4**

**CONCLUSION AND FUTURE SCOPE**

**Conclusion**

In this project, various machine learning models were implemented and evaluated to predict the likelihood of Autism Spectrum Disorder (ASD) using behavioral and demographic data. Among the models tested, the Random Forest Classifier delivered the most impressive performance, achieving an accuracy of 0.91 and a high R² score of approximately 0.96, indicating excellent predictive power and reliability. Its ensemble nature allowed it to handle noise, prevent overfitting, and generalize well across unseen data. Although the Decision Tree Classifier also performed reasonably well with an accuracy of 0.83, it was comparatively less robust

**Future Scope**

Although the Random Forest Classifier demonstrated excellent performance with a high R² score (~0.96) and an accuracy of 0.91, the system can be further enhanced by incorporating additional features such as genetic, behavioral, or neurological data to improve predictive power. In the future, the model could be deployed as a web or mobile application to assist healthcare professionals in early autism detection, especially in remote areas. Integrating real-time data, adding explainable AI techniques for better interpretability, and expanding the model to predict severity levels of ASD or other related conditions are also promising directions. Additionally, implementing continuous learning mechanisms can ensure that the model remains up-to-date with evolving data trends.

## REFERENCES

1. Scikit-learn Developers. (2023). Scikit-learn: Machine Learning in Python. Retrieved from <https://scikit-learn.org>
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DOI: 10.1145/2939672.2939785
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
4. Kaggle. (n.d.). Autism Screening Adult Dataset. Retrieved from <https://www.kaggle.com>
5. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
6. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830*.*

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru

**Dept of**

## CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

## Program Outcomes (POs)

**Engineering Graduates will be able to:**

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems anddesign system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant tothe professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need forsustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write

effective reports and design documentation, make effective presentations, and give and receive clear instructions.

1. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
2. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

## Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems. PSO2 : Design and develop web sites, web apps and mobile apps.

**PROJECT PROFORMA**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification of**  **Project** | **Application** | **Product** | **Research** | **Review** |
| √ |  |  |  |

**Note: Tick Appropriate category**

|  |  |
| --- | --- |
| **Machine Learning Outcomes** | |
| Course Outcome (CO1) | Describe machine learning and different forms of learning. |
| Course Outcome (CO2) | Use statistical learning techniques to solve a class of problems. |
| Course Outcome (CO3) | Build support vector machine for the given data to create optimal boundary  that best classifies the data. |
| Course Outcome (CO4) | Design neural networks to simulate the way human brain analyzes and processes information. |
| Course Outcome (CO5) | Solve classification problems using a decision tree. |

**Mapping Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CS3509 : MACHINE LEARNING** | | | | | | | | | | | | | | | |
| **Course Outcomes** | **Program Outcomes and Program Specific Outcome** | | | | | | | | | | | | | | |
| **PO 1** | **PO 2** | **PO 3** | **PO 4** | **PO 5** | **PO 6** | **PO 7** | **PO 8** | **PO 9** | **PO 10** | **PO 11** | **PO 12** |  | **PSO 1** | **PSO 2** |
| CO1 | 1 | 1 |  |  |  |  |  |  |  |  |  | 1 |  |  |  |
| CO2 | 1 |  |  |  |  |  |  |  |  |  |  | 1 |  |  |  |
| CO3 | 2 | 3 | 2 |  |  |  |  |  |  |  |  | 2 |  | 1 |  |
| CO4 | 2 | 2 | 3 | 2 |  |  |  |  |  |  |  | 2 |  | 2 |  |
| CO5 | 1 | 2 | 3 | 1 |  |  |  |  |  |  |  | 2 |  | 1 |  |

**Note: Map each Data Mining outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:**

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped