Analysis and Prediction of Autism Spectrum Disorder Using Machine Learning

[1] SATYABRATA PATRO,
[1]M.tech Assistant Professor, Department of CSE, Raghu Engineering College, Dakamarri, Andhrapradesh
Email: - Satyabrata.patro@raghuenggcollege.in

[2] SRI TEJA RAYALA,
[2] B Tech Student, Department of CSE, Raghu Institute of Technology, Dakamarri, Andhrapradesh
Email: - sritejarayala@gmail.com

[3] P. RAMESH,
[3] B Tech Student, Department of CSE, Raghu Institute of Technology, Dakamarri, Andhrapradesh
Email: - pramesh4613@gmail.com

[4] P. PUNEETH,
[4] B Tech Student, Department of CSE, Raghu Institute of Technology, Dakamarri, Andhrapradesh
Email: - parimipuneeth2002@gmail.com

[5] P. VARUN KUMAR,
[5] B Tech Student, Department of CSE, Raghu Institute of Technology, Dakamarri, Andhrapradesh
Email: - P.Varun2410@gmail.com

ABSTRACT

Early diagnosis and intervention are greatly hampered by autism spectrum disorder (ASD), which calls for creative methods of precise prediction. This study uses a variety of classification techniques to create a strong prediction model for ASD identification by utilizing data mining. We start by gathering and preparing a large amount of data, then we go through a painstaking process of feature selection and model training. We employ thorough preprocessing methods to make sure our data collection is ready for analysis. By using a rigorous feature selection procedure, we determine which features are most discriminative and important for ASD prediction. Our study carefully assesses the effectiveness of a suite of cutting-edge classification algorithms, including KNearestNeighbor, NaiveBayes, SGDClassifier, Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, Support Vector Machine, and AdaBoost Classifier, in predicting ASD. To ensure optimal performance, each algorithm is painstakingly trained and adjusted using the training data set. A wide range of performance criteria, such as accuracy, sensitivity, specificity, precision, and F-measure, are used to thoroughly evaluate the prediction models. We identify the advantages and disadvantages of each algorithm by doing a thorough experimentation and comparative study, ultimately determining which model is the most proficient for ASD prediction. Our results highlight the revolutionary potential of data mining in the field of healthcare, especially when it comes to enabling early detection and intervention for ASD. This effort clarifies the performance landscape of various classification algorithms, which highlights the importance of using data-driven approaches in healthcare decision-making and opens up new avenues for improving ASD prediction.

Keywords:-ASD, KNearest Neighbor, Naive Bayes, SGDClassifier, Decision Tree Classifier, Random Forest Classifier

1. INTRODUCTION

The process of going through massive databases to find patterns and linkages so that data analysis can be used to address problems is known as data mining. Using data mining technologies, businesses may forecast future trends. Numerous fields of study, including genetics, marketing, mathematics, and cybernetics, use data mining techniques. While data mining techniques can be used to increase productivity and forecast consumer behavior, predictive analysis can also be utilized to help a company stand out from the competition. The capacity to find hidden patterns and relationships in data that can be utilized to generate predictions that have an impact on enterprises is generally where data mining benefits originate from. Six typical task classes are involved in data mining: anomaly detection, association rule learning, summarization, regression, classification, and clustering. Data mining can, however, be inadvertently abused and result in seemingly important findings that are not predictive of future behavior and are not repeatable on fresh data samples, making them of little utility. This frequently happens when too many hypotheses are investigated and inadequate statistical hypothesis testing is done. Overfitting is a term used to describe a basic form of this issue in machine learning, while it can also occur at other stages of the procedure. Examining, purifying, converting, and modeling data in order to find relevant information, draw conclusions, and aid in decision-making is the process of data analysis. Data analysis is a multifaceted field that encompasses various techniques under numerous titles, and it is applied in various business, scientific, and social science sectors. Data analysis is essential in today's corporate environment for helping organizations operate more effectively and for making decisions more scientific. While business intelligence encompasses data analysis that primarily focuses on business information and extensively relies on aggregation, data mining is a specific type of data analysis that focuses on modeling and knowledge discovery for predictive rather than just descriptive reasons. Data analysis in statistical applications can be separated into three categories: exploratory data analysis (EDA), descriptive statistics, and confirming or falsifying existing hypotheses While text analytics employs statistical, linguistic, and structural techniques to extract and categorize information from textual sources—a type of unstructured data—predictive analytics focuses on the application of statistical models for predictive forecasting or categorization. These are all different types of data analysis. Finding hidden, legitimate, and maybe helpful trends in large databases is the goal of data mining. Finding unexpected or previously unknown relationships among the data is the main goal.

Autism Spectrum Disorder (ASD) is a developmental disorder characterized by challenges in communication and social interaction, often accompanied by repetitive behaviors. While ASD can be diagnosed at any age, symptoms typically manifest within the first two years of life, impacting daily routines and interactions within the surrounding environment.

This initiative uses classification techniques to try and determine whether children show symptoms of ASD. Furthermore, it uses performance metrics including accuracy, sensitivity, specificity, precision, and F-measure values to determine which classification system predicts ASD the best. Data cleaning using mean approaches is the next step in the preparation phase of the acquired data set, which includes adult autism screening data. Label Encoder and OneHotEncoder are used to convert texts to numerical data. To ensure consistency and improve algorithm performance, the data set is then put through a series of feature scaling algorithms, such as StandardScaler, RobustScaler, QuantileTransformer, MinMaxScaler, MaxAbsScaler, Power Transformer, and Normalizer. Testing data is used to verify the accuracy of the model, whereas training data is used to identify patterns within the dataset. ASD prediction requires the application of feature selection approaches, such as the ones previously discussed, to find relevant features.

The chosen features are then used in Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, Support VectorMachine, AdaBoost Classifier, KNearestNeighbor, NaiveBayes, and SGDClassifier, among other classification methods. The training data is used to train these algorithms, and then tests are conducted to determine how well they predict ASD. The efficacy of the classification algorithm is evaluated by taking into account important factors including F-measure values, sensitivity, specificity, and accuracy. The algorithm that performs better on all of these metrics is the one that is most appropriate for ASD prediction.

This research aims to improve early identification and intervention options for afflicted individuals by improving ASD diagnosis methodologies through rigorous testing and evaluation.

2. LITERATURE SURVEY

2.1 In 2016, Osman Altay, Mustafa Ulas presented a research paper on title was **Prediction of the Autism Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbor in Children**. This proposed system used to detect Autism patient using data mining concepts. Autism Spectrum Disorder (ASD) negatively affects the whole life of people. The main indications of ASD are seen as lack of social interaction and communication, repetitive patterns of behavior, fixed interests and activities. In this paper, it was tried to find out whether children have ASD by using classification methods. As a result of the classification, there are two classes of cases in which the child is ASD or not ASD.90.8% accuracy was obtained as a result of the LDA algorithm and 88.5% accuracy was obtained from the KNN algorithm.

This paper using two methodology Linear Discriminant Analysis Classifier and K-Nearest Neighbor for classification which is used to predict the ASD diagnosis.The data set consists of 292 samples with 19 different attributes. In the dataset, there are 10 questions directly related to ASD and a score attribute consisting of the sum of these questions. Ethnicity and country of residence which have a string value has been transformed to numerical values to make it suitable for LDA and KNN algorithms.The KNN algorithm is mainly based on the distance calculation. In the LDA algorithm, it is necessary to calculate the scatter matrices within classes and between classes. Performance Evaluation is calculated for test the classification algorithms as accuracy, sensitivity, specificity, precision, and F-measure.

The advantage of this paper is F-measure value of LDA algorithm attained 1.95% better success rate than the KNN algorithm. The F-measure value is calculated as 0.9091 for the LDA algorithm and 0.8913 for the KNNalgorithm.The disadvantage of this paper is Increased number of feature extracted from children During data collection, which takes more time. In Proceedings of the International Journal of Advance Research in Science and Engineering IJARSE.

2.2 In 2010, Siriwan Sunsirikul, Tiranee Achalakul presented a paper on the title **Associative Classification Mining in the Behavior Study of Autism Spectrum Disorder**. This proposed system aim is to develop a data analysis tool to aid doctors in the diagnosis process in the future. In this research, attempted to extract patterns from behavioral data and develop a classifier for patients' behaviors. A sufficient number of patients' behavior records, it maybe possible to discover the association between some particular behaviors and the autistic symptoms. This paper discusses data mining techniques aimed at providing an array of tools to assist doctors in analyzing patients' data intelligently.

The proposed methodology is an associative classification method. A classification-based association(CBA) technique is used to find association of behavioral patterns for autistic and PDD-NOS children. The results present some useful information that can be used in the future to guide clinicians in selecting appropriate treatments, which in turn can help autistic children function better in a society as well as enable early detection and intervention. CBA includes two main parts: (1) a rule generator (CBA-RG) that is used to generate a complete set of class association rules (2) a classifier builder (CBA-CB) that is used to produce a classifier. Patients' behavior records will be used as input. The output is a set of accuracy rules with support and confidence measures.

The advantage of this paper is the relationship of the behavior pattern for autistic and PDD NOS children can be identified. a set of impairments, a disorder type can be suggested with a relatively high confidence level. The disadvantage is Prediction error in some cases because small number of samples tend to overfit the solution. Lack of clinical data of normal children for use in the training phase. In Proceedings of the The 2nd International Conference on Computer and Automation Engineering (ICCAE).

2.3 In 2017, FadiThabtah presented a paper on the title is **Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment**. This proposed system aim is to Reducing the screening time, improving sensitivity and specificity, Identifying the smallest number of ASD codes to simplify the problem. ASD diagnosis is considered a typical classification problem in machine learning in which a model is constructed based on previously classified cases and controls. This model can then be employed to guess the new case diagnosis type (ASD, No-ASD).

This paper solving ASD diagnosis as a classification problem. The input will be a training dataset of cases and controls that have already been diagnosed. The cases and controls have been generated using a diagnostic instrument such as ADOSR, ADI-R. The aim is identifying the best ASD features, or reducing computing resources used during the data processing. Once initial data is processed then a machine learning algorithm can be applied. here are different measures that the end user can use to evaluate the effectiveness of the chosen machine learning method on guessing the type of diagnosis.

In this paper, we focused on recent machine learning studies that tackled ASD as a classification problem and critically analysed their advantages and disadvantages. It showed the necessary steps required to claim the development of intelligent diagnostic tools based on machine learning by replacing the

handcrafted rules inside the ASD screening tools with a predictive model. In Proceedings of the 1st International Conference on Medical and Health Informatics 2017 (pp. 1-6)

2.4 In 2015, Mohana e1, Poonkuzhali.s2 presented a paper on the title is **Categorizing the risk level of autistic children using data mining techniques**. Autism spectrum disorders (ASD) are enclosure of several complex neurodevelopmental disorders characterized by impairments in communication skills and social skills with repetitive behaviors. It is widely recognized for many decades, yet there are no definitive or universally accepted diagnostic criteria. This paper focuses on finding the best classifier with reduced features for predicting the risk level of autism. The dataset is pre-processed and classified. It produced high accuracy of 95.21% using Runs Filtering.

This paper using four feature selection algorithms and several classification algorithms. feature selection algorithms such as Fisher filtering, ReliefF, runs filtering and Stepdisc are used to filter relevant feature from the dataset. Ball Vector Machine, CS-CRT, Core Vector Machine, K-Nearest Neighbor classification algorithms are applied on this reduced features. Finally, performance evaluation is done on all the classifier results.Finally, Error rate, Accuracy, Recall, Precision is calculated.

The advantage of this paper is BVM (ball vector machine), CVM (core vector machine), and MLR achieved high accuracy of 95.21%. The disadvantage of this paper is Fisher Filtering and Stepdisc does not filter any features.In Proceedings of the International Journal of Advance Research in Science and Engineering IJARSE.

2.5 In 2016 Khalid Al-jabery1, Tayo Obafemi-Ajayi1, Gayla R. Olbricht 2, T. Nicole Takahashi3, Stephen Kanne3 and Donald Wunsch1 presented a paper on title is **Ensemble Statistical and Subspace Clustering Model for Analysis of Autism Spectrum Disorder Phenotypes**. The results provide useful evidence that is helpful in elucidating the phenotype complexity within ASD. Our model can be extended to other disorders that exhibit a diverse range of heterogeneity In this paper, present an ensemble model for analyzing ASD phenotypes using several machine learning techniques and a k-dimensional subspace clustering algorithm. Our ensemble also incorporates statistical methods at several stages of analysis. A key phase in any clustering framework is feature selection. Ensemble statistical and clustering model to analyze a population of ASD patients. The ensemble model consists of five phases. Data Processing, Correlation Analysis, Uni-dimensional Clustering, k-dimensional Clustering, Clusters Evaluation.

The advantages of this method isfive stages of statistical and machine learning approaches, to achieve a subspace clustering of ASD data. The clustering results show promise for sorting out the heterogeneity that ischaracteristic of these patients. Multiple techniques were also combined for the validation of the identified clusters.In Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

2.6 In 2018, Fatiha Nur Buyukoflaz, Ali Ozturk presented a paper on the title is **Early Autism Diagnosis of Children with Machine Learning Algorithms**. Autistic Spectrum Disorder (ASD) is a neuro-developmental disorder that is one of the major health problems and its early diagnosis is of great importance for controlling the disease. In this existing system aim to performance comparisons using the machine learning classifying method. As a result of the experiment, it was shown that Random Forest method was more successful than Naive Bayes, IBk and RBFN methods.This paper using four machine learning methodology such as Naive Bayes classifier, IBk (k-nearest neighbours) classifier, Random Forest classifier, RBFN (radial basis function network)classifier.

The KNN, which aims to classify a new instance of x, selects the closest examples in the education database.NB, which is assumed to be based on the properties of the class. The Random Forest algorithm is an integrated class consisting of a set of decision tree classifiers. The advantage of this paper is Random Forest achieved 100% accuracy. Disadvantage of this paper is IBk achieved 89.65% accuracy.

2.7 In 2013, Mengwen Liu, Yuan An, Xiaohua Hu, Debra Langer, Craig Newschaffer, Lindsay Shea presented a paper on the title is **an evaluation of identification of suspected Autism Spectrum Disorder(ASD) cases in early intervention (EI) records**.This paper aim is toused EI records to evaluate classification techniques to identify suspected ASD cases. It improved the performance of machine learning techniques by developing and applying a unified ASD ontology to identify the most relevant features from EI records. It shows that developing automatic approaches for quickly and effectively detecting suspected cases of ASD from non standardized EI records earlier than most ASD cases are typically detected is promising.

This paper using three classification algorithms. Such as Naïve Bayes (NB), Bayesian Logistic Regression (BLR), and Support Vector Machine (SVM). Data preprocessing and feature selection techniques also used in this paper. Naïve Bayes uses probability distribution of features to estimate the label of an instance by assuming the independency between features. Bayesian Logistic Regression is extended from a logistic regression model by adding a prior probability distribution. Support Vector Machine thebasic idea is to find a maximum marginal hyperplane, which gives the largest separation between classes.

The advantage of this paper is that resultsindicate that Information improve the performance of an SVM classifier. it shows that text classification from EI records is a real possibility and could be useful to state EI systems. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine.

3. Implementation Study

The existing system described in the selected base paper focuses on data classification using various machine learning techniques. It involves two main steps: data classification and evaluation metrics. In the data classification step, a learning process is conducted where a model is built based on a set of data instances belonging to predefined classes. This model is then tested using different machine learning techniques to assess its classification accuracy, AUC value, precision, recall, and F1 score. The classifiers employed in this system include Naïve Bayes, K-Nearest Neighbor (kNN), Logistic Regression, Gradient Boosting, Support Vector Machine, Decision Tree, and MLP Classifier.

3.1 Proposed Methodology

This paper using four methodologies for predicting heart disease such as Decision tree, Support vector machine (SVM) Neural network, K-nearest neighbor algorithm. It can handle multi-dimensional data. It still suffering by repetition and replication. Therefore, some steps are needed to handle repetition and replication. Attribute selection is used to improve the performance of this technique. It is used to classify both linear and non-linear data. It classifies into two classes. Hyper-plane is used to separate the given classes. The classification task is performed by maximizing the margin of hyper-plane. Neural network consists of artificial neurons and process information. In neural network, basic elements are nodes or neurons. It can minimize the error by adjusting its weights and by making changes in its structure. KNN classification algorithm works by finding K training instances that are close to the unseen instance. This is done by using distance measurements such as Euclidean, Manhattan, maximum dimension distance, and others. Advantage of this paper is that Support Vector Machine (SVM) technique is an efficient method for predicting heart disease. It gives good accuracy by observing various research papers.In Proceedings of 2018 Conference on Emerging Devices and Smart Systems (ICEDSS)

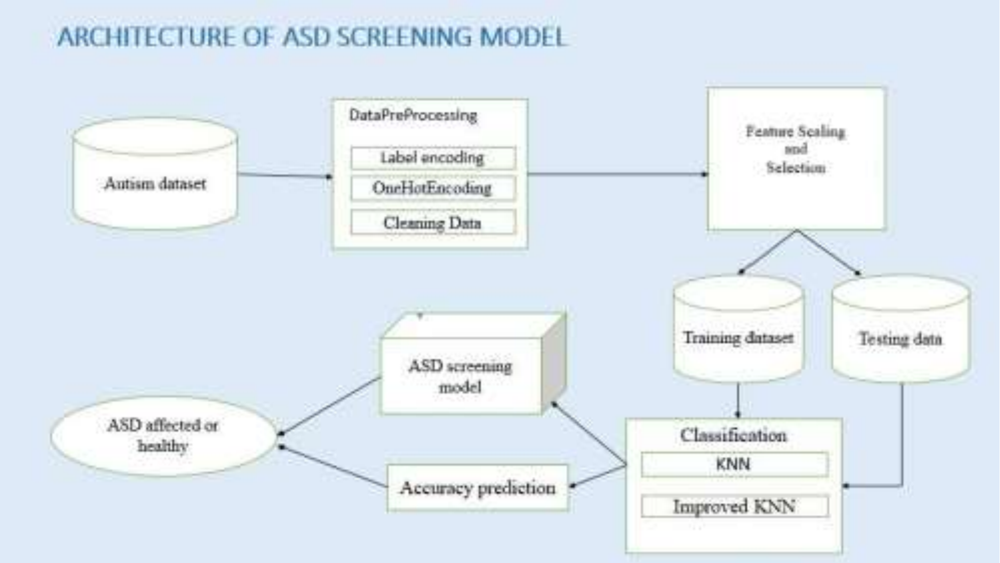


Fig 1:- Proposed System Architecture

The system architecture of our ASD prediction model is designed to streamline the process of data preprocessing, feature selection, model training, and prediction. At the core of the architecture lies a robust data pipeline that ingests raw data containing behavioral and clinical attributes of individuals. This data is then cleaned and transformed through various preprocessing modules, including data cleaning, encoding categorical variables, and feature scaling. These preprocessing steps ensure that the data is in a suitable format for analysis and model training. Subsequently, feature selection techniques such as chi-square are applied to identify the most relevant attributes for predicting ASD, enhancing the efficiency of the predictive model.

Once the data is preprocessed and feature-selected, the system leverages multiple classification algorithms, including Decision Tree, Random Forest, Support Vector Machine, and others, to train predictive models. Each classifier undergoes training using a portion of the preprocessed data and is evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. The architecture facilitates comparative analysis of different classifiers, allowing for the selection of the best-performing model for ASD prediction. Finally, the selected model is deployed to predict ASD diagnosis for new instances based on their behavioral and clinical attributes. The system architecture ensures scalability, flexibility, and accuracy, enabling efficient ASD prediction and aiding in early intervention and treatment planning.

4. METHODOLOGY & ALOGRITHAM

Module 1: Data Cleaning
Input: Raw data
Process: Cleaning the raw data by performing tasks such as handling missing values, removing duplicates, and addressing outliers.

Output: Cleaned dataset ready for further processing.

Module 2: Label Encoding
Input: Cleaned dataset
Process: Converting categorical variables into numerical values using label encoding.
Output: Dataset with categorical values encoded into numerical format.

Module 3: One-Hot Encoding
Input: Label encoded dataset
Process: Transforming categorical variables into binary vectors using one-hot encoding, creating separate columns for each category.
Output: Dataset with expanded columns for categorical variables.

Module 4: Feature Selection
Input: Dataset with encoded features
Process: Selecting relevant features using techniques such as Chi-square to predict outcomes.
Output: Dataset with selected features for prediction.

Module 5: Data Splitting

Input: Dataset with selected features
Process: Splitting the dataset into training and testing sets.
Output: Training and testing datasets.

Module 6: Classification with Various Algorithms
Input: Training and testing datasets
Process: Implementing classification algorithms such as DecisionTreeClassifier, RandomForestClassifier, ExtraTreesClassifier, SupportVectorMachine (SVM), AdaBoostClassifier, KNearestNeighbor (KNN), NaiveBayes, and SGDClassifier.
Output: Classification results for each algorithm.

Module 7: Feature Scaling
Input: Selected Features Data
Process: Scaling the features using techniques such as StandardScaler, RobustScaler, QuantileTransformer, MinMaxScaler, MaxAbsScaler, PowerTransformer, and Normalizer.
Output: Scaled dataset ready for classification.

Module 8: Prediction
Input: Scaled dataset
Process: Predicting outcomes using the trained classification models.
Output: Predicted results for each classifier.

Module 9: Performance Evaluation
Input: Predicted results
Process: Evaluating the performance of each classifier using metrics such as accuracy, precision, recall, and F1-score.
Output: Performance metrics for each classifier.

Module 10: Comparison and Selection
Input: Performance metrics of each classifier
Process: Comparing the performance of different classifiers to identify the one with the highest accuracy.
Output: Identification of the classifier with the best accuracy for the given dataset.

4.2 Algorithm

4.1.1 Random Forest:

Random Forest is a powerful ensemble learning method that employs multiple decision trees to enhance prediction accuracy and mitigate overfitting. During training, each tree is constructed using a random subset of the training data and features, ensuring diversity among the trees. When making predictions, the algorithm aggregates the results of individual trees through a majority voting mechanism, producing a robust final prediction. Moreover, it can provide insights into feature importance, aiding in the identification of relevant features for the classification task.

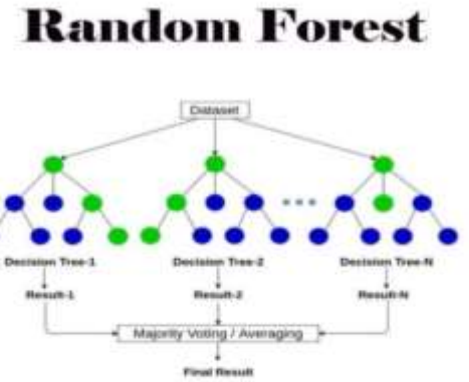


Figure 2: - Random Forest

4.1.2 K-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) is a simple yet effective algorithm used for classification and regression tasks in machine learning. It operates on the principle of similarity, where an unlabeled instance is classified based on the class labels of its nearest neighbors in the feature space. Here's how KNN works and its application in detecting phishing websites:

In the context of detecting phishing websites, KNN can be applied by first defining relevant features extracted from websites, such as URL characteristics, domain attributes, and content-based features. These features serve as dimensions in the feature space, and each website is represented as a point in this space.

During the classification phase, when a new website is encountered, KNN identifies its k nearest neighbors based on the similarity of their features to the features of the new website. The class labels of these neighbors are then used to predict the class of the new website. For example, if the majority of the k nearest neighbors are labeled as phishing websites, the new website is classified as phishing as well. By considering the characteristics and attributes of known phishing websites, KNN can effectively identify similarities between new and existing instances, enabling accurate classification decisions

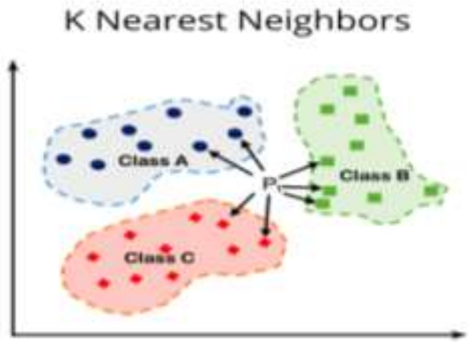


Figure 3: - K Nearest Neighbor

4.1.3 Naïve Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of feature independence. This means that it calculates the probability of a hypothesis (e.g., a website being phishing) given the observed evidence (e.g., features extracted from the website) using conditional probability. Despite its simplifying assumption, Naive Bayes is widely used in machine learning for classification tasks due to its simplicity and efficiency. By modeling the relationship between features and classes using probabilities, Naive Bayes can effectively classify instances based on their observed attributes.

In the context of detecting phishing websites, Naive Bayes proves to be useful for several reasons. Firstly, it can analyze various features extracted from websites, including URL attributes (e.g., length, presence of certain keywords), domain characteristics (e.g., age, registration information), and content (e.g., presence of suspicious links or language). This multifaceted analysis allows Naive Bayes to capture diverse patterns associated with phishing activity, contributing to its effectiveness in classification. Additionally, Naive Bayes requires minimal training data compared to more complex algorithms, making it particularly useful in scenarios where labeled datasets are limited or costly to obtain.

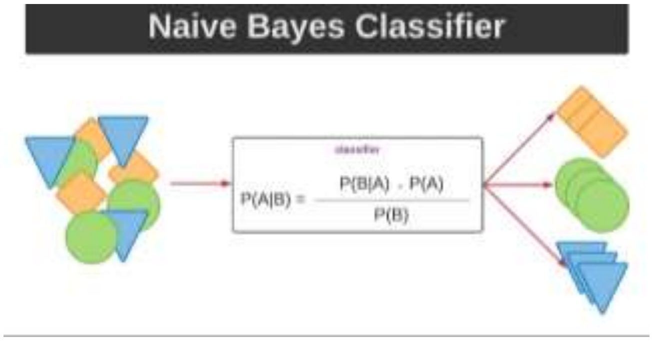


Figure 4: - Naïve Bayes

4.1.4 SGD Classifier

The Stochastic Gradient Descent (SGD) algorithm is an optimization technique commonly used in machine learning for training models, particularly in scenarios with large datasets. It's a variant of the gradient descent algorithm that updates the model parameters incrementally based on individual training examples (hence "stochastic").

Here's how the SGD algorithm works in a general sense:

Initialization: Initialize the model parameters (weights and biases) randomly or with some predefined values.

Iterative Optimization:

Iterate through the training data one example at a time (or in mini-batches).

For each example:

Compute the gradient of the loss function with respect to the model parameters using the current example.

Update the model parameters in the direction opposite to the gradient to minimize the loss function.

The update rule typically involves multiplying the gradient by a learning rate and subtracting it from the current parameter values.

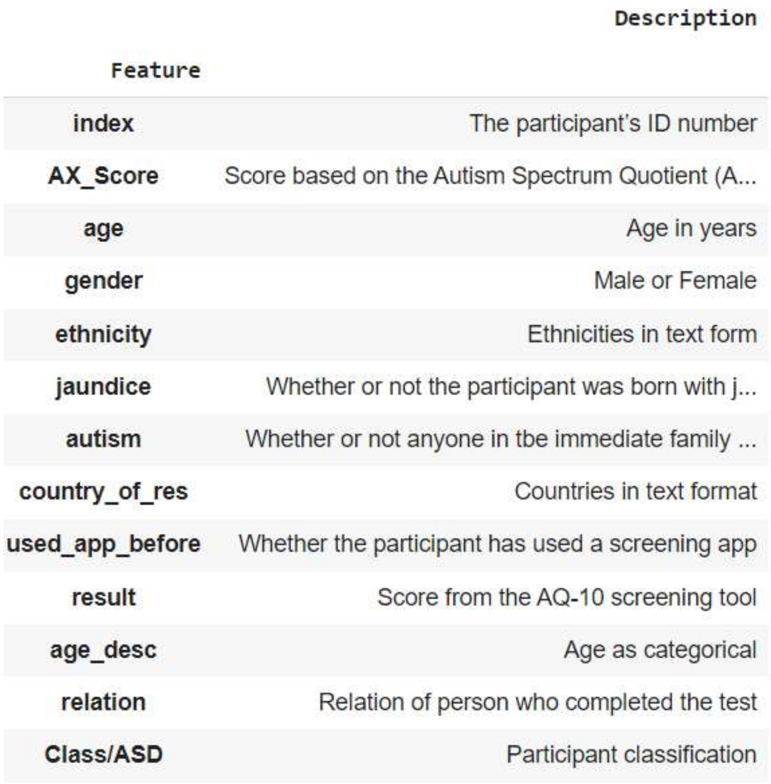
Convergence:

Repeat the iterative optimization process until a stopping criterion is met, such as a maximum number of iterations or when the improvement in the loss function becomes negligible.

The key advantages of using SGD include its ability to handle large datasets efficiently and its convergence properties. However, SGD can be sensitive to the choice of learning rate and may require careful tuning of hyperparameters for optimal performance.

```
Input: Training dataset D = {(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)}
    Learning rate alpha
    Number of iterations max_iter
Initialize model parameters w randomly or with zeros
for iter = 1 to max_iter do
  Shuffle the training dataset D
  for each example (x_i, y_i) in D do
    Compute the gradient of the loss function with respect to the parameters:
      grad = -2 * x_i * (y_i - w*x_i)
    Update the model parameters:
      w = w - alpha * grad
  end for
  end for
  Output: Learned parameters w
```

5. RESULTS AND DISCUSSION SCREEN SHOTS



	Description
Feature	
index	The participant's ID number
AX_Score	Score based on the Autism Spectrum Quotient (A...
age	Age in years
gender	Male or Female
ethnicity	Ethnicities in text form
jaundice	Whether or not the participant was born with j...
autism	Whether or not anyone in tbe immediate family ...
country_of_res	Countries in text format
used_app_before	Whether the participant has used a screening app
result	Score from the AQ-10 screening tool
age_desc	Age as categorical
relation	Relation of person who completed the test
Class/ASD	Participant classification

Fig 1:- Sample Dataset

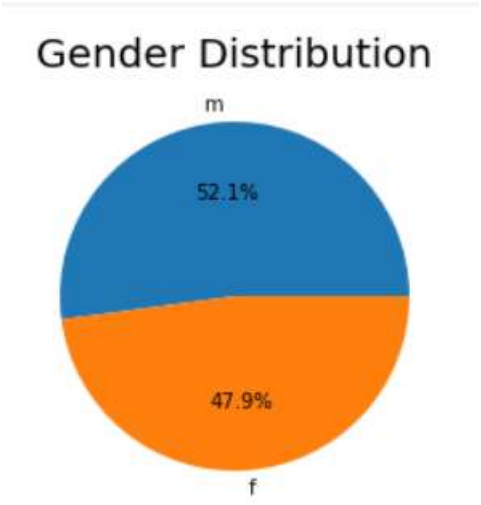


Fig 2:- Gender Distribution

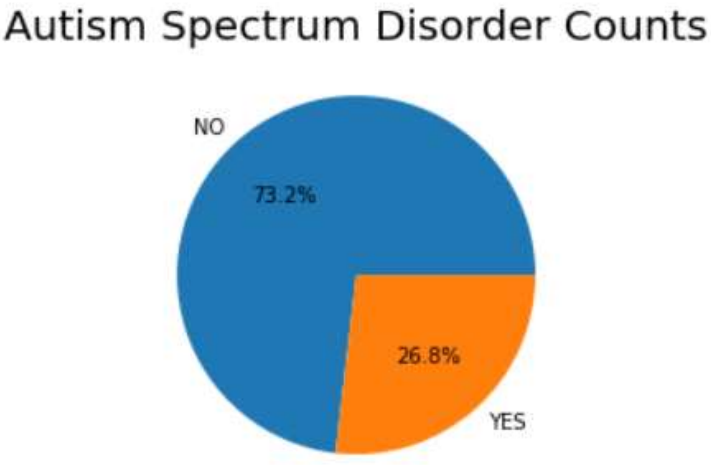
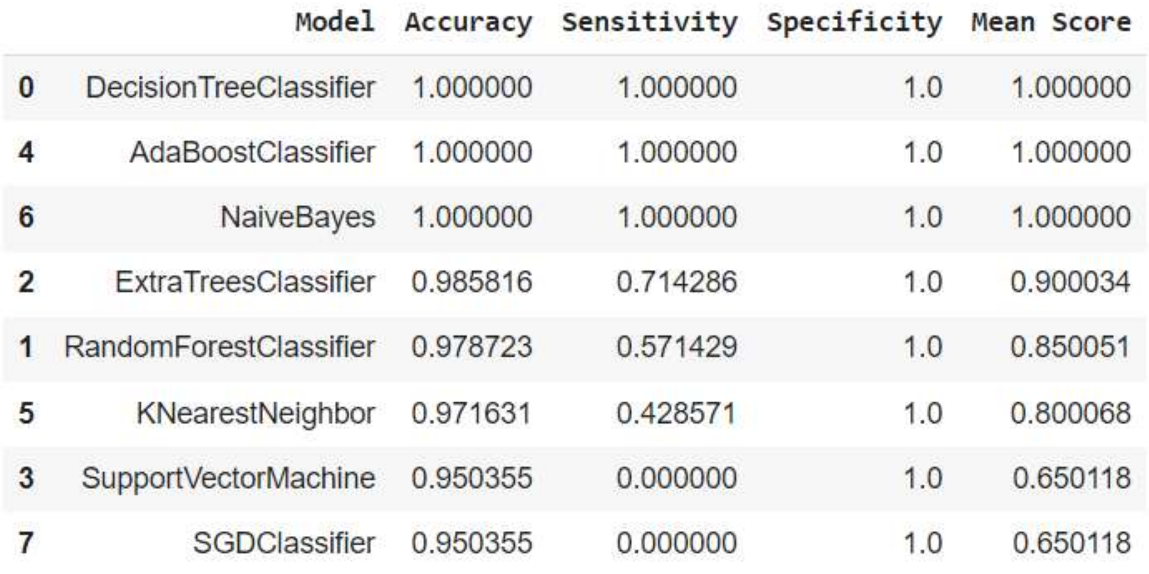


Fig 3:- ASD Count pie chart



	Model	Accuracy	Sensitivity	Specificity	Mean Score
0	DecisionTreeClassifier	1.000000	1.000000	1.0	1.000000
4	AdaBoostClassifier	1.000000	1.000000	1.0	1.000000
6	NaiveBayes	1.000000	1.000000	1.0	1.000000
2	ExtraTreesClassifier	0.985816	0.714286	1.0	0.900034
1	RandomForestClassifier	0.978723	0.571429	1.0	0.850051
5	KNearestNeighbor	0.971631	0.428571	1.0	0.800068
3	SupportVectorMachine	0.950355	0.000000	1.0	0.650118
7	SGDClassifier	0.950355	0.000000	1.0	0.650118

Fig 4:- Comparative Accuracy and Evaluation results of Machine Learning Algorithms

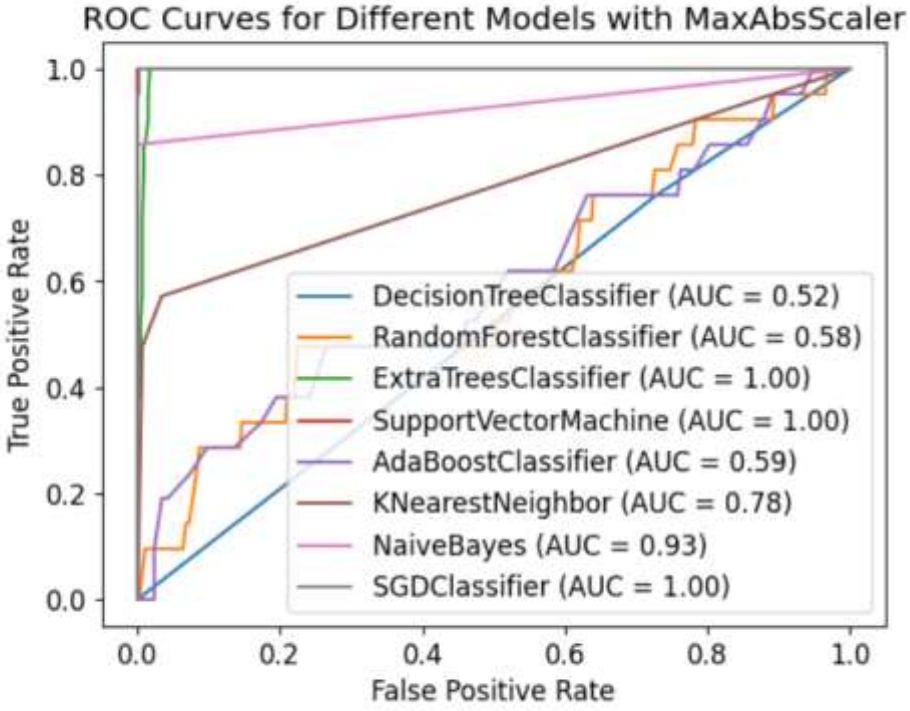


Fig 5:-ROC graph of different machine learning model using Max Abs Scaler feature selection technique

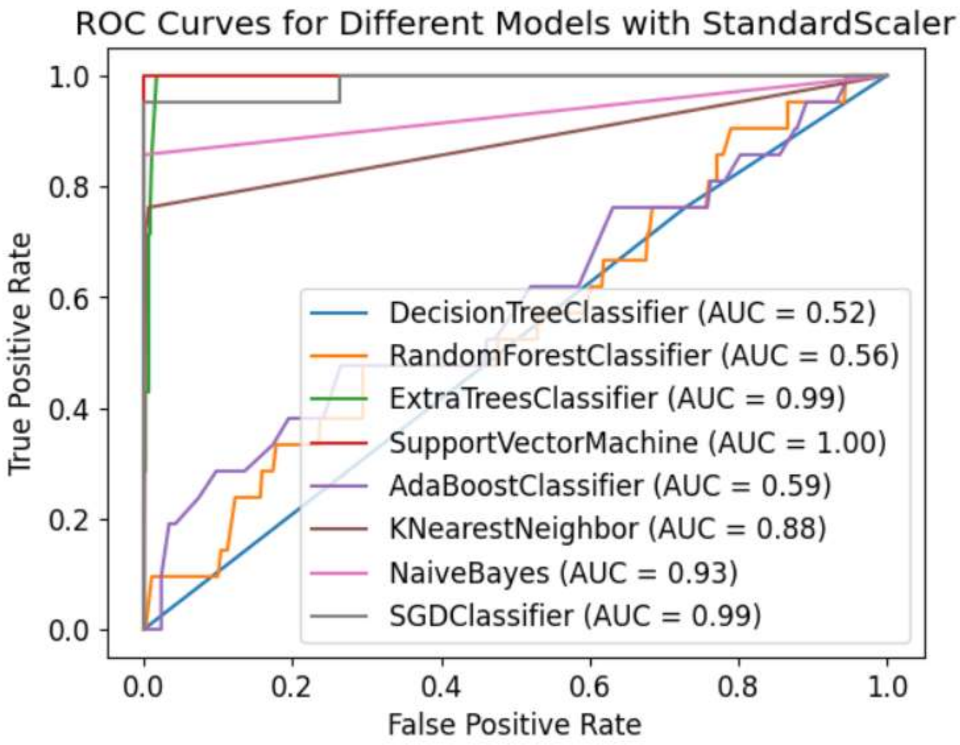


Fig 6:-ROC graph of different machine learning model using Standard Scaler feature selection technique

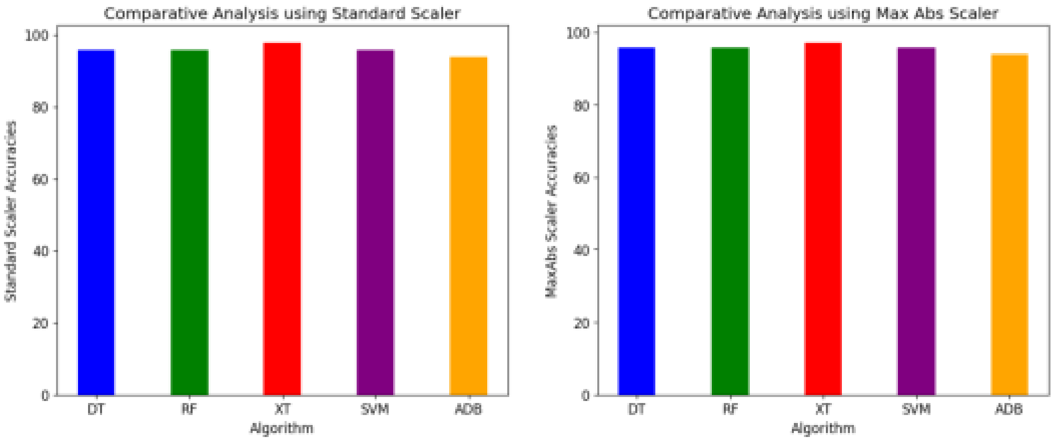


Fig 7:-Comparative Analysis Graph of different feature Selection Techniques

6. CONCLUSION

These accuracies demonstrate the effectiveness of our approach in accurately predicting ASD based on the selected features and classification algorithms. The high accuracy scores obtained from Extra Trees (XT) classifier highlight its robustness and suitability for this classification task. Additionally, the consistent performance across Decision Tree, Random Forest, and Support Vector Machine classifiers underscores the reliability of our predictive models.

By leveraging feature scaling techniques such as StandardScaler, RobustScaler, MinMaxScaler, and others, we were able to preprocess the data effectively and improve the performance of our classifiers. Furthermore, the inclusion of diverse classifiers allowed us to explore different modeling approaches and identify the most suitable ones for predicting ASD.

Overall, our project provides valuable insights into the application of machine learning techniques for ASD prediction, offering a potential tool for early diagnosis and intervention. Future work could focus on expanding the dataset, refining feature selection methods, and exploring advanced modeling techniques to further enhance the accuracy and applicability of the predictive models.

6. 1 Future Enhancements:

1. Integration of Deep Learning Models: Explore the integration of deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) for ASD prediction. Deep learning techniques have shown promising results in various healthcare applications and may offer improvements in accuracy and generalization.
2. Ensemble Learning: Implement ensemble learning techniques such as stacking or boosting to combine the predictions of multiple classifiers. Ensemble methods often lead to better performance by leveraging the strengths of individual models and mitigating their weaknesses.
3. Feature Engineering: Conduct more extensive feature engineering to extract relevant features from the dataset. Consider incorporating domain knowledge and exploring advanced feature selection methods to identify the most informative features for ASD prediction.
4. Cross-Validation and Hyperparameter Tuning: Perform thorough cross-validation and hyperparameter tuning to optimize the performance of the classifiers. Fine-tuning the hyperparameters of the models can lead to better generalization and robustness.
5. Incorporation of Additional Data: Expand the dataset by incorporating additional data sources such as genetic information, neuroimaging data, or behavioral assessments. Integrating diverse data modalities can provide a more comprehensive understanding of ASD and improve the predictive models.

6. Real-Time Prediction and Deployment: Develop a real-time prediction system that can quickly assess individuals for ASD based on their behavioral and clinical data. Deploy the predictive models in clinical settings or mobile applications to facilitate early diagnosis and intervention.
7. Interpretability and Explainability: Enhance the interpretability and explainability of the predictive models to gain insights into the underlying factors contributing to ASD prediction. Utilize techniques such as feature importance analysis or model explainability methods to interpret the model's decisions.
8. Longitudinal Data Analysis: Explore longitudinal data analysis techniques to track the progression of ASD symptoms over time and predict future outcomes. Longitudinal studies can provide valuable insights into the developmental trajectory of ASD and inform personalized intervention strategies.

7 REFERENCES

[1] 1.Linstead, Erik, Ryan Burns, Duy Nguyen, and David Tyler. "AMP: A platform for managing and mining data in the treatment of Autism Spectrum Disorder." In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, pp. 2545-2549. IEEE, 2016.

[2].Mohana, E. "Poonkuzhali. S,“Categorizing The Risk Level Of Autistic Children Using Data Mining Techniques”." International Journal of Advance Research In Science And Engineering IJARSE 4: 223-230.

[3].Sunsirikul, Siriwan, and Tiranee Achalakul. "Associative classification mining in the behavior study of Autism Spectrum Disorder." In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, vol. 3, pp. 279-283. IEEE, 2010.

[4].Raju, Cincy, E. Philipsy, Siji Chacko, L. Padma Suresh, and S. Deepa Rajan. "A Survey on Predicting Heart Disease using Data Mining Techniques." In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 253-255. IEEE, 2018.

[5].Altay, Osman, and Mustafa Ulas. "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children." In Digital Forensic and Security (ISDFS), 2018 6th International Symposium on, pp. 1-4. IEEE, 2018..

[6].Pattini, Elena, and Dolores Rollo. "Response to stress in the parents of children with autism spectrum disorder." In Medical Measurements and Applications (MeMeA), 2016 IEEE International Symposium on, pp. 1-7. IEEE, 2016.