Predicting and classifying the Levels of Autism Spectrum Disorder

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Technology Degree in Information Technology
By

ABHISHEK SINGH - 38120003
VAIBHAV VISHAL - 38120087

DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MAY 2022

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of ***Abhishek Singh (38120003) & Vaibhav Vishal (38120087)*** who carried out the project entitled ***"PREDICTING AND CLASSIFING THE LEVELS OF AUTISM SPECTRUM DISORDER"*** under my supervision from Jan 2022 .to Apr 2022

Internal Guide
Dr. Y. Bevish Jinila. Y, ME., Ph.D.,

Head of the Department
Dr. R. Subhashini, ME., Ph.D.,

Submitted for Viva voce Examination held on 02nd May,2022

Internal Examiner
(Name in Capital letters with Signature)

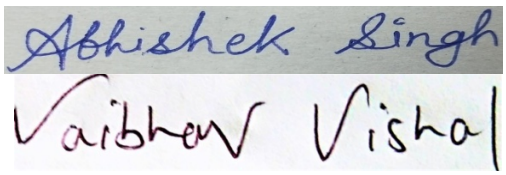
External Examiner
(Name in Capital letters with signature)

DECLARATION

I ***Abhishek Singh & Vaibhav Vishal*** hereby declare that the Project Report entitled "**PREDICTING AND CLASSIFING THE LEVELS OF AUTISM SPECTRUM DISORDER**" done by us under the guidance of **Dr. *Bevish Jinila*** at **Sathyabama Institute of Science and Technology** is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in ***INFORMATION TECHNOLOGY***.

DATE: 18-04-2022

PLACE: Chennai **SIGNATURE OF THE CANDIDATE**



ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to ***Dr. T. Sasikala M.E. , Ph.D.,*** *Dean,* School of Computing *and* **Dr. R. Subhashini, M.E., Ph.D.,** Head of the Department, Dept. of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. Bevish Jinila, M.E., Ph.D,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the ***Department of Information Technology*** who were helpful in many ways for the completion of the project.

ABSTRACT

Data Mining is defined as extracting information from huge sets of data. Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analyze data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. Data mining involves effective data collection and warehousing as well as computer processing. Data mining is also known as Knowledge Discovery in Data (KDD). The knowledge discovery in databases (KDD) process is commonly defined with the stages of Selection, Pre-processing, Transformation, Data mining, Interpretation/evaluation. Data mining is useful for discovering patterns and relationships in data to help make better decisions. The most popular algorithm used for data mining are classification and regression algorithm, which are used to identify relationships among data elements. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively.

This project aims for which classification algorithm is best suitable for predicting Autism spectrum disorder based on Accuracy, sensitivity, specificity, precision, and F-measure values. Collected dataset values are preprocess using preprocessing technique. Then cleaning data values are separating as training data used to recognizes patterns in the data, and testing data used to verify the data. Then the feature selection algorithm is applied. The selected features areK-Nearest Neighbor. Then testing data is applied to the classified model for predicting Autism.

Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates with

a number of prominent packages as NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, as well as other open source software. It Support for multiple IPython consoles.Python is free, open-source software, and consequently anyone can write a library package to extend its functionality. Data science has been an early beneficiary of these extensions, particularly Pandas, the big of them all. Pandas is the Python Data Analysis Library, used for everything from importing data from Excel spreadsheets to processing sets for time series analysis. Pandas puts pretty much every common data munging tool at your fingertips. This means that basic clean up and some advanced manipulation can be performed with Pandas' powerful data frames.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	i
	LIST OF FIGURES	iv
	LIST OF TABLES	iv
	LIST OF ABBREVIATIONS	v
1	**INTRODUCTION**	**1**
	1.1. About Domain	1
	1.2. About Project	3
	1.3. Organization of the Report	4
2.	**LITERATURE SURVEY**	**5**
3.	**SYSTEM DESIGN (PRILIMINARY DESIGN)**	**15**
	3.1 Overall System Architecture	15
	3.2 Process Flow Diagram	16
	3.3 System requirements	17
4	**DETAILED SYSTEM DESIGN**	**20**
	4.1 Data flow diagram	21
	4.2 UML	29
5	**EXPERIMENTAL RESULTS**	32
	5.1 Pre-Processing	32

5.2 Feature Selection 33
5.3 Splitting Data 34

6 **SUMMARY AND CONCLUSIONS** 41

6.1 Performance Analysis 41

REFERENCES 45

APPENDIX 54

A.SCREENSHOTS 43
B. SOURCE CODE 47

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
3.1	Overall System Architecture Design	15
3.2	Process Flow Diagra	16
4.1	Data Flow Diagram	22
4.2	Sequence Diagram	30
5.1	Graph	38

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO.
S.1	Accuracy	43
S.2	Module-1	43
S.5	Ethnicity Analysis	45
S.6	Heat Map	45

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
✓ EDA	exploratory data analysis
✓ CDA	confirmatory data analysis
✓ ASD	Autism spectrum disorder
✓ BVM	Ball vector machine
✓ CVM	Core vector machine
✓ SVM	Support vector machine
✓ IDE	Integrated development environment

CHAPTER – 1
INTRODUCTION

1.1 ABOUT DOMAIN

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. Data mining techniques are used in many research areas, including mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to drive efficiencies and predict customer behavior, if used correctly, a business can set itself apart from its competition through the use of predictive analysis. In general, the benefits of data mining come from the ability to uncover hidden patterns and relationships in data that can be used to make predictions that impact businesses. Data mining involves six common classes of tasks such as anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization. Data mining can unintentionally be misused and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this result from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as overfitting, but the same problem can arise at different phases of the process.

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. In today's business, data analysis is playing a role in making decisions more scientific and helping the

business achieve effective operation. Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information .In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses.

Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis. Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

1.2 ABOUT PROJECT

Autism spectrum disorder (ASD) is a developmental disorder that affects communication and normal behavior of the human beings. Although autism can be diagnosed at any age, it is said to be a “developmental disorder” because symptoms generally appear in the first two years of life. ASD people have difficulty with communication and interaction with other people. They have interest withrestrictive mannerand their behaviors may be repetitive. Their daily routine life in the surrounding environment.

This project aims to find out whether the children have ASD by using classification methods and also find out theclassification algorithm which is best suitable for predicting ASD based on Performance measurements such as Accuracy, sensitivity, specificity, precision, and F-measure values. The Collected Autism screening Adult dataset values are undergone preprocessing technique like conversion of Tring to numerical data using LabelEncoder and OneHotEncoder and then mean technique for cleaning the data. Training data is used to recognizes patterns in the data, and testing data used to verify the data. Thenthe feature selection technique (chi square) are applied to the dataset for selecting important features. The selected features are classified using K-Nearest Neighbor. Then testing data is applied to the classified model for predicting Autism. All trained algorithms are then applied for testing.

The performance of the classification algorithm is evaluated using parameters such as Accuracy, Sensitivity, Specificity, and F-measure values. Finally, the classification algorithm with high performance values considered as a most suitable classification algorithm for ASD.

1.3 ORGANIZATION OF THE REPORT

The overall project report is organised as follows.

Chapter 1: Introduction

Chapter 2: Literature Survey

This chapter is used about the literature survey, where the work related with ASD.

Chapter 3: System Design

This chapter provides overall system for the processed system.

Chapter 4: Detailed System Design

This chapter gives the detailed design for each process in the system.

Chapter 5: Experimental Results

This chapter discusses about the experimental results and analyse the performance of each classifier.

Chapter 6: Conclusion

Finally, this chapter concludes with overall summary of the ASD classification with future enhancement.

CHAPTER – 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.*

2.8 In 20016**,** Erik Linstead, Ryan Burns, Duy Nguyen, David Tyler presented a paper on the title is **AMP: A platform for managing and mining data in the treatment of Autism Spectrum Disorder**. This paper aim is to introduce AMP (Autism Management Platform), an integrated health care information system for capturing, analyzing, and managing data associated with the diagnosis and treatment of Autism Spectrum Disorder in children. AMP provides an intelligent web interface and analytics platform which allow physicians and specialists to aggregate and mine patient data in real-time, as well as give relevant feedback to automatically learn data filtering preferences over time.

This paper produced AMP, which implements a client-server architecture using a standard LAMP(Linux, Apache, MySQL, PhP) stack which is free and open source software. The AMP relies on open source software for its back endimplementation allows the system to be easily extensible. Data in the AMP system is managed using event feeds. The feed is where the user interacts with the items that have been posted, as well as communicates with the authorized users who have created those posts. If a particular user (a physician, for example)

is authorized for more than one child, he or she can easily switch to the context of a second or third child and view their feeds respectively. Items submitted to the feed are persisted in a secured relational database, and this same database manages the permissions that ensure only authorized users can see data associated for a particular patient. AMP utilizes existing Internet infrastructure and standards. Any transfer of sensitive data is performed over the HTTPS (HTTP / SSL) protocol.

The advantage of this paper is AMP fills a gap in existing autism technologies, and shows promise for improving the ease with which pertinent data can be collected, shared, and analyzed. Whether deployed in a clinic, school, or home, AMP and health information systems like it, provide a practical mechanism to improve the treatment of those living with an ASD diagnosis. *In Proceedings of 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*

2.9 In 2018**,** Thy Nguyen, Kerri Nowell, Kimberly E. Bodner, Tayo Obafemi-Ajayi presented a paper on the title is **Ensemble validation paradigm for intelligent data analysis in autism spectrum disorders.**In this paper aim is to apply varied clustering methods to subgroup an ASD simplex sample based on relevant phenotype features that may uncover meaningful subtypes. We present a detailed cluster validation analysis using an ensemble validation paradigm and visualization techniques. It presents a rigorous clinical/behavioral analysis of the top highly ranked results.

This paper using methodology such as Ensemble Validation, Normalization Methods in Cluster Analysis. Ensemble Validation addresses an important need to determine appropriate metrics for identifying optimal partitions in

clusteranalysisas Ensemble Validation. Our ensemble method selects the top result with highest aggregated ranks for further domain specific analysis. One of the objectives of this work is to evaluate the effect of the data preprocessing technique for normalization and missing values.

The advantage of this paper is method to cluster analysis of ASD phenotypes using different normalization techniques and multiple clustering algorithms. It presents a rigorous clinical/behavioral analysis of the top highly ranked results. *In Proceedings of IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).*

2.10 In 2018, Cincy Raju, E Philipsy, Siji Chacko, L Padma Suresh, S Deepa Rajan presented a paper on title is **A Survey on Prediction Heart Disease using Data Mining Techniques**. Heart disease is a most harmful one that will cause death. It has a serious long term disability. This disease attacks a person so instantly. The motivation of this paper is to develop an efficacious treatment using data mining techniques that can help remedial situations. There are many classification algorithms are used. Among these algorithms Support Vector Machine (SVM) gives best result.

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)*

CHAPTER 3

SYSTEM DESIGN (PRILIMINARY DESIGN)

This chapter gives complete overview of the system. The overall architecture of the system to figure in 3.1. The process flow diagram to figure in 3.2.

3.1 OVERALL SYSTEM ARCHITECTURE:

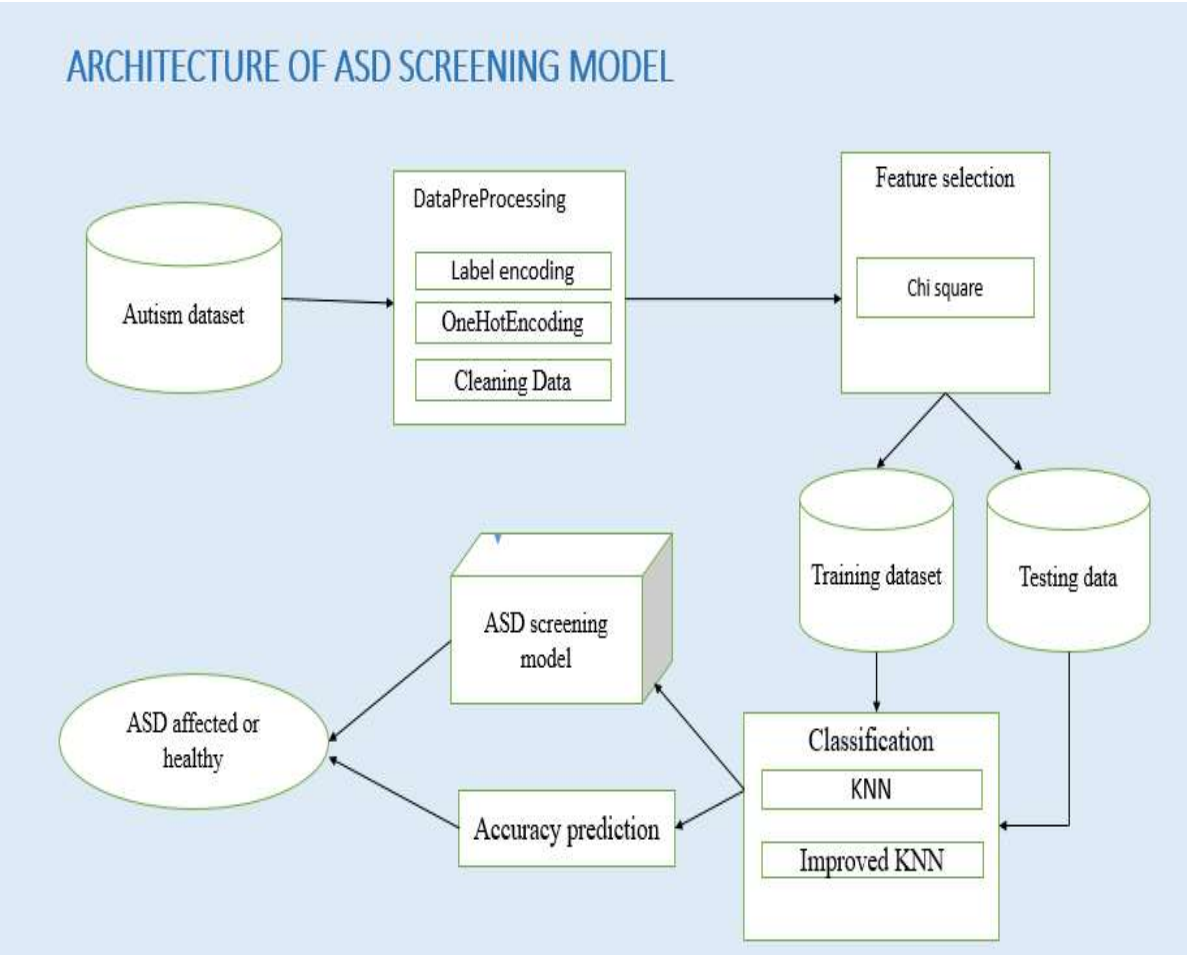


Fig – 3.1 Overall System Architecture Diagram

3.2 PROCESS FLOW DIAGRAM:

The process flow diagram is to demonstrate the flow of process in the project withmodule integrations

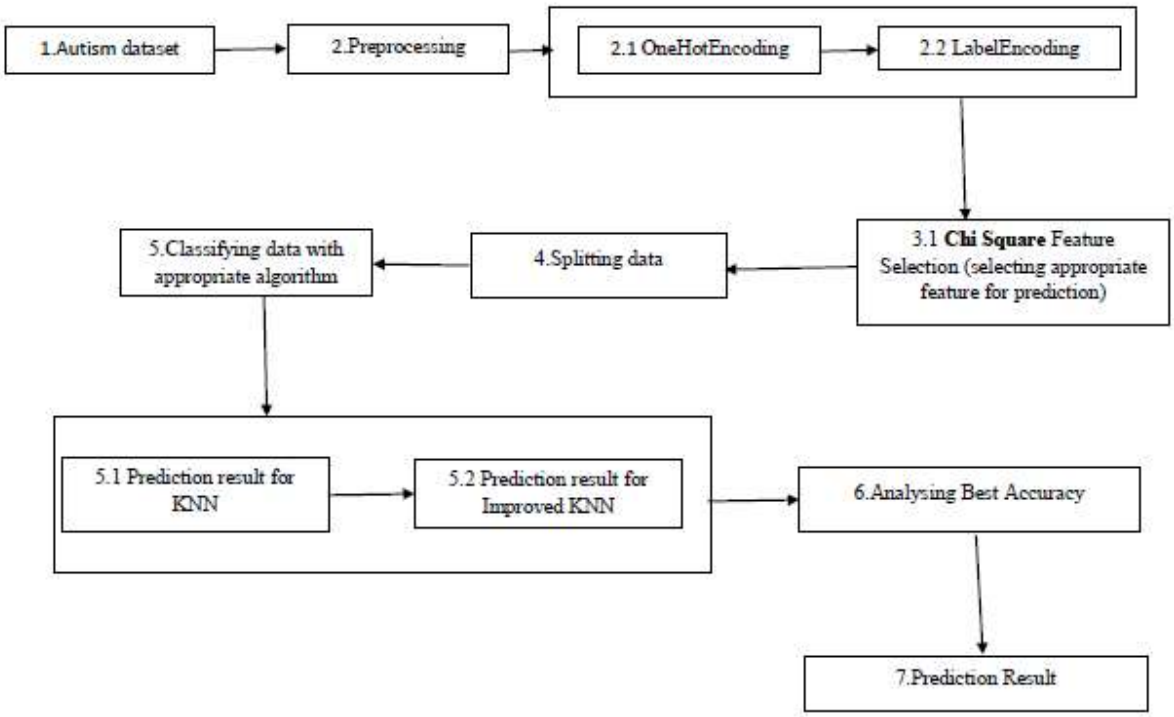


Fig 3.2: Process Flow Diagram

The modules in each phase aredetailed below: From Autism Screening data set we pre-processing the data with data cleaningtechniques, then Feature selection is for filtering irrelevant or redundant features from dataset,with that features we do classification using algorithms like SVM,LDM and KNN then we predict theresults, then we analyze which algorithm gives better accuracy.

3.3 SYSTEM REQUIREMENT

The system requirement is a main part in the analyzing phase of the project. The analyzer of the project has to properly analyze the hardware and the software requirements, otherwise in future the project designer will face more trouble with the hardware and software required. Below specified are the project hardware and software requirements.

3.3.1HARWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by the software engineers as the starting point for the system design. It shows what the system does and not how it should be implemented.

- SYSTEM : PENTIUM DUAL CORE
- HARD DISK : 120 GB
- MONITOR : 15” LCD
- INPUT DEVICE : KEYBOARD, MOUSE
- RAM : 8 GB

3.3.2SOFTWARE REQUIREMENTS

The software requirements document is the software specification of the system. It should include both a definition and a specification of a requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team

activities, performing tasks and tracking the team's progress throughout the development activity.

- OPERATING SYSTEM : WINDOWS 10
- PROGRAMMING LANGUAGE : PYTHON
- TOOLS : ANACONDA,SPYDER,PYTHON
- DATABASE : KERAS LIBRARY

3.3.3SOFTWARE FEATURES

PYTHON

Python features a dynamic type system and automatic memory management. It supportsmultipleprogrammingparadigms,including objectoriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open sourcesoftware and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming .Many other paradigms are supported via extensions, including design by contract and logic programming.

Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution, which binds method and variable names during program execution.Python's design offers some support for functional

programming in the Lisp tradition. It has `filter()`, `map()`, and `reduce()` functions; list comprehensions, dictionaries, and sets; and generator expressions. The standard library has two modules that implement functional tools borrowed from Haskell and Standard ML.

CHAPTER 4

SYSTEM DESIGN (DETAILED DESIGN)

This chapter explains the various modules of the system descriptions along with their Input, process flow, and output in an algorithmic way.

4.1 Module 1

From the raw data set we have cleaned the data for further process

Input: raw data

Process: Label encoding simply converting each value in a column to a number.

Output: categorical value changed into numeric value.

4.2 Module 2

Input: label encoded values

Process: separate each categorical value into separate column

Output: include many column as per the categorical value.

4.4 Module 3

Input: Cleaned dataset.

Process: Selecting appropriate features using Chi square which used to predict the results.

Output: Selected features for prediction.

4.5 Module 4

Input: Predicted feature columns.

Process: Splitting the data.

Output: Training and testing set.

4.6 Module 5.1

Classifying the data set using KNN

Input:Selected Features Data.

Process: Classifying the data using KNN algorithm.

Output: KNN Classification result.

4.7 Module 5.2

Classifying the data set using Improved KNN

Input: Selected Features Data.

Process: Classifying the data using Improved KNN algorithm.

Output: KNNClassification result.

4.9 Module 6

Predicting results

Input: Classification result

Process: Predicting results for each classifier

Output: Accuracy result

4.10 Module 7

Finding out the better accuracy

Input: Accuracy of each Algorithm

Process: Finding out the better accuracy

Output: better accuracy result.

SYSTEM DESIGN

The term "Design" is defined as the technical kernel of the software engineering process and is applied regardless of the development paradigm and area of application. Design is the first step in the development phase for any engineered product or system. The designer's goal is to produce a model or representation of an entity that will later be built.

4.1 DATA FLOW DIAGRAM

DFDs are used to Specify Functions of the Information System and how data flow from function to function. It's a collection of function that manipulates data. On a DFD, data items flow from an external data source or an internal data

store to an internal data store or an external data sink, via an internal process. A DFD provides no information about the timing of processes, or about whether processes will operate in sequence or in parallel.

4.1 DATA FLOW DIAGRAM AT THE INITIAL LEVEL

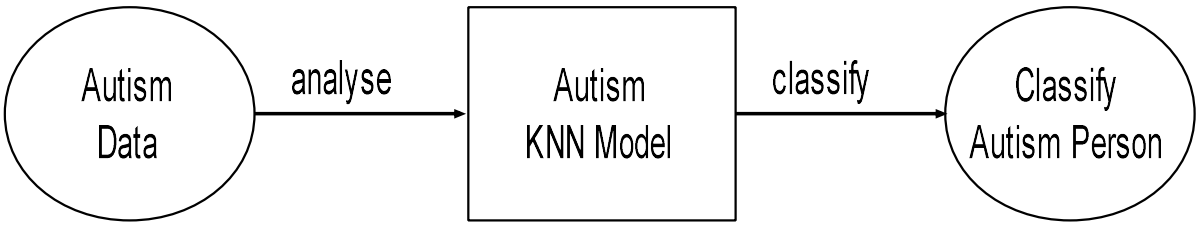


Fig 4.1 :Data Flow Diagram Level 0 for the entire project

Fig 4.1 explains the outer overall functionality of the proposed system application this indicates that have collected the weather dataset from various source and analysis and processing of data have done by predicting the autism and classifying the rate of accuracy from this information by plotting the graph is predicted.

4.2 DATA FLOW DIAGRAM LEVEL 1 FOR ENTIRE PROJECT

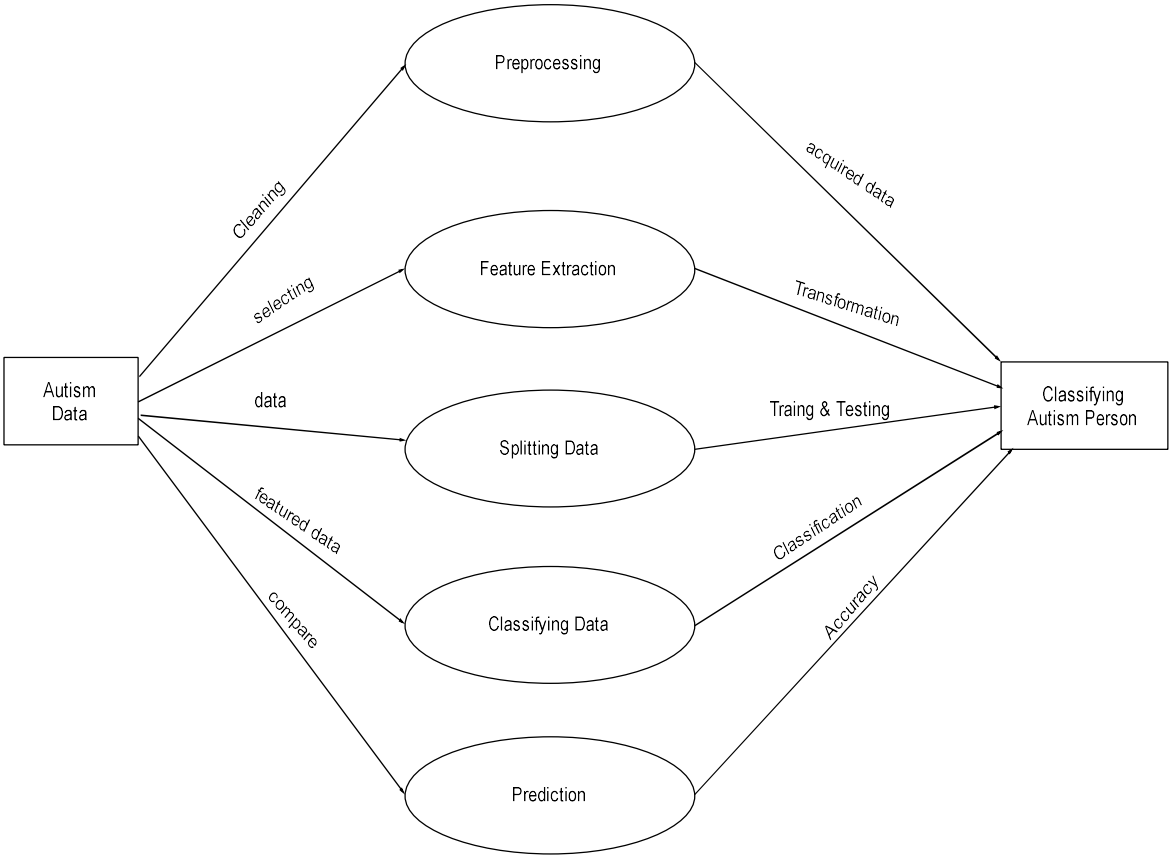


Fig 4.2 :Data Flow Diagram Level 1 for the entire project

Fig 4.2 above is an overall representation of each module and their functionality in this project. This Level one Data Flow Diagram shows how the various modules help in the providing the output which predicts the accurate weather information data with high frequency dataset.

4.3 DATA FLOW DIAGRAM LEVEL 1 FOR DATA PREPROCESSING

Fig 4.3 below clearly explains the data is cleaned before loading to the Neural Network Classifier.Import the Label Encoder class from the sclera library, fit and transform the first column of the data, and then replace the existing text data with the new encoded data.One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. This work fitting and

transforming StandardScaler method on train data.

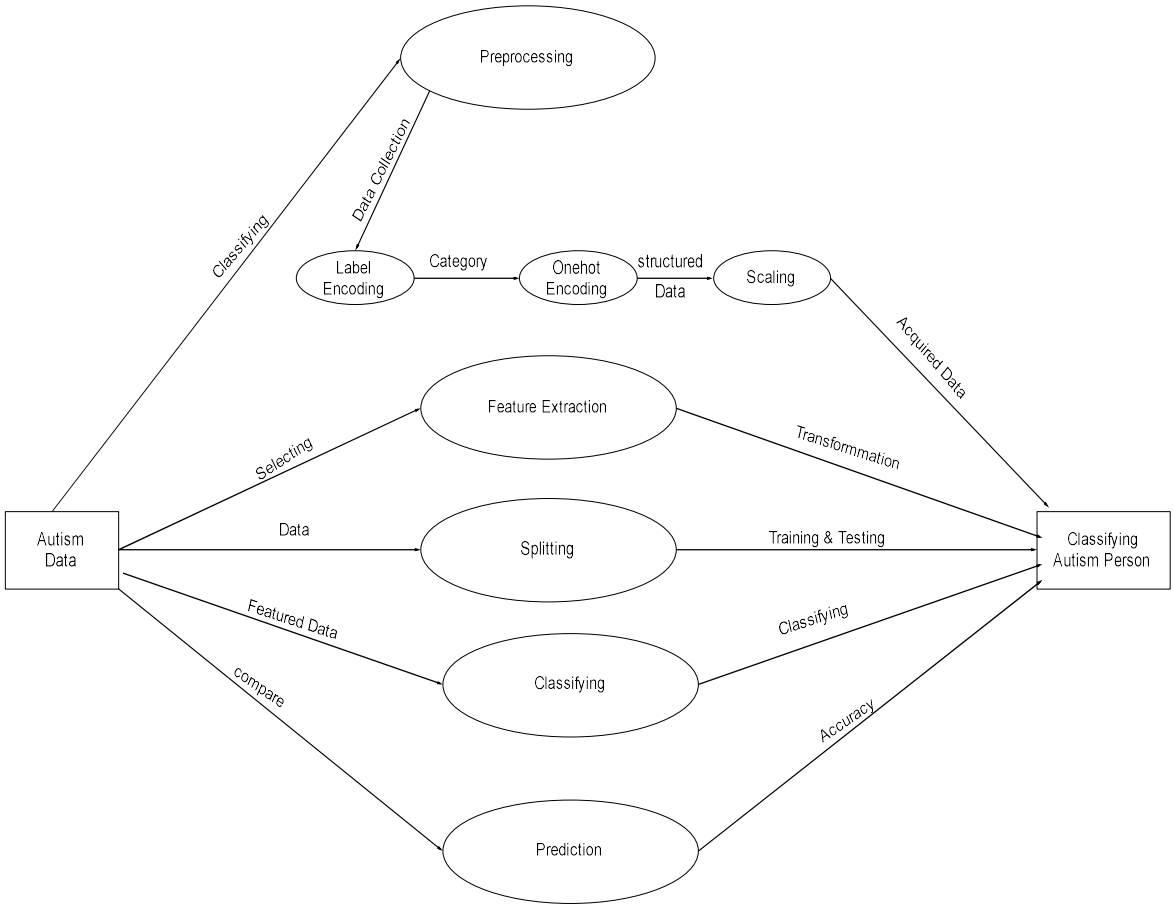


Fig 4.3 :Data Flow Diagram Level 1 for Pre-processing Data

4.4DATA FLOW DIAGRAM LEVEL 2 FOR FEATURE EXTRACTION

Figure 4.4 below explains the Feature extraction by Layering of data.For feature extraction, first select the data and process data for original feature. Then reduce the feature by transform refining data. This can be done by selecting the data by CSV file and process the data for original feature and it was transformed and refining data to reduced feature, finally it predicts the accuracy of weather.

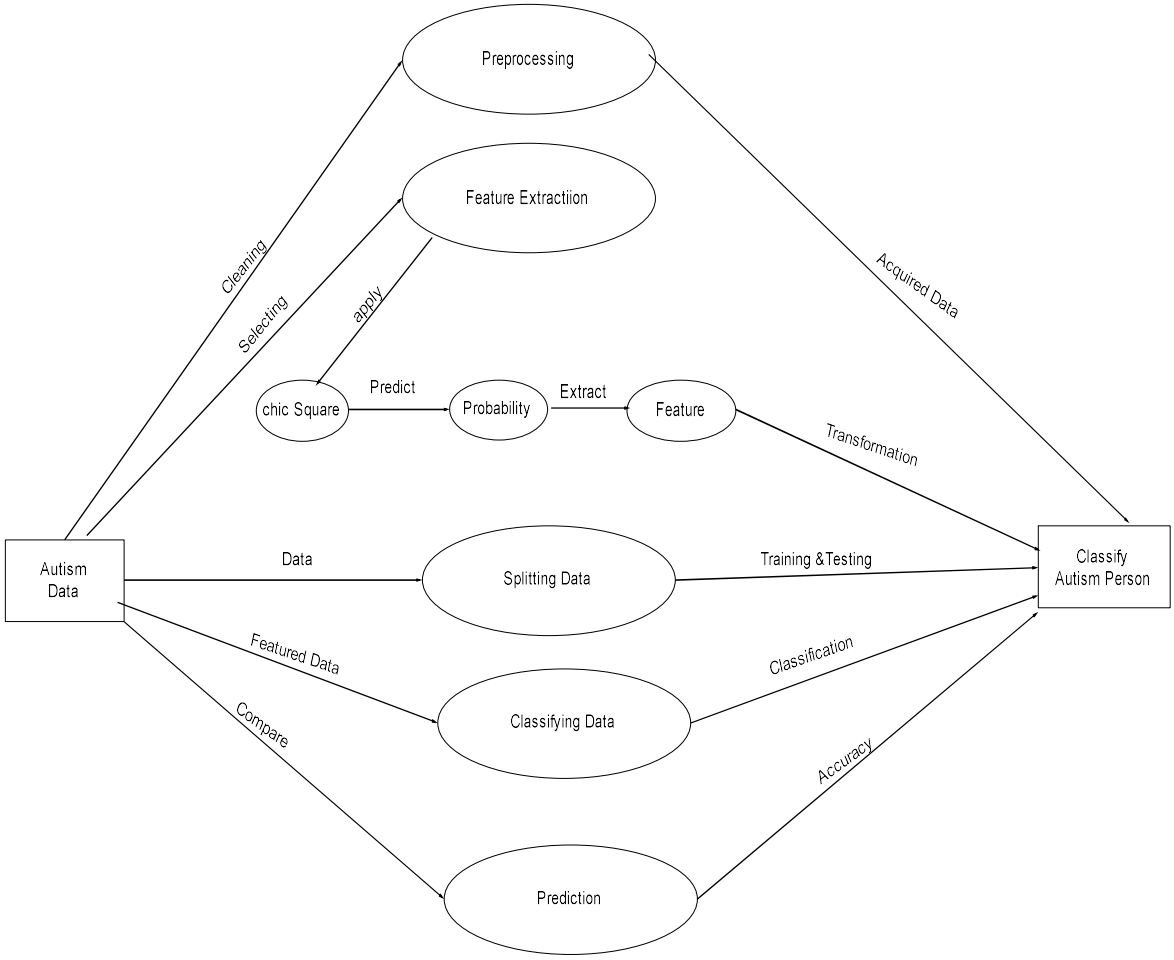


Fig 4.4 : Data Flow Diagram Level 2 for Feature Extraction

4.5DATA FLOW DIAGRAM LEVEL 2 FOR SPLITTING DATA

Figure 4.5 below explains the Splitting data by two ways are Training dataset and Testing dataset. In statistics and machine learning we usually split our data into two subsets: **training data** and **testing data** (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. When this work do that, one of two thing might happen: we overfit our model or we underfit our model. It don't want any of these things to happen, because they affect the predictability of our model—Work might be using a model that has lower accuracy and/or is ungeneralized .

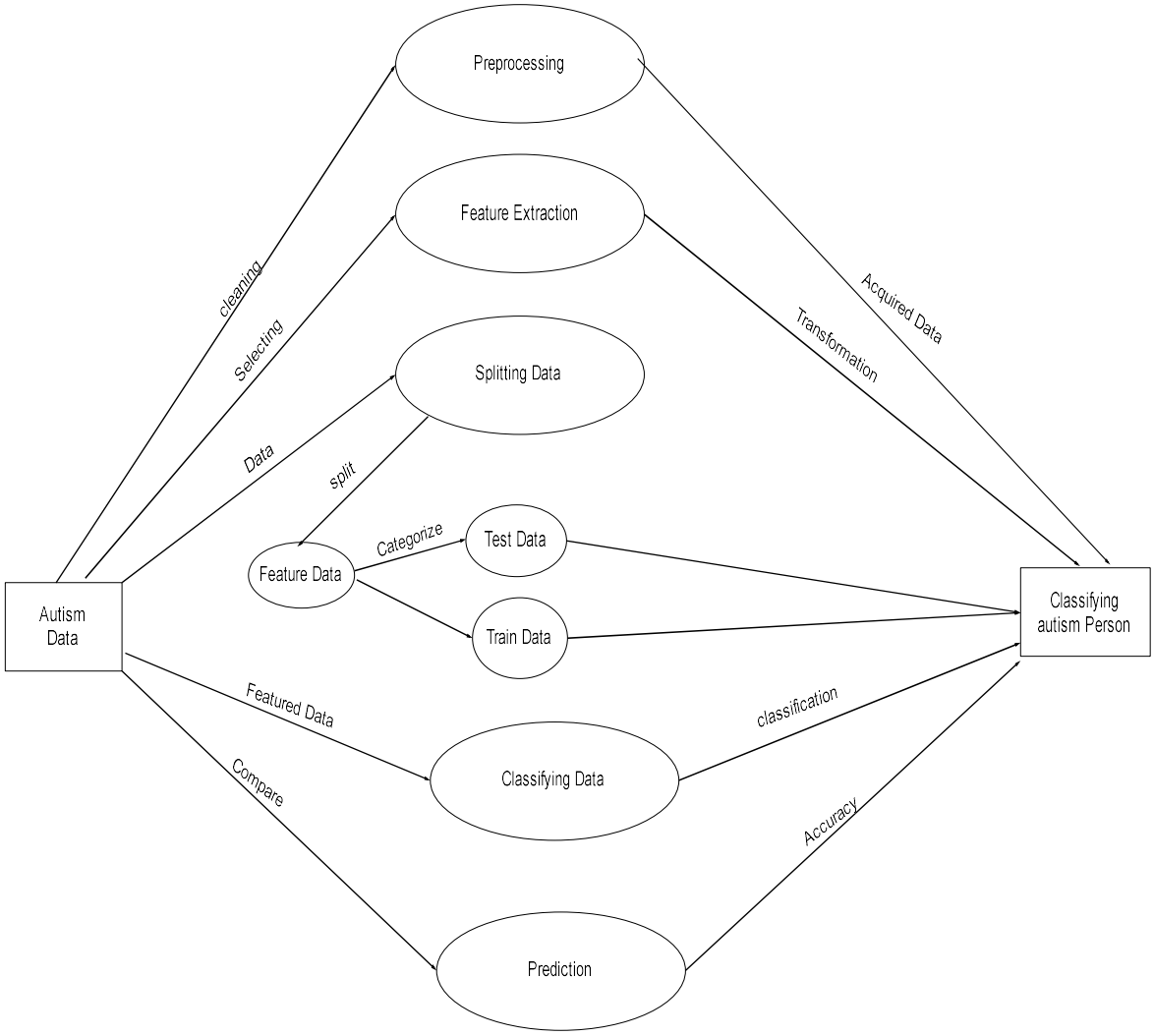


Fig 4.5 : Data Flow Diagram Level 2 for Splitting Data

4.6 DATA FLOW DIAGRAM LEVEL 2 FOR CLASSIFYING DATA

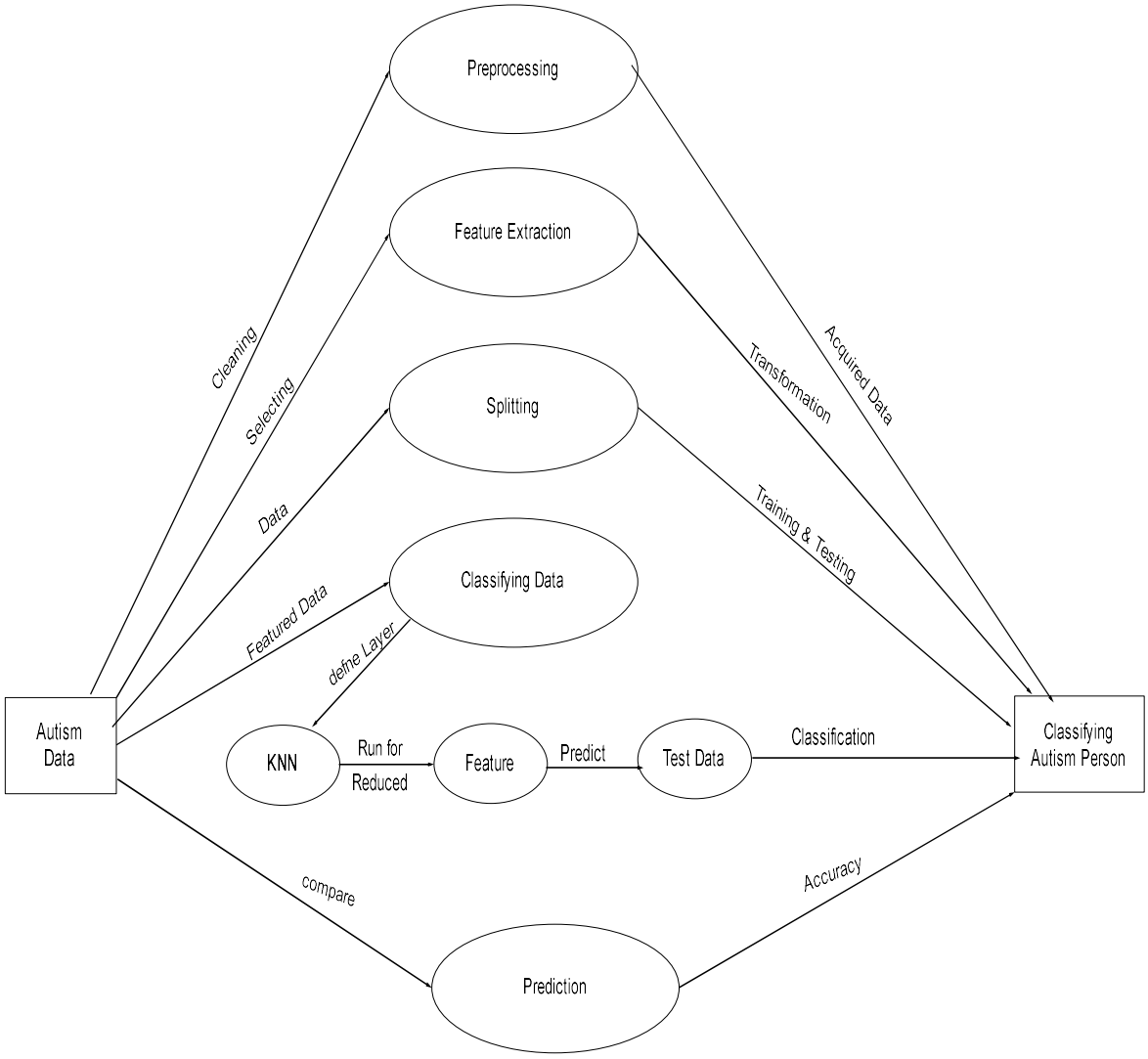


Fig 4.6 : Data Flow Diagram Level 2 for Classifying Data

Figure 4.6 above explains that Classification of data through KNN use randomness by design to ensure they effectively learn the function being approximated for the problem. Randomness is used because this class of machine learning algorithm performs better with it than without.

4.7 DATA FLOW DIAGRAM LEVEL 2 FOR PREDICTION

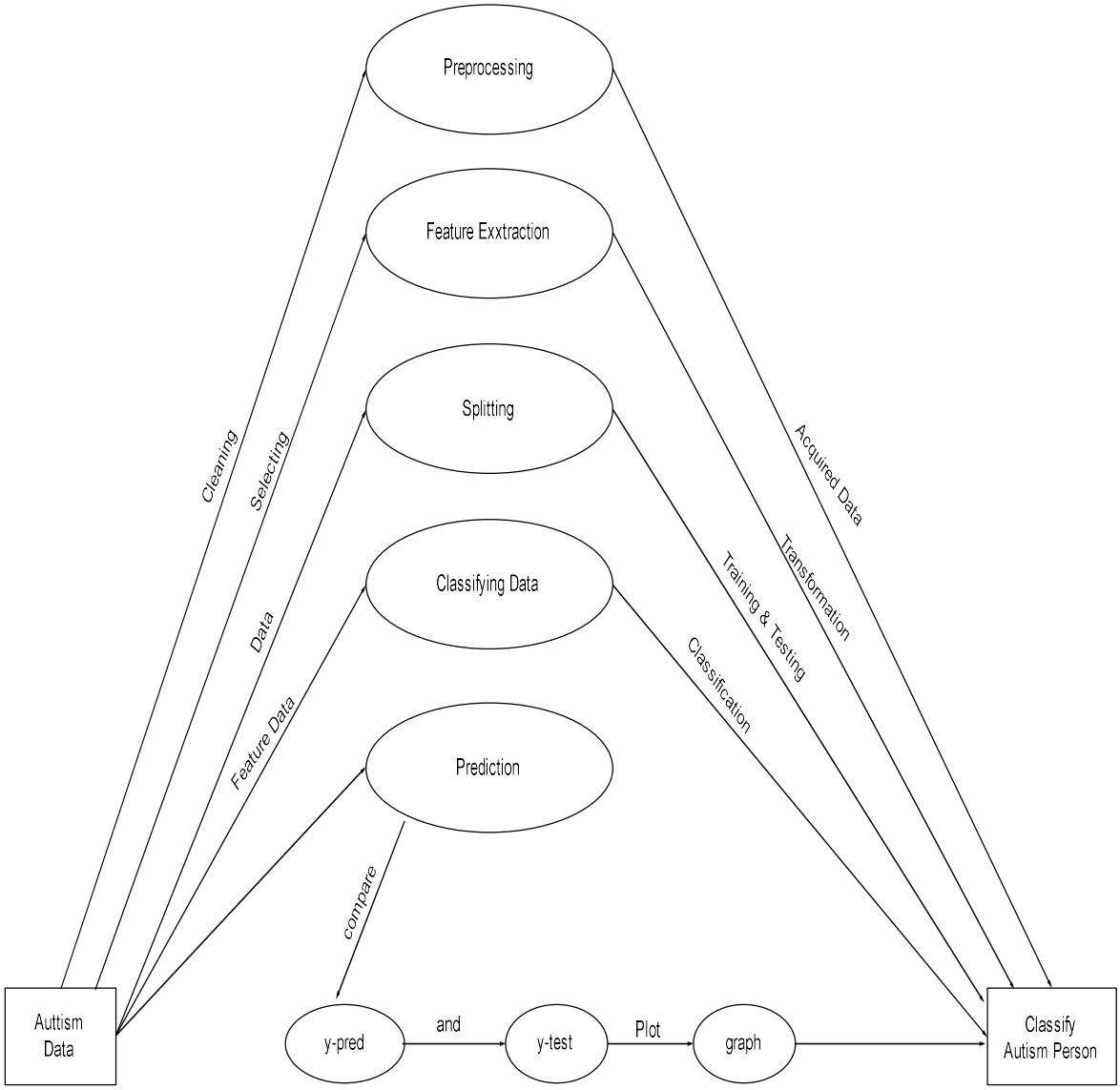


Fig 4.7 : Data Flow Diagram Level 2 for Accuracy Weather Prediction.

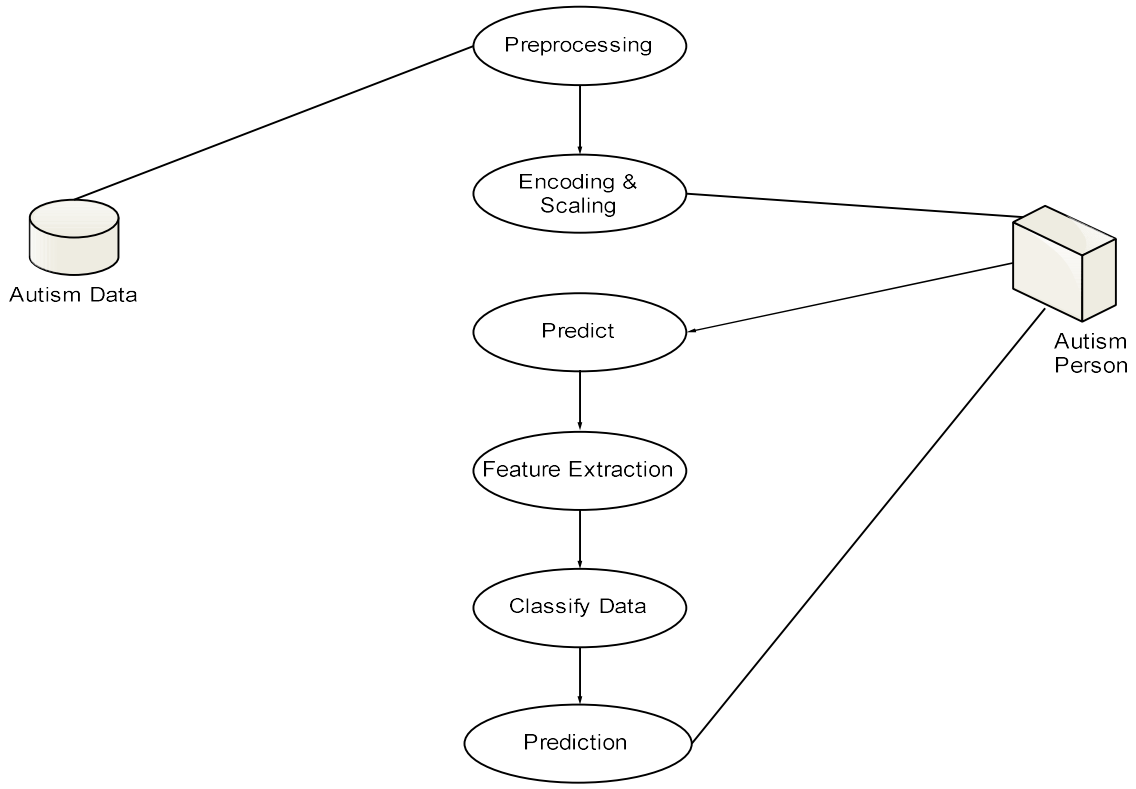
Figure 4.7 above explains how the accuracy weather prediction occurs using the classifier, list with predicted values and plot the autism graph.**Predicting** the test set result. The prediction result will give you probability of a person having autism. This will convert that probability into binary 0 and 1. 1 for Having Autism and 0 for Not . This is the final step where this work evaluating our model performance. It predicts the Confusion matrix to check the

accuracy of model.

4.8 UNIFIED MODELING LANGUAGE

The Unified Modeling Language (UML) is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software intensive system under development.UML combines techniques from data modeling (entity relationship diagrams), business modeling (work flows), object modeling, and component modeling. It can be used with all processes, throughout the software development life cycle, and across different implementation technologies.

4.8.1 USE CASE DIAGRAM FOR LOAD FORECASTING



Above the figure shows that it provides a global view of the actors involved in a system and the actions that the system performs, which in turn provide an observable result that is of value to the actors. Actors are an abstraction of the entity that interacts with the software. Here the system is partitioned simpler functions. The use case diagram of this software is as follows.

4.8.2 SEQUENCE DIAGRAM FOR LOAD FORECASTING

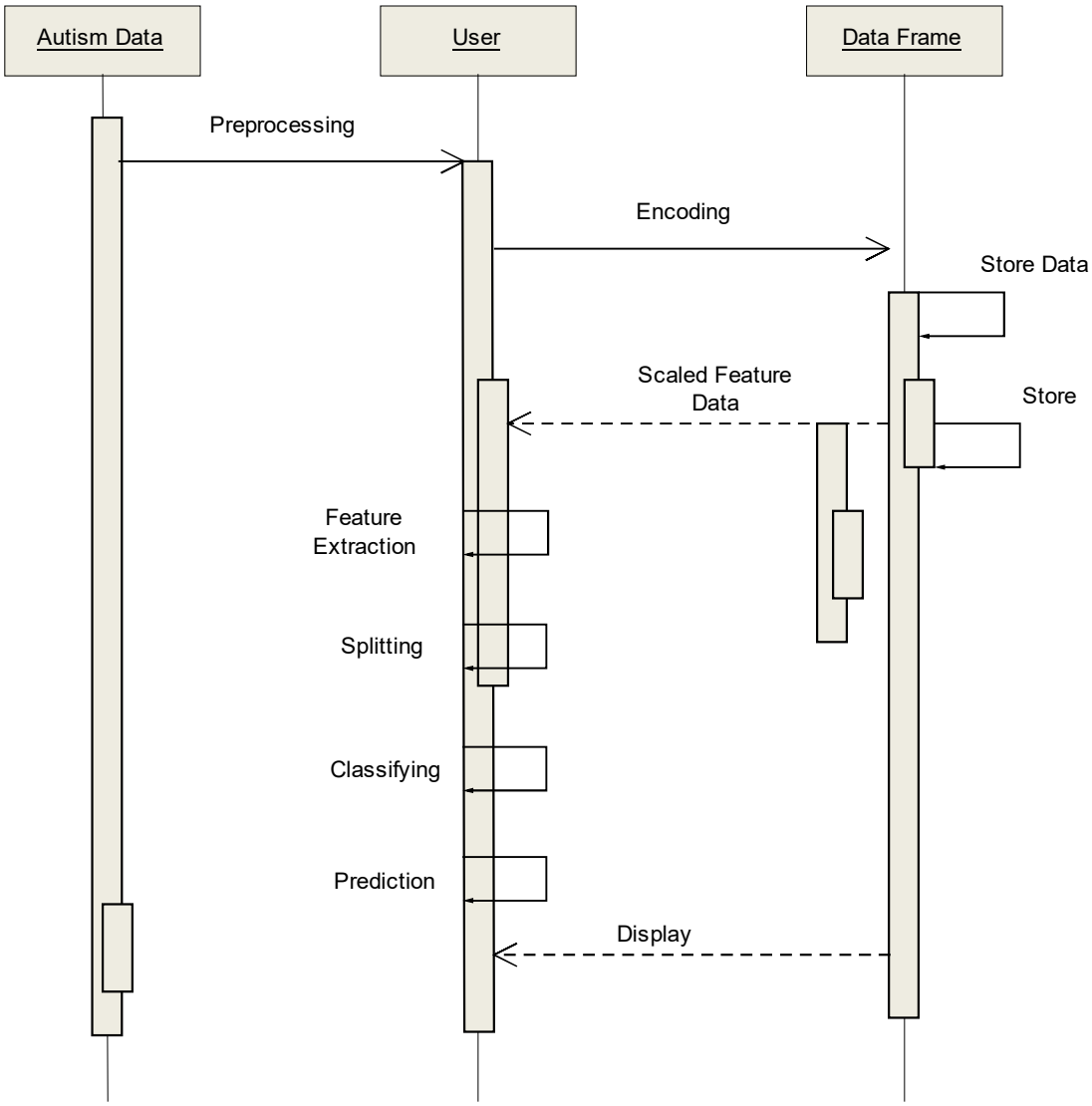


Fig 4.1.2 Sequence Diagram of module 1

In the below sequence diagram, the entities used are data sources, extract, transform, load. The data source extract data from database which normalize the data and metadata management that performs mapping of datafileds. DataMining Algorithm were used to generate RDF triple subnet dataset which transforms the datarate. This data results in each module to visualize the load dataset in effective and distributed way.

CHAPTER – 5

EXPERIMENTAL RESULTS

5.1 Pre-Processing

Data preprocessing used to transform raw data into an understandable format. The data is often incomplete, inconsistent, and contain many errors and noisy data. Data preprocessing is a proven method of resolving such issues. In this diagram shows Pre-Processed data using LabelEncoding and OneHotEncoding. This technique used to transform non numerical data into numerical data.

X - NumPy array

	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0	0	1	1	0	1	16	3	0	8	1	0	4
1	0	1	1	0	0	0	24	4	1	5	1	0	4
2	0	0	1	1	0	1	24	4	1	8	1	0	4
3	1	1	1	1	1	1	12	10	1	0	0	0	4
4	1	1	1	1	1	1	8	9	0	5	0	1	4
5	1	1	1	1	1	1	9	8	1	7	0	0	4
6	1	1	0	0	1	0	21	5	1	6	1	0	4
7	1	0	1	1	1	1	21	6	1	6	1	0	4
8	0	0	1	0	0	1	24	2	1	6	0	0	4
9	1	1	0	1	1	1	10	8	1	10	0	0	0
10	0	1	1	0	1	1	24	6	1	0	1	1	4
11	0	1	1	1	0	1	5	8	1	8	1	0	4
12	0	0	0	0	0	0	13	0	0	8	1	0	4
13	0	0	1	0	1	1	3	7	0	8	1	0	4
14	0	0	0	0	0	0	6	0	1	8	0	0	4
15	1	0	1	1	0	1	0	7	1	7	0	0	4
16	0	0	0	0	0	0	24	0	1	8	0	1	4
17	1	1	1	1	0	1	0	8	0	8	1	0	4
18	1	0	0	0	0	1	17	3	0	8	0	0	4

Format Resize Background color

Save and Close Close

5.2 Feature Selection

Feature Selection is used to select the most useful features in a dataset. Unnecessary features decrease training speed. In this diagram shows important features are selected using **Chi square** algorithm. Here features with height Chi-squared values are selected.

X_kbest_features_train - NumPy array

	0	1	2	3	4	5	6	7	8	9
0	1	1	1	1	1	1	1	1	1	10
1	0	0	0	0	0	0	0	0	0	1
2	0	0	1	1	1	1	1	0	1	7
3	0	0	0	0	1	1	0	0	1	4
4	0	1	1	1	1	1	0	0	1	6
5	1	0	0	0	0	0	0	0	0	2
6	1	0	0	1	0	1	0	0	1	4
7	1	1	0	0	1	1	1	1	1	7
8	0	0	1	1	0	0	0	0	0	2
9	0	0	0	0	1	0	1	1	0	4
10	0	0	0	1	1	0	0	1	0	3
11	1	1	0	0	1	0	0	1	0	4
12	0	1	0	0	0	0	0	0	0	2
13	1	1	1	1	1	1	1	1	1	10
14	1	1	1	1	0	1	1	0	1	8
15	1	1	1	0	1	1	1	1	1	8
16	1	1	1	1	0	1	0	0	0	6
17	1	1	1	1	1	0	1	1	0	8
18	0	0	1	1	1	1	1	1	1	8

Format Resize Background color

Save and Close Close

5.3 Splitting data

5.3.1 Training data

Separating data into training and testing sets most of the data is used for training, and a smaller portion of the data is used for testing. Training set is used to train and fit the model the model. In this diagram shows Training data.

X_train - NumPy array

	4	5	6	7	8	9	10	11	12	13	14	15	16
824	1	1	1	1	1	1	18	10	1	5	1	1	4
825	0	0	1	1	0	1	20	7	1	6	0	0	4
826	0	0	0	0	0	0	24	3	1	6	0	1	4
827	1	1	1	1	0	1	24	8	1	8	0	0	4
828	0	1	1	0	0	1	18	6	1	8	0	0	4
829	1	1	1	1	1	1	17	10	1	5	0	0	4
830	0	1	1	1	0	0	0	3	1	1	0	0	4
831	1	0	1	1	1	1	6	8	0	5	0	0	4
832	0	0	0	0	1	1	24	4	1	5	1	0	4
833	0	0	0	0	0	0	24	0	1	6	0	0	4
834	1	1	1	1	1	1	21	8	0	6	0	0	4
835	0	1	0	1	0	0	13	3	1	5	0	0	4
836	0	0	0	0	0	0	24	0	1	6	0	0	4
837	0	1	0	1	1	1	24	6	1	0	0	0	4
838	1	1	1	1	1	1	18	9	1	1	0	0	4
839	1	1	1	1	1	1	8	9	0	5	0	1	4
840	1	1	1	0	1	1	16	9	1	8	0	0	4
841	0	0	0	1	0	1	24	2	0	6	0	0	1
842	1	1	0	1	1	0	7	7	0	8	0	0	4

Format Resize Background color

Save and Close Close

5.3.2 Testing data

Test data is used to validate model. In this diagram shows Testing data

X_test - NumPy array

	0	1	2	3	4	5	6	7	8	9	10	11	12
192	0	1	0	0	1	1	0	0	1	1	18	5	1
193	1	1	1	1	1	1	1	0	1	0	6	8	1
194	0	1	1	1	1	1	0	1	0	0	22	6	1
195	0	0	1	1	0	1	1	1	1	0	21	6	1
196	0	0	0	0	0	0	0	0	0	1	19	1	0
197	1	1	0	1	0	1	1	0	0	1	15	6	0
198	0	0	1	1	1	1	1	0	0	0	19	5	1
199	0	0	0	0	0	0	1	1	1	0	22	3	0
200	1	1	1	1	1	1	1	0	1	0	24	8	1
201	1	1	1	1	1	1	1	1	1	0	23	9	1
202	1	1	0	0	0	0	0	1	1	0	14	4	0
203	0	1	0	0	0	1	1	0	1	0	24	4	1
204	0	0	0	0	0	0	0	0	0	0	24	0	1
205	0	1	0	0	0	1	0	0	0	0	24	2	1
206	0	0	1	1	1	1	1	1	1	0	3	7	1
207	0	1	1	1	0	1	1	1	1	0	18	7	1
208	1	1	1	1	1	1	1	0	1	1	13	9	0
209	0	1	1	1	1	1	1	0	0	0	10	6	0
210	0	0	0	0	0	0	1	0	0	1	24	2	1

Format Resize Background color

Save and Close Close

5.4 KNN Classification

KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. It selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong. The task is to predict whether the patient is having autism or not and classify accordingly.The K-nearest neighbours (KNN) algorithm is a

type of supervised machine learning algorithms. KNN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real-world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

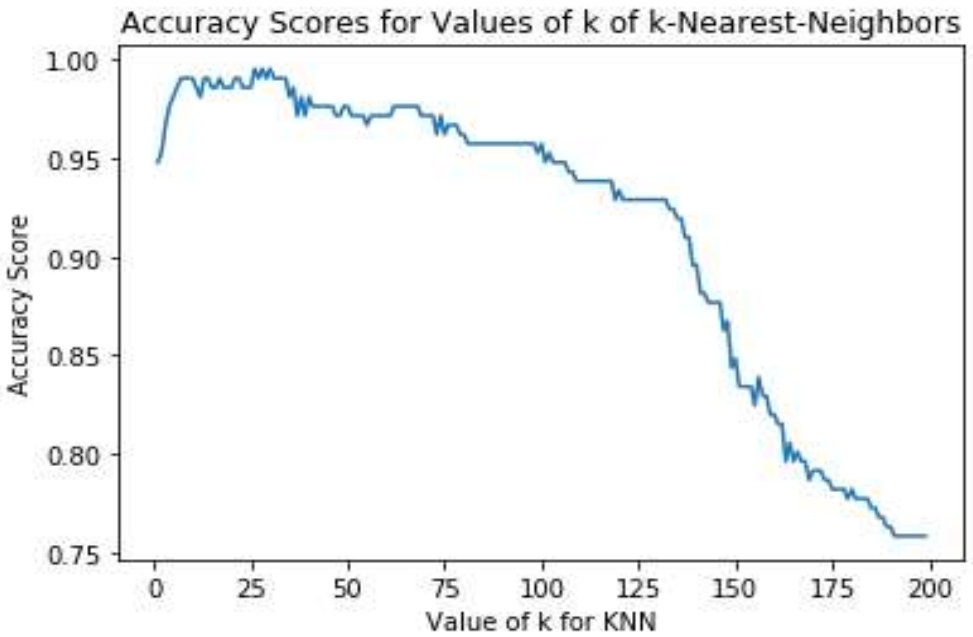
The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. It selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong.

Pros

1. It is extremely easy to implement
2. As said earlier, it is lazy learning algorithm and therefore requires no training prior to making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g SVM, linear regression, etc.
3. Since the algorithm requires no training before making predictions, new data can be added seamlessly.
4. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

We are going to use the autism data set for our KNN. The dataset consists of many attributes: sex, age , how the patient react to communication , speaking reacting test etc. The task is to predict whether the patient is having autism or not and classify accordingly.

This diagram shows accuracy scores for KNN whether Ten features. Accuracy Score is 0.985781990521327



This diagram shows accuracy scores for KNN whether nine features. Accuracy Score is 0.985781990521327.

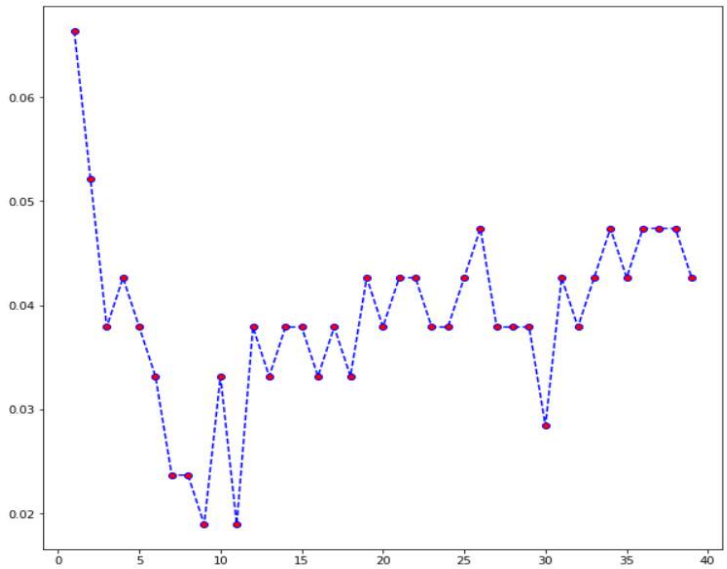


Fig. 5 Error rate in Naïve Bayes

Chi Square Feature Selection :

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. The benefits of performing feature selection before modeling your data are:

- Avoid Overfitting: Less redundant data gives performance boost to the model and results in less opportunity to make decisions based on noise
- Reduces Training Time: Less data means that algorithms train faster

The Chi-Square test of independence is a statistical test to determine if there is a significant relationship between 2 categorical variables. In simple words, the Chi-Square statistic will test whether there is a significant difference in the observed vs the expected frequencies of both variables.

The Chi-Square statistic is calculated as follows:

$$\chi = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$Expected = \frac{RowTotal \times ColumnTotal}{OverallTotal}$$

$$DF = (Rows - 1) \times (Columns - 1)$$

The **Null** hypothesis is that there is NO association between both variables.

The **Alternate** hypothesis says there is evidence to suggest there is an association between the two variables.

In our case, we will use the Chi-Square test to find which variables have an association with the Autism detection . If we reject the null hypothesis, it's an important variable to use in your model.

Rules to use the Chi-Square Test:

1. Variables are Categorical

2. Frequency is at least 5

3. Variables are sampled independently

Chi-Square Test in Python

We will now be implementing this test in an easy to use python class we will call ChiSquare. Our class initialization requires a panda's data frame which will contain the dataset to be used for testing. The Chi-Square test provides important variables such as the P-Value mentioned previously, the Chi-Square statistic and the degrees of freedom.

The first step is to import the Chi Square libraries

```
from sklearn.feature_selection import SelectKBest

from sklearn.feature_selection import chi2
```

This Chi Feature works over the Probability of finding , it requires all its input in the form of integer , hence converting the testing and training input variables to integer format .

```
X = X.astype(int)
```

```
X_train= X_train.astype(int)

X_test= X_test.astype(int)
```

We first compute the observed count for each class. This is done by building a contingency table from an input XX (feature values) and yy (class labels). Each entry ii, jj corresponds to some feature ii and some class jj, and holds the sum of the ithith feature's values across all samples belonging to the class jj.
Scikit-learn provides a **`SelectKBest`** class that can be used with a suite of different statistical tests. It will rank the features with the statistical test that we've specified and select the top performing ones.
Then the best outcome among all the class is taken and stored in a variable chi2_features .

```
chi2_features = SelectKBest(chi2)
```

For the Chi-Square feature selection we should expect that out of the total selected features, a small part of them are still independent from the class. It rarely matters when a few additional terms are included the in the final feature set. The trained and test data are feature selected as below

```
X_kbest_features_train = chi2_features.fit_transform(X_train, y_train)

X_kbest_features_test = chi2_features.fit_transform(X_test, y_test)
```

CHAPTER - 6

CONCLUTION

In this project gives better Accuracy score using KNN.Data preprocessing is important phase while making prediction from dataset. Data preprocessing includes LabelEncoding, OneHotEncoding. This techniqueis used to change categorical data into numerical data. This project using chi-square method for feature selection. Chi-square checks the dependence of each single attribute with class label. If the dependence is high chi-square keeps the attribute otherwise discard that feature. KNN has been used to improve the accuracy of classifiers. This Project used different set of features selected by chi-square method. Combination of ten (nine, eight,etc..,) features produced most accurate sameresults. Irrelevant features have been eliminated after feature selection using chi-square method. After the selected features are classified using KNN algorithm. It gives better Accuracy.

PERFORMANCE ANALYSIS

6.1 PERFORMANCE ANALYSIS FOR K Nearest Neighbour

Here various predicting models described for finding Autism. This work has chosen KNN –K Nearest Neighbour.KNN use randomness by design to ensure they effectively learn the function being approximated for the problem. Randomness is used because this class of machine learning algorithm performs better with it than without.

6.2 LDM

LDM is one of the most prominent supervised machine learning approaches. This method helps in predicting prediction of a categorical dependent variable from a group of independent variables.". Linear Regression and Logistic Regression are very similar in terms of how they are employed."

6.3 SVM

For dealing with unlabeled data, supervised learning is impossible, so an unsupervised technique is required, in which the data is naturally clustered into groups, and then fresh data is mapped to these created groups, which is where SVM comes in. It categorizes unlabeled data using support vector statistics generated in the support vector machines technique. The efficiency of SVM is determined on the kernel chosen, its parameters, and the soft margin parameter. A Gaussian kernel, which has only one parameter, is a popular choice.Potential drawbacks of the SVM include the following aspects:

- Input data must be fully labelled.
- Probabilities of belonging to a class that aren't calibrated SVM is based on Vapnik's theory, which eliminates the need to estimate probabilities on finite data.
- Only two-class tasks are directly applicable to the SVM. As a result, techniques must be used to reduce the multi-class task to a series of binary problems.
- It's tough to interpret the parameters of a solved model.

6.4 Naïve Bayes

Naive Bayes is a prominent method for building classifiers. The "event model" of the naive Bayes classifier is made up of assumptions about feature distributions. Multinomial and Bernoulli distributions are common for discrete features like those found in document categorization (including spam filtering). These assumptions result in two unique models that are frequently misunderstood.

Samples (feature vectors) in a multinomial event model represent the frequency with which certain events have been generated by a multinomial.

(d) Training and Testing

After selecting different models, the dataset is trained based on different Construction Model. Then the data is tested based over all the Model and one with better accuracy is considered to be used further.

Appendix

A. Screenshots

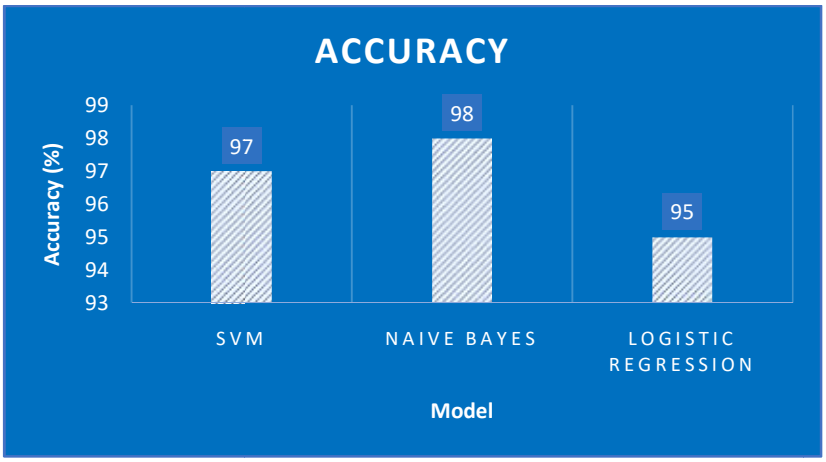


Fig. S.1 Accuracy

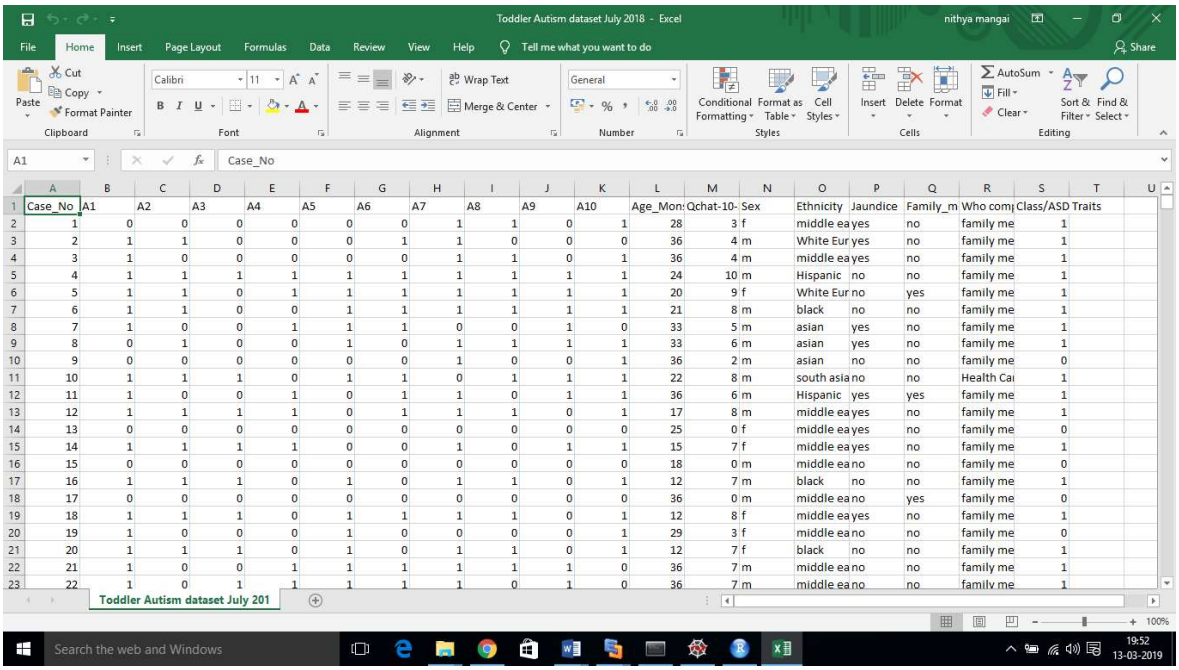
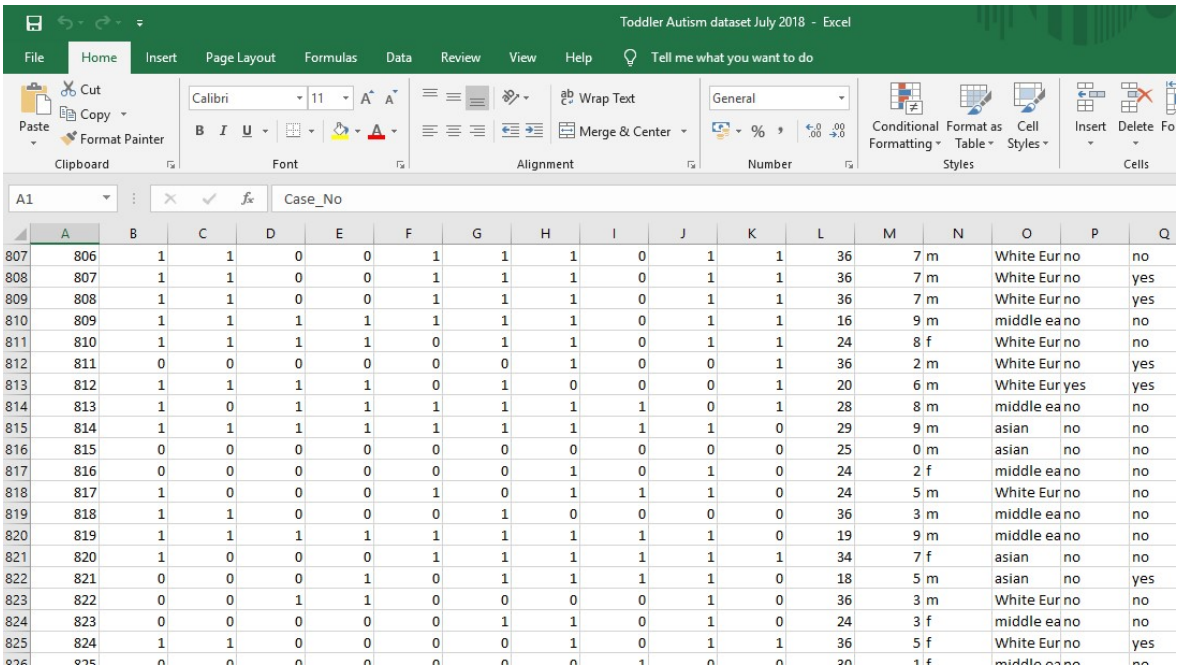


Fig S.2 SCREENSHOT FOR MODULE 1



Toddler Autism dataset July 2018 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
807	806	1	1	0	0	1	1	1	0	1	1	36	7	m	White Eur	no	no
808	807	1	1	0	0	1	1	1	0	1	1	36	7	m	White Eur	no	yes
809	808	1	1	0	0	1	1	1	0	1	1	36	7	m	White Eur	no	yes
810	809	1	1	1	1	1	1	1	0	1	1	16	9	m	middle ea	no	no
811	810	1	1	1	1	0	1	1	0	1	1	24	8	f	White Eur	no	no
812	811	0	0	0	0	0	0	1	0	0	1	36	2	m	White Eur	no	yes
813	812	1	1	1	1	0	1	0	0	0	1	20	6	m	White Eur	yes	yes
814	813	1	0	1	1	1	1	1	1	0	1	28	8	m	middle ea	no	no
815	814	1	1	1	1	1	1	1	1	1	0	29	9	m	asian	no	no
816	815	0	0	0	0	0	0	0	0	0	0	25	0	m	asian	no	no
817	816	0	0	0	0	0	0	1	0	1	0	24	2	f	middle ea	no	no
818	817	1	0	0	0	1	0	1	1	1	0	24	5	m	White Eur	no	no
819	818	1	1	0	0	0	1	0	0	0	0	36	3	m	middle ea	no	no
820	819	1	1	1	1	1	1	1	1	1	0	19	9	m	middle ea	no	no
821	820	1	0	0	0	1	1	1	1	1	1	34	7	f	asian	no	no
822	821	0	0	0	1	0	1	1	1	1	0	18	5	m	asian	no	yes
823	822	0	0	1	1	0	0	0	0	1	0	36	3	m	White Eur	no	no
824	823	0	0	0	0	0	1	1	0	1	0	24	3	f	middle ea	no	no
825	824	1	1	0	0	0	0	1	0	1	1	36	5	f	White Eur	no	yes

Fig S.3 SCREENSHOT FOR MODULE 1

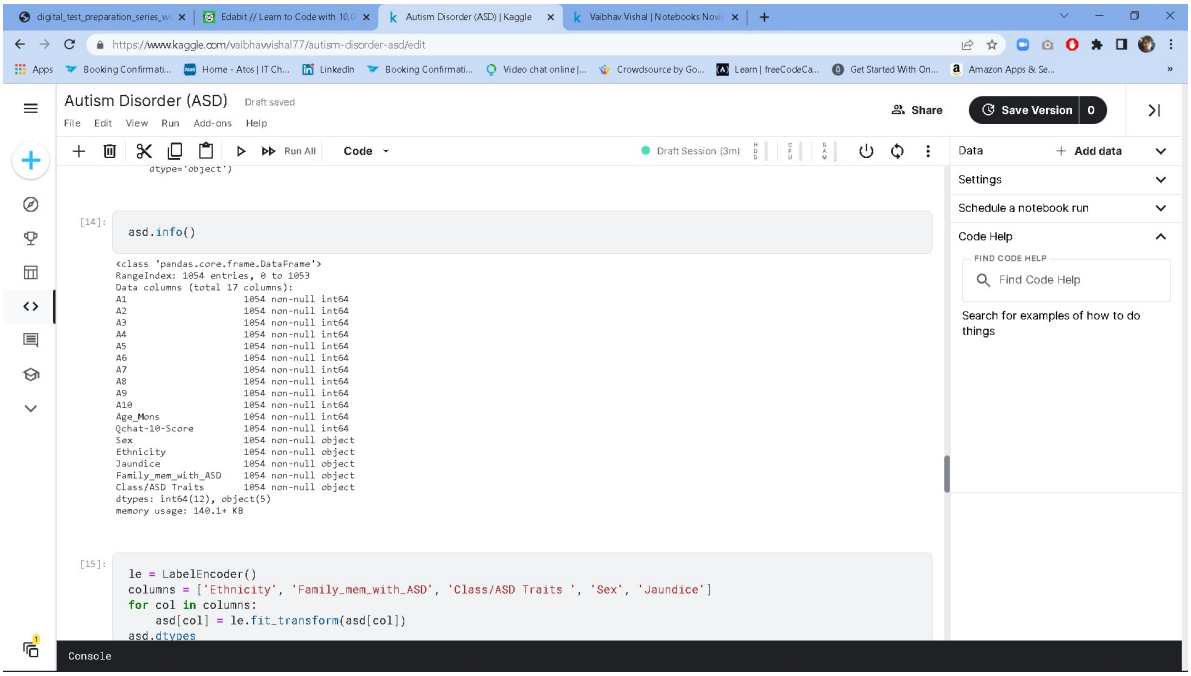
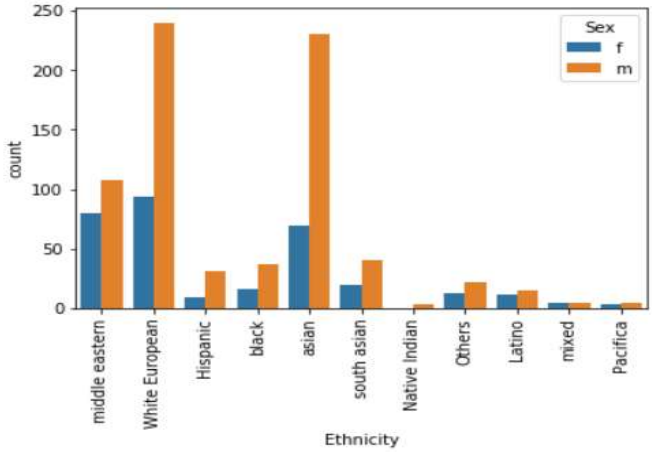


Fig. S.4 Data Analysis

```
[11]: sns.countplot(x='Ethnicity',hue='Sex',data=asd)
      plt.xticks(rotation=90)
```

```
[11]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10]),
       <a list of 11 Text xticklabel objects>)
```



+ Code + Markdown

Fig. S.5 Ethnicity Analysis

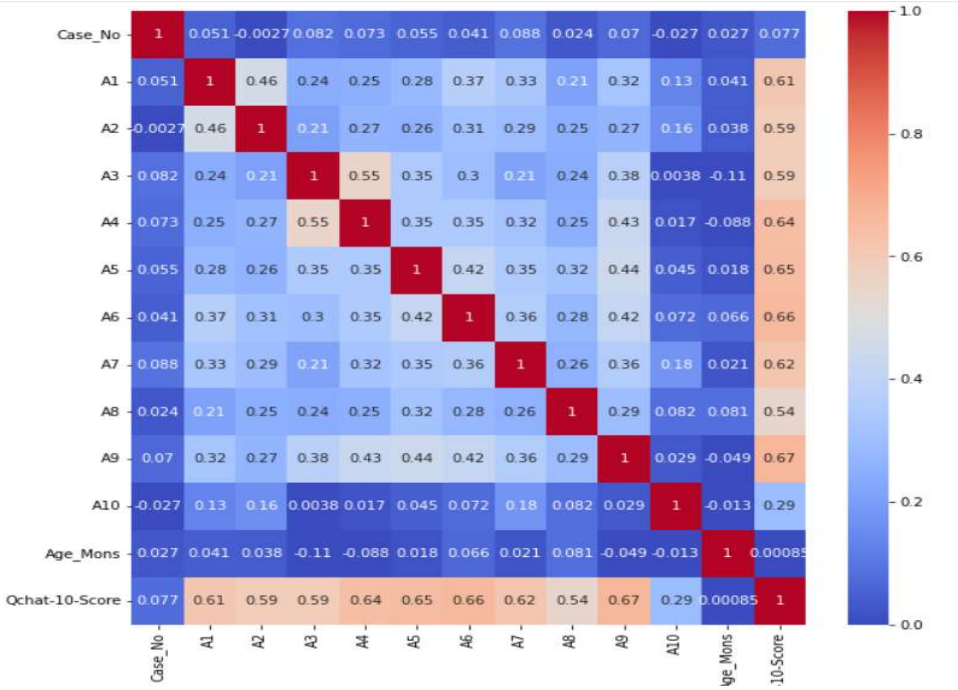


Fig. S.6 Heat Map

In [86]:

```
sns.scatterplot(x='Age_Mons',y='Qchat-10-Score',data=df)
```

Out[86]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d031b55d0>
```

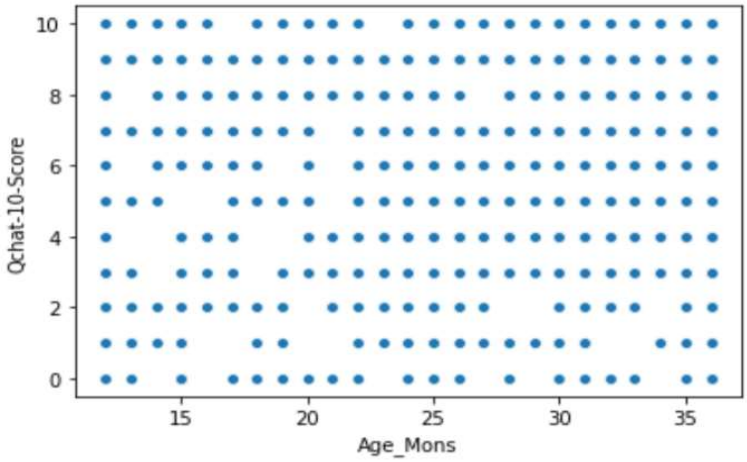


fig. S.7 Age Analysis

```
models = []
models.append(('Logistic Regression:', LogisticRegression()))
models.append(('Naive Bayes        :', GaussianNB()))
models.append(('SVM                :', SVC()))

for name, model in models:
    model.fit(x_train, y_train)
    pred = model.predict(x_test).astype(int)
    print(name, accuracy_score(y_test, pred))
```

```
Logistic Regression: 0.957345971563981
Naive Bayes        : 0.981042654028436
SVM                : 0.9715639810426541
```

+ Code + Markdown

Fig. S.8 Result

B. Source Code:-

```
# Importing libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Supress warnings
import warnings
warnings.filterwarnings("ignore")

# Classification
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier



# Regression
from sklearn.linear_model import LinearRegression,Ridge,Lasso,RidgeCV, ElasticNet,
LogisticRegression
from sklearn.ensemble import
RandomForestRegressor,BaggingRegressor,GradientBoostingRegressor,AdaBoostRegressor
from sklearn.preprocessing import Imputer , Normalizer , scale
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import RFECV
from sklearn.model_selection import GridSearchCV , KFold , cross_val_score,
ShuffleSplit, cross_validate

# Preprocessing :
from sklearn.preprocessing import MinMaxScaler , StandardScaler, Imputer, LabelEncoder

# Metrics :
# Regression
from sklearn.metrics import mean_squared_log_error,mean_squared_error,
r2_score,mean_absolute_error
# Classification
from sklearn.metrics import accuracy_score,precision_score,recall_score,f1_score,
classification_report

print("Setup complete...")
asd = pd.read_csv("../input/Toddler Autism dataset July 2018.csv")
print("Dataset loaded...")
le = LabelEncoder()
columns = ['Ethnicity', 'Family_mem_with_ASD', 'Class/ASD Traits ', 'Sex', 'Jaundice']
for col in columns:
    asd[col] = le.fit_transform(asd[col])
asd.dtypes
X = asd.drop(['Class/ASD Traits '], axis = 1)
Y = asd['Class/ASD Traits ']
```

```
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.20,
random_state = 7)
print('Processed...')
#logistic method
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
preds = logreg.predict(x_test)
logreg.score(x_train, y_train)
#classification report
#confusion matrix
from sklearn.metrics import classification_report, confusion_matrix
classification_report(y_test, preds)
#to get the optimum n_neighbors
import matplotlib.pyplot as plt
error_rate = []
for i in range(1,40):
    knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(x_train, y_train)
    pred = knn.predict(x_test)
    error_rate.append(np.mean(y_test != pred))
plt.figure(figsize=(10,10))
plt.plot(range(1,40), error_rate,
color='blue',linestyle='dashed',marker='o',markerfacecolor='red')
models = []
models.append(('Logistic Regression:', LogisticRegression()))
models.append(('Naive Bayes        :', GaussianNB()))
models.append(('SVM                :', SVC()))


for name, model in models:
    model.fit(x_train, y_train)
    pred = model.predict(x_test).astype(int)
    print(name, accuracy_score(y_test, pred))
```

REFERENCES

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.

7. Umair Ayub, Syed Atif Moqurrab.” Predicting crop diseases using data mining approaches:Classification.” 1st International Conference on Power, Energy and Smart Grid (ICPESG).