PREDICTION AND MULTI LABEL CLASSIFICATION OF LEARNING DISABILITIES IN CHILDREN

Project report submitted to the Amrita Vishwa Vidyapeetham University in partial fulfillment of the requirement for the Degree of

BACHELOR of TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

R. OJASWITHA MOUNICA - AM.EN.U4CSE15150
SARATH CHANDRIKA K - AM.EN.U4CSE15154
KROVVIDI SINDHU - AM.EN.U4CSE15233
VATTI SOUMYA - AM.EN.U4CSE15262

AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
(Estd. U/S 3 of the UGC Act 1956)
AMRITAPURI CAMPUS
KOLLAM - 690525

MAY 2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

AMRITA VISHWA VIDYAPEETHAM
(Estd. U/S 3 of the UGC Act 1956)

Amritapuri Campus

Kollam - 690525

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **"PREDICTION AND MULTI LABEL CLASSIFICATION OF LEARNING DISABILITIES IN CHILDREN "** submitted by

R. OJASWITHA MOUNICA (AM.EN.U4CSE15150),
SARATH CHANDRIKA K (AM.EN.U4CSE15154),
KROVVIDI SINDHU (AM.EN.U4CSE15233),
VATTI SOUMYA (AM.EN.U4CSE15262)

in partial fulfillment of the requirements for the award of Degree of Bachelor of Technology in Computer Science and Engineering from Amrita Vishwa Vidyapeetham, is a bonafide record of the work carried out by them under my guidance and supervision at Amrita School of Engineering, Amritapuri during Semester 8 of the academic year 2018-2019.

Gayathri RG
Project Guide

Sandhya Harikumar
Project Coordinator

Dr. Jayaraj Poroor
Chairperson,
Dept. of Computer Science & Engineering

External Examiner

Place : Amritapuri
Date : 31-05-2019

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

AMRITA VISHWA VIDYAPEETHAM
(Estd. U/S 3 of the UGC Act 1956)

Amritapuri Campus

Kollam - 690525

DECLARATION

We,

R. OJASWITHA MOUNICA (AM.EN.U4CSE15150),
SARATH CHANDRIKA K (AM.EN.U4CSE15154),
KROVVIDI SINDHU (AM.EN.U4CSE15233),
VATTI SOUMYA (AM.EN.U4CSE15262)

hereby declare that project entitled **"PREDICTION AND MULTI LABEL CLASSIFICATION OF LEARNING DISABILITIES IN CHILDREN"** is a record of original work done by us under the guidance of **GAYATHRI RG**, Dept. of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, that this work has not formed the basis for any degree / diploma / associationship / fellowship or similar awards to any candidate in any university to the best of our knowledge.

Place : Amritapuri

Date : 31-05-2019

Signature of the student

Signature of the Project Guide

Acknowledgements

Firstly, I would like to offer my pranams at the lotus feet of our beloved Amma who showered her blessings all throughout the life at Amrita.

My sincere gratitude to Dr. Jyothi S N, Principal, Amrita school of Engineering, Dr. Jayaraj Poroor, Chairperson, Department of Computer science and Engineering for providing necessary facilities.

I would like to thank my guide Ms. Gayathri RG for providing continual encouragement through a relaxed approach, support and proper guidance with regard to development of the project and holding us to a higher standard.

My heartfelt thanks to my project coordinator, Sandhya Harikumar, Assistant professor, Department of Computer science and Engineering for her guidance and coordination.

I would like to thank my friends for giving us valuable information and also our family members for the support they have shown.

I thank the Almighty god for his blessings without which I could not have finished this project.

Abstract

The aim of the study is to exhibit the importance of ensemble techniques in the prediction of learning disabilities in school-age children and to classify the type of learning disability using multi label classification techniques. Learning disability (LD) is a neurological disorder in which students have persistent difficulty in learning. About 1 out of 10 school age children in India suffer from learning disorders. Children with problems in LD are of a higher concern to the parents and teachers. Therefore, diagnosis of LD at the early stages play a major role. Along with this finding out the type of LD in children is also useful.

This study helps to identify the important parameters of LD thereby estimating the relevance of each symptom of LD and effectively predict the possibility of LD using ensemble technique and type of LD using multi label classification techniques.

In this study, stacking ensemble technique is implemented to predict whether a child is suffering from learning disabilities. Initially, the results of individual machine learning models that is logistic regression, k-nearest neighbour (knn) and support vector machine (svm) are calculated. Later, the above algorithms are used in stacking ensemble model along with decision tree and final results are obtained. Finally, the results of individual machine learning algorithms and stacking ensemble algorithm are compared. Further multi label classification is implemented to predict the type of LD. In this the problem transformations methods binary relevance, classifier chain, label powerset are used to construct stacking ensemble model. All the individual transformation models, adaption model, rakel algorithms and stacking ensemble of transformation models are implemented and comparative results are produced.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Learning disability (LD) is a neurological disorder in which students have persistent difficulty in learning. In simple terms, a learning disability results from a difference in the way a person's brain is weird. They are neurologically-based processing problems. These processing problems can interfere with learning basic skills such as reading, writing and/or math. They can also have difficulty with higher level skills such as organizing information, time planning, abstract reasoning, long or short term memory and attention. Children with learning disabilities are as smart or smarter than their peers. It cannot be cured or fixed but with right support and intervention children with learning disabilities can succeed in school and go on to be successful. About 1 out of 10 school-age children in India suffer from learning disorders. The problems of children with these learning disabilities cause a greater concern to parents and teachers. Therefore early diagnosis of LD in children is important. Our study deals with predicting whether a child is suffering from LD along with multi label classification into Dyslexia, Dysgraphia, Dyscaulcia and ADHD.

1.1 Motivation

In India at least 10 Percent of children have learning disability. It is a tremendous challenge to identify and diagnose and assist children with learning disability. As the concept is still new, in many developing countries including India, the research conducted in learning disability has been primarily done over the last two decades are yet in the infancy stage. Since no national census of the learning disabled has been taken in India, it is difficult to collect their actual number.

We do not have a clear idea about incidence and prevalence of learning

disability in India. The problems of LD affected children have been a cause of concern to parents and school authorities for some time. With the right help at right time, right assessment and remediation, children with LD can learn successfully and become winners in the society later. These facts suggest that early diagnosis of learning disability in children is very important. Research works done in this area using computer based methods is found very little compared to the magnitude of learning disability affected children. If the LD determination facility is attached with schools and the check ups are arranged as a routine process, LD can be identified at an early stage. Under these circumstances, it is felt to design a tool based on machine learning techniques for prediction and also classification of learning disability in School-aged children.

1.2 Concept of Learning Disability

Learning disability is usually caused by an unknown factor or factors. The unknown factor is the disorder that affects the brain's ability to receive and process information. This disorder can make it problematic for a child to learn as quickly or in the same way as some child who isn't affected by a learning disability

1.2.1 Causes of Learning Disability

It is still uncertain about the causes of learning disabilities. Often, learning problems can run in families (genetic), but environmental factors can also play a role. Mostly, learning disabilities occur because there is an enormous range of variation that occurs normally in peoples cognitive strengths and weaknesses. If we think about our physical development, nearly everyone has two eyes, a nose and a mouth, yet each of our faces has its own distinctive features. The same is true of brain development. Those children who encounter difficulty in meeting age and grade level expectations, then the problem can be identified as a learning disability.

Few factors that might influence the development of learning disorders include:

- Family history and genetics
  A family history of learning disorders increases the risk of a child developing a disorder.

- Prenatal and neonatal risks
  Poor growth in the uterus (severe intrauterine growth restriction), ex-

posure to alcohol or drugs before being born, premature birth, and very low birth weight have been linked with learning disorders.

- Psychological trauma
  Psychological trauma or abuse in early childhood may affect brain development and increase the risk of learning disorders.

- Physical trauma
  Head injuries or nervous system infections might play a role in the development of learning disorders.

- Environmental exposure
  Exposure to high levels of toxins, such as lead, has been linked to an increased risk of learning disorders.

1.2.2 Types of Learning Disability

Learning disabilities are also termed as learning differences, based on the fact that certain individuals learn differently - they are not unable to learn, but respond best to ways of learning that are different from traditional teaching methods. They tend to be diagnosed only when a child reach school age. This is because school focuses on the many things that may be difficult for the child reading, writing and math, listening, speaking and reasoning. Types of learning disabilities are often grouped by school area skill set or cognitive weakness.

The common types of learning disabilities are:

Dyslexia

Dyslexia is one of the most common learning disability. The word dyslexia implies the meaning difficulty with words. Some of the most common signs of reading disability include: difficulty in associating or recognizing sounds that go with letters and separating the sounds within words, difficulty in sounding out words, trouble in rhyming, poor spelling and problems in understanding and using both words and grammar.

Dyscalculia

A specific learning disability that affects a persons ability to understand numbers and learn math facts. Some signs include slow in counting and math problem solving skills, computing problems, problems with time concepts and poor sense of direction.

Dysgraphia

The word dysgraphia implies difficulty with writing. Its a persons ability to express their thoughts in writing. Some of the common signs include: awkward or tight grip on pencil, illegible handwriting, speaking the words out loud while writing, difficulty with grammar and difficulty in organizing thoughts while writing.

Attention deficit hyperactivity disorder (ADHD)

Children with ADHD show signs of inattention, hyperactivity and/or impulsitivity in specific ways. They may have trouble sitting still, following directions and completing tasks at home or school.

1.2.3 Assessment of learning disability

LD can be a lifetime condition. There may be several apparent overlapping learning disabilities in some children while others may have a single, isolated learning problem that has little impact on their lives. LD is diagnosed by a qualified child psychologist in association with a pediatrician. The process of diagnosing a learning disability can be confusing. It shall be started with the child's school. It involves testing, history taking and observation by a trained specialist. A series of tests may be required to be done to identify the affected areas. Special education brings some solution to the problems the children affected with LD. Finding a reputable referral is important. Since the educational needs of such children are different they will be given special academic sessions in an integrated set up.

Assessment of LD is the systematic process of collecting information about a child, his past and current levels of performance his strength and weakness, in order to help make better education decisions [10]. Assessment needs to be relevant to the teaching goals and interventions that the child will receive. Assessment is directly linked with how one will go about helping the child. It is linked with intervention methods. The information collected through the assessment must be relevant and of practical help in the class room. Assessment helps parents to better understand their childs problems and adjust their expectations on the basis of the assessment data. It is pertinent to note that, in India the history of LD assessment is still in its infancy. Many types of assessment tests are available. Child's age and the type of problem determine tests that child needs. Before any formal testing, a conference is usually arranged between the child's parents and representatives from the special education department. A factor that prevents accurate

diagnosis of twice exceptional is the prevalent practice of comparing gifted children with the norms for average children. In psychology, as well as in other therapeutic fields, such as audiology, speech pathology, occupational therapy and optometry, the diagnostic question that is usually asked is how this childs performance compares with the norm. If the child scores within the normal range, no disabilities are detected.

The purpose of any evaluation or assessment for LDs is to determine child's strengths and weaknesses and to understand how he or she best learns and where they have difficulty. A major factor that makes it difficult to assess a learning disabled child is the confusing nature of the disability itself. The absence of testing instrument relevant to Indian students is another major drawback. Most tests are designed for native English speakers and have items which lie outside the cultural experience of the average Indian student. Assessment can be expensive too. In order to overcome this situation, this study is a basic step in predicting whether a child has learning disability or not using machine learning algorithms based on the data collected from local hospitals.

1.3 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Machine Learning algorithms are often categorized as Supervised or unsupervised.

Supervised Learning

Supervised learning is a learning in which we train the machine using data which is well labeled. Then the machine is provided with a new set of data so that the algorithm analyses the training data and produces a correct outcome from labeled data.

Supervised learning is classified into two set of algorithms:

- Classification
  A classification problem is when the output variable is a category. It is an approach in which the program learns from the data input given

to it and uses this learning to classify new observation. It is used for predicting discrete responses.

- Regression
  A regression problem is when the output variable is a real value. It is a statistical approach to find the relationship between variables.

Unsupervised Learning

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. It learns from test data that has not been labeled, classified or categorized. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unsupervised learning classified into two categories of algorithms:

- Clustering
  Clustering involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. It is a common technique for statistical data analysis used in many fields. Often however, cluster analysis overestimates the similarity between groups and does not treat data points as individuals.

- Association
  Association mining is a rule based method for finding interesting relations between variables in large databases. It identifies set of items that frequently occur together in dataset.

1.3.1 Binary classification

Binary classification is a task of classifying the elements of given set into two groups on the basis of a classification rule (Classification rule is a procedure by which elements of a set are each predicted to belong to one of the classes). The actual output of many binary classification algorithms is a prediction score. The score indicates the systems certainty that the given observation belongs to the positive class. To make the decision about whether the observation should be classified as positive or negative, as a consumer of this score, you will interpret the score by picking a classification threshold (cut-off) and compare the score against it. Any observations with scores higher than the threshold are then predicted as the positive class and scores lower than the threshold are predicted as the negative class.

1.3.2 Multi Class classification

Multi class classification is a task with multiple classes i.e more than two classes. Each training point belongs to one of the N classes. This classification task assumes that each sample is assigned to one and only one label. Given a new data point, the goal is to construct function which will correctly predict the class to which the new point belongs. In multiclass problems the classes are exclusive.

1.3.3 Multi Label classification

Multi label classification is a classification problem where multiple target labels can be assigned to each observation instead of only one like in multi class classification. This is a generalization of multi class classification, in the multi-label problem there is no constraint on how many of the classes the instance can be assigned to.For multi label problems each label represents a different classification task,but the tasks are some how related.

1.4 Ensemble Machine Learning techniques

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, non stationary learning and error-correcting. We use such an approach routinely in our daily lives by asking the opinions of several experts before making a decision. For example, we typically ask the opinions of several doctors before agreeing to a medical procedure, we read user reviews before purchasing an item, we evaluate future employees by checking their references, etc. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new examples. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that ensembles are often much more accurate than the individual classifiers that make them up. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse. An accurate classifier is one that has an error rate of better than random guessing on new

'x' values. Two classifiers are diverse if they make different errors on new data points.

1.4.1 Advantages of Ensemble

- Statistical
  A learning algorithm can be viewed as searching a space 'H' of hypotheses to identify the best hypothesis in the space. The statistical problem arises when the amount of training data available is too small compared to the size of the hypothesis space. Without sufficient data, the learning algorithm can find many different hypotheses in H that all give the same accuracy on the training data. By constructing N ensemble out of all these accurate classifiers, the algorithm can average their votes and reduce the risk of choosing the wrong classifier.

- Computational
  Many learning algorithms work by performing some form of local search that may get stuck in local optima. In cases where there is enough training data, it may still be very difficult computationally for the learning algorithm to find the best hypothesis. An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual classifiers.

- Representational
  In most applications of machine learning, the true function f cannot be represented by any of the hypotheses in H. By forming weighted sums of hypotheses drawn from H, it may be possible to expand the space of re-presentable functions. The representational issue is somewhat subtle because there are many learning algorithms for which H is in principle space of all possible classifiers. We must consider the space H to be the effective space of hypotheses searched by the learning algorithm for a given training data set.

  These three fundamental issues are the three most important ways in which existing single learning algorithms fail. Hence, ensemble methods have the promise of reducing (and perhaps even eliminating) these three key shortcomings of standard learning algorithms.

1.4.2 Types of Ensemble

Bagging or Bootstrap Aggregating

Bootstrapping is a sampling technique in which we create subsets of observations from the original dataset, with replacement. The size of the subsets is the same as the size of the original set.

Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set. Multiple subsets are created from the original dataset, selecting observations with replacement. A base model (weak model) is created on each of these subsets. The models run in parallel and are independent of each other. The final predictions are determined by combining the predictions from all the models. The image below will help explain:

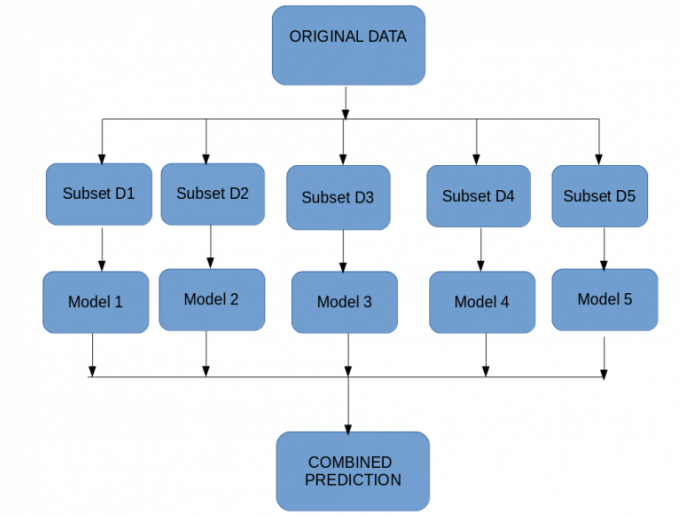

Figure 1.1: Bagging

Boosting

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Lets understand the way boosting works in the below steps.

1. A subset is created from the original dataset.
2. Initially, all data points are given equal weights.
3. A base model is created on this subset.

4. This model is used to make predictions on the whole dataset.
5. Errors are calculated using the actual values and predicted values.
6. The observations which are incorrectly predicted, are given higher weights.
7. Another model is created and predictions are made on the dataset. (This model tries to correct the errors from the previous model)
8. Similarly, multiple models are created, each correcting the errors of the previous model.
9. The final model (strong learner) is the weighted mean of all the models (weak learners).

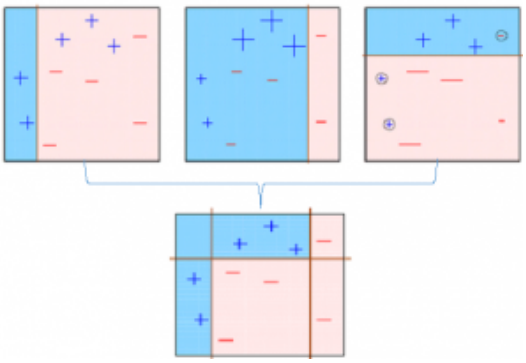

Figure 1.2: Boosting

Stacking

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features.

The base level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous. This developed model is used for making predictions on the test set. Below is a step-wise explanation for a simple stacked ensemble:

1. The train set is split into k parts.
2. A base model is fitted on k-1 parts and predictions are made for the Kth part. This is done for each part of the train set.

3. The base model (in this case, decision tree) is then fitted on the whole train dataset.
4. Using this model, predictions are made on the test set.
5. Steps 2 to 4 are repeated for another base model resulting in another set of predictions for the train set and test set.
6. The predictions from the train set are used as features to build a new model.
7. This model is used to make final predictions on the test prediction set.

Basic concepts of Ensemble

- Averaging
  Multiple predictions are made for each data point in averaging. In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.
- Majority Vote
  The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a vote. The predictions which we get from the majority of the models are used as the final prediction.
- Weighted Average
  This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction. For instance, if two of your colleagues are critics, while others have no prior experience in this field, then the answers by these two friends are given more importance as compared to the other people.

1.5 Ensemble MultiLabel Classification techniques

1.5.1 Problem transformation

Problem transformation methods map the multi-label learning task into one or more single-label learning tasks.

- Binary Relevance
  The technique, which treats each label as a separate single class classification problem.

  In considering the Fig 1.3, we have the data set like this, where X is the independent feature and Ys are the target variable.

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

Figure 1.3: Binary Relevance - Dataset

In binary relevance, this problem is broken into 4 different single class classification problems as shown in the Fig 1.4

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Figure 1.4: Binary Relevance

- Classifier Chains
  In Classifier Chains the first classifier is trained just on the input data and then each next classifier is trained on the input space and all the previous classifiers in the chain.

  In the dataset given in Fig 1.5, X is the input space and Ys are the labels.

  In classifier chains, this problem would be transformed into 4 different single label problems, as shown in Fig 1.6. Here yellow colored is the input space and the white part represent the target variable.

- Label Powerset
  In this, we transform the problem into a multi-class problem with one

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Figure 1.5: Classifier Chains - Dataset

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

Classifier 4

Figure 1.6: Classifier Chains

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

Figure 1.7: Label Powerset - Dataset

multi-class classifier is trained on all unique label combinations found in the training data.

In this, we find that x1 and x4 have the same labels, similarly, x3 and x6 have the same set of labels. So, label powerset transforms this problem into a single multi-class problem as shown below.

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Figure 1.8: Label Powerset

Ensemble Multi label classification

RAkEL, random k label sets is the ensemble multi label classification algorithm and is explained as follows.

The RAkEL algorithm iteratively constructs an en- semble of m Label Powerset (LP) classifiers. At each iteration, i = 1..m, it randomly selects a k-labelset, Y i , from L k without replacement. It then learns an LP classifier h i : X P (Y i). The pseudocode of the ensemble production phase is given in Figure 1.9.

Input: Number of models m, size of labelset k, set of labels L, training set D
Output: An ensemble of LP classifiers h_i and corresponding k-labelsets Y_i
$R \leftarrow L^k$;
for $i \leftarrow 1$ **to** $\min(m, |L^k|)$ **do**
  $Y_i \leftarrow$ a k-labelset randomly selected from R;
  train an LP classifier $h_i : X \rightarrow P(Y_i)$ on D;
  $R \leftarrow R \setminus \{Y_i\}$;

Figure 1.9: RAkEL Pseudo code

The number of iterations (m) is a user-specified parameter with acceptable values ranging from 1 to —L k —. The size of the labelsets (k) is another user- specified parameter with meaningful values ranging from 2 to —L— 1. For k = 1 and m = —L— we get the binary classifier ensemble of the Binary Relevance (BR) method, while for k = —L— (and consequently m = 1) we

get the single- label classifier of the LP method. We hypothesize that using labelsets of small size and an adequate number of iterations, RAKEL will manage to model label correlations effectively. The experimental study in Section 5 provides evidence in support of this hypothesis and guidelines on selecting appropriate values for k and m. For the multilabel classification of a new instance x, each model h i provides binary decisions h i (x, j) for each label j in the corresponding k-labelset Y i . Subsequently, RAKEL calculates the average decision for each label j in L and outputs a final positive decision if the average is greater than a user-specified threshold t. An intuitive value for t is 0.5, but RAKEL performs well across a wide range of t values as it shown by the experimental results. The pseudocode of the ensemble production phase is given in Figure 1.10.

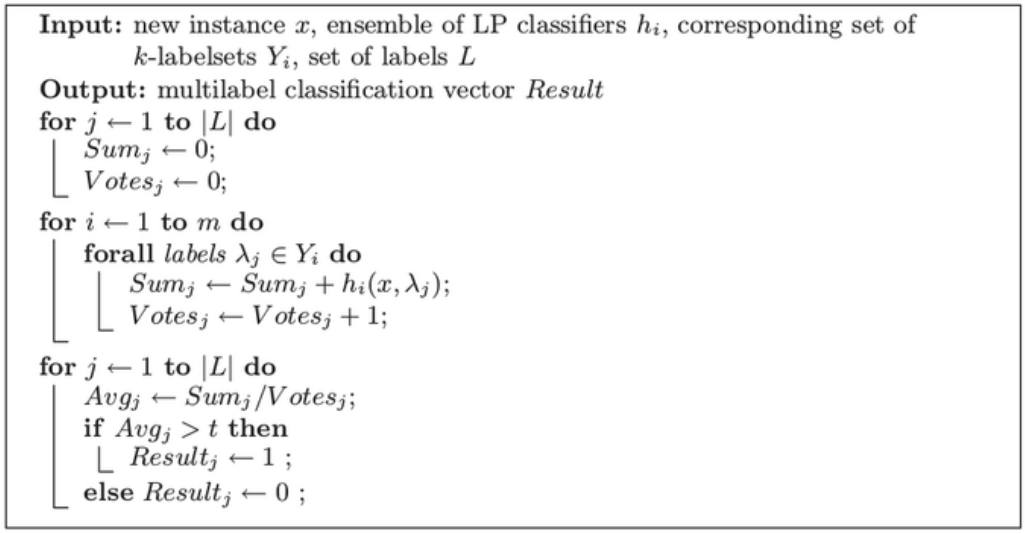

Figure 1.10: RAkEL code

1.6 Road Map

The rest of this thesis is organized as follows; Chapter 2 describes about the problem statement of the study. A detailed literature survey on learning disabilities, ensemble learning and multi label classification is described in chapter 3. Chapter 4 deals with the proposed system of the study which includes proposed stacking ensemble model for prediction of learning disability and multi label classification to find out the type of learning disabilities in children. Chapter 5 consists of all the experimental results. Chapter 6 discusses about the project conclusion.

Chapter 2

Problem Definition

Given the data of school aged children with neurological disorders, the aim is to

- Identify the important parameters of learning disability (LD) for predicting the occurrence of LD using ensemble technique.
- Predict the multiple learning disabilities for a single child using ensemble approaches of multi label classification techniques.

Chapter 3

Related Work

3.1 Literature Survey on prediction of Learning Disability

Julie M. David and Kannan Balakrishnan [1] conveyed the importance of two classification techniques (decision tree and clustering) in the prediction of Learning Disabilities in school-age children and exhibited the importance of clustering in finding different signs and symptoms in Ld affected children. Decision tree implemented using the J48 algorithm on consistent data produced a better result in terms of efficiency and complexity. Clustering implemented using K means helped in developing clusters that predict a binary yes or no output for the existence of LD.

Julie M. David and Kannan Balakrishnan [2] implemented rough sets and decision tree for the prediction of learning disabilities in children. The rough Set based algorithm will find a rule with a reduced number of conditions so that only those combinations of input values which appear in the data can be included. For attribute reduction and analysis, Johnsons Reduction and Naive Bayes Classifier are performed. The J48 algorithm is used in Decision Tree implementation. Results are compared with a similar study on Naive Bayes and SVM. It was observed that decision trees are poorer in certain aspects compared to Rough Sets, as Rough Sets deal with inconsistent data and also provides information about attribute correlation. In Decision Tree the main objective of an attribute is based on information gain and in Rough sets evaluation is based on the elimination of redundant attribute.

Julie M. David and Kannan Balakrishnan [3] prioritised the applications

of Data Mining by proposing SVM using Sequential Minimal Optimisation which provided an accuracy of 97.86% and Decision Tree using J48 that provided an accuracy of 97.47% in predicting LD in school-age children. Even though time taken by SVM is large as compared to DT, SVM is proved to give an accurate result since it can handle inconsistent data in a better way in comparison with DT.

Pooja Manghirmalani et al [4] conducted a comparative study of different soft computing models to diagnose Learning Disabilities and proposed a fuzzy approach in classifying LD into Dyslexia, Dysgraphia and Dyscalculia which helped in diagnosing LD with accuracy but is difficult to judge specific LD among the children.

The research study of Ambili K, Afsar P [5] focus to apply a data mining approach to predict the child's learning behavior and skills using machine learning algorithms. The information gained from this evaluation is furthur used to predict learning disbility found in children. The observations show that the prediction model developed using a Hybrid Naive Bayes and Decision Tree fusion technique-NB Tree is found to be the best among classification and prediction algorithms for child development analysis and learning disability prediction.

The author [6] aims at analyzing various data mining techniques for the prediction of learning disability. The observations show that the fusion technique of naive bayes and neural network is found to be the best among classification and prediction algorithms in the diagnosis of learning disability when compared to other machine learning algorithms

3.2 Literature Survey on Ensemble Techniques

C.V. Krishna Veni, T. Sobha Rani [7] developed a frame work which uses less than one-third of the data set for training and tests the remaining two-thirds of the data and still gives results comparable to other classifiers. To achieve good classification accuracy with small training sets, they focused on three issues: The first is that, one-third of the data should represent the entire data set. The second is on increasing the classification accuracy even with these small training sets, and the third issue is on taking care of deviations in the small training sets like noise or outliers. First issue is addressed by

proposing three methods: divide the instances into 10 bins based on their distances from the centroid, based on their distance from a reference point 3/2(min+max) and a distribution specific binning. In all these methods, training sets are formed using stratified sampling approach which ensures that the samples chosen are from the entire distribution. Second issue is dealt with using the concept of ensemble based weighted majority voting for classification. Third issue is tackled by implementing four filters on training sets. The filters used are removing outliers using Inter Quartile Range option and removing misclassified instances applying Naive Bayes, IB3, IB5 as filters. Experiments are conducted on seven binary and multi-class data sets taking only 6 to 18 percent of the total data for training and implemented the proposed three methods without any filters

In [8], a comparative study of different Ensemble Learning techniques has been presented using the Wisconsin Breast Cancer data set. The primary objective behind using Ensemble learning here is a classification task. This comparative study should help the researchers to find the suitable Ensemble Learning technique for improving their results.

Rammohan Mallipeddi and Ponnuthurai N. Suganthan [9] proposed an ensemble of constraint handling techniques (ECHT) to solve constrained real-parameter optimization problems, where each constraint handling method has its own population. A distinguishing feature of the ECHT is the usage of every function call by each population associated with each constraint handling technique. Being a general concept, the ECHT can be realized with any existing EA. In this paper, they presented two instantiations of the ECHT using four constraint handling methods with the evolutionary programming and differential evolution as the EAs. Experimental results show that the performance of ECHT is better than each single constraint handling method used to form the ensemble with the respective EA, and competitive to the state-of-the-art algorithms.

The article [10] targets the task of content-based multiple-instance people retrieval from video surveillance footage. This task is particularly challenging when applied on such datasets as the available samples to train the decisioning system and formulate the query are insufficient (one image, few frames, or seconds of video recording). To cope with these challenges in [4] they investigated three established ensemble-based learning techniques, e.g., boosting, bagging and blending (stacking). Such methods are based on a set

of procedures employed to train multiple learning algorithms and combine their outputs, while functioning together as a unified system of decision making. The approach was evaluated on two standard data sets (accounting for 16 people searching scenario on ca. 53000 labeled frames). Performance in terms of F2-Score attained promising results while dealing with our current task.

Monisha Kanakaraj1 and Ram Mohana Reddy Guddeti [11] analyzed the mood of the society on a particular news from Twitter posts. The key idea was to increase the accuracy of classification by including Natural Language Processing Techniques (NLP) especially semantics and Word Sense Disambiguation. The mined text information is subjected to Ensemble classification to analyze the sentiment. Ensemble classification involves combining the effect of various independent classifiers on a particular classification problem. Experiments conducted demonstrate that ensemble classifier outperforms traditional machine learning classifiers by 3-5 percent

Lihua Hao, Christopher G. Healey and Steffen A.[12] Bass demonstrated techniques on an ensemble studying matter transition from hadronic gas to quark-gluon plasma during gold-on-gold particle collisions.

3.3 Literature Survey on Multi Label Classification

Hamed Bonab and Fazli Can[13] introduced a novel online and dynamically-weighted stacked ensemble for multi-label classification, called GOOWEML, that utilizes spatial modeling to assign optimal weights to its component classifiers.Their model can be used with any existing incremental multi-label classification algorithm as its base classifier and conducted experiments with 4 GOOWE-ML-based multi-label ensembles and 7 baseline models on 7 real-world datasets from diverse areas of interest.Their experiments show that GOOWE-ML ensembles yield consistently better results in terms of predictive performance in almost all of the datasets, with respect to the other prominent ensemble models.

In[14]the authors proposed a novel Ensemble Label Power-set Pruned datasets Joint Decomposition (ELPPJD).First,they transformed the multilabel classi-

fication into a multiclass classification. Then, proposed the pruned datasets and joint decomposition methods to deal with the imbalance learning problem. Two strategies size balanced (SB) and label similarity (LS) are designed to decompose the training dataset. In the experiments, the dataset is from the real physical examination records.They contrasted the performance of the ELPPJD method with two different decomposition strategies. Moreover, the comparison between ELPPJD and the classic multilabel classification methods RAkEL and HOMER is carried out. The experimental results show that the ELPPJD method with label similarity strategy has outstanding performance.

The simplest form of ensemble learning is to train the base-level algorithms on random subsets of data and then let them vote for the most popular classifications or average the predictions of the base-level algorithms.An ensemble learning method is proposed for improving multi-label classification evaluation criteria[15].They have compared their method with well-known base-level algorithms on some data sets. Experiment results show the proposed approach outperforms the base well-known classifiers for the multi-label classification problem.

Rokach, L., Schclar, A. and Itach [16] selected the minimum required subsets of k labels that cover all labels and meet additional constraints such as coverage of inter-label correlations. Construction of the cover is achieved by formulating the subset selection as a minimum set covering problem (SCP) and solving it by using approximation algorithms. Every cover needs only to be prepared once by offline algorithms. Once prepared, a cover may be applied to the classification of any given multi-label dataset whose properties conform with those of the cover. The contribution is two-fold. First, they introduced SCP as a general framework for constructing label covers while allowing the user to incorporate cover construction constraints.They demonstrated the effectiveness of this framework by proposing two construction constraints whose enforcement produces covers that improve the prediction performance of random selection by achieving better coverage of labels and inter-label correlations. Second,they provided theoretical bounds that quantify the probabilities of random selection to produce covers that meet the proposed construction criteria. The experimental results indicate that the proposed methods improve multi-label classification accuracy and stability compared to the RAKEL algorithm and to other state-of-the-art algorithms.

In [17] the authors discussed some interesting properties of BR, mainly that it produces optimal models for several ML loss functions. Additionally, They presented an analytical study of ML benchmarks datasets and point out some shortcomings. As a result, they proposed the use of synthetic datasets to better analyze the behavior of ML methods in domains with different characteristics. To support this claim, They performed some experiments using synthetic data proving the competitive performance of BR with respect to a more complex method in difficult problems with many labels, a conclusion which was not stated by previous studies.

Martin Boros, Jiri Marsik and Franky [18] applied ensemble techniques, which have proven to be effective in solving other multi-label classification problems, to combine them. They implemented seven ensemble techniques presented in previous work and evaluated their performance. They found that some of the ensemble classifiers outperform all of the individual classifiers. Ensemble techniques have thus proven themselves to be applicable to the domain of text classification

The authors in [19] proposed breaking the initial set of labels into a number of small random subsets, called labelsets and employing LP to train a corresponding classifier. The labelsets can be either disjoint or overlapping depending on which of two strategies is used to construct them. The proposed method is called RAkEL (RAndom k labELsets), where k is a parameter that specifies the size of the subsets. Empirical evidence indicates that RAkEL manages to improve substantially over LP, especially in domains with large number of labels and exhibits competitive performance against other high-performing multilabel learning methods.

Jesse Read et al[20], stated that binary relevance-based methods have much to offer, especially in terms of scalability to large datasets. They exemplified this with a novel chaining method that can model label correlations while maintaining acceptable computational complexity. Empirical evaluation over a broad range of multi-label datasets with a variety of evaluation metrics demonstrates the competitiveness of our chaining method against related and state-of-the-art methods, both in terms of predictive performance and time complexity.

The authors in [21] introduced the task of multi-label classification, organized the sparse related literature into a structured presentation and performed

comparative experimental results of certain multi-label classification methods. They also contributed the presentation of an undocumented method and the definition of a concept for the quantification of the multi-label nature of a data set.

Chapter 4

Proposed System

The main aim of this project is to predict whether a child is facing learning disability or not and if so classify it into different types of learning disabilities.

4.1 Multi-layer ensemble model Establishment

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. The proposed stacking ensemble model fuses multiple algorithms by considering the combination of these base learners to train the top level learner. An ensemble model used in the implementation boosts up the accuracy of classification and prediction in comparison with individual algorithms. Not only for better accuracy but combining multiple algorithms to form an ensemble model can also be done when input dataset is extensive and spatially distributed. Therefore, proposed ensemble model helps to develop a model that produces low variance without compromise in bias.

In this study a two-layer stacking ensemble model is implemented. Stacking also known as a stacked generalization uses algorithms in multiple levels and gives best results for small as well as large data sets. The two layered stacking ensemble model comprising of base and top level learners is shown. In stacking predictions taken from base level learners are given to top level learners. Initially, at the base layer, the data with m instances and n features are trained on M models. The predictions from each model on 'm' instances are taken as a dataset with a dimension of m rows and M columns. This formed data set obtained is sent to the top level as its input. Using this input data, the top level model is trained and produces the final outcome. Stacking uses K-fold cross validation in taking the predictions from base layer.

4.1.1 K-fold Cross Validation

Cross-validation is a re-sampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model

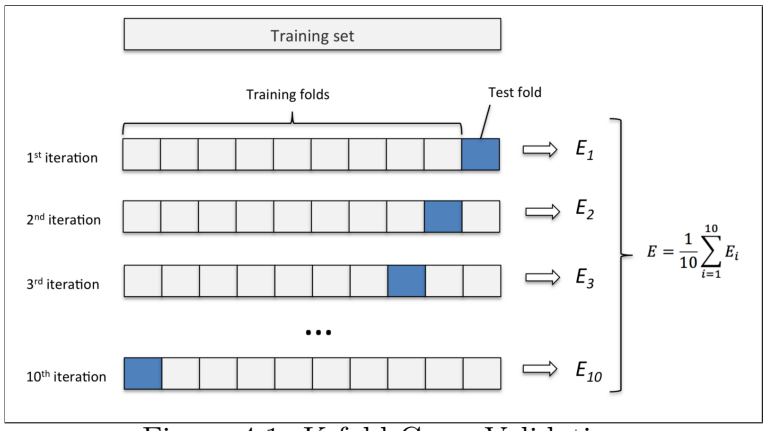

Figure 4.1: K-fold Cross Validation

The k value must be chosen carefully for your data sample. A poorly chosen value for k may result in a miss-representative idea of the skill of the model, such as a score with a high variance (that may change a lot based on the data used to fit the model), or a high bias, (such as an overestimate of the skill of the model).

Suppose in a model, if different folds are applied, the predictions of the model for the training and test data may change, and one fold may result in a better CV score than another fold. Since the same model and training data can be used, the predictive power of the same models trained with different K-folds must be the same for the unseen test data. If the predictions of the same models are used with different folds as inputs for stacking, the ensemble model will weigh one with better CV score higher than one with worse CV score, while two are exactly the same predictors for the test data. Even worse, if the predictions of different models with different folds as inputs are used for stacking, again, the ensemble model will weigh one with better CV

score higher than one with worse CV score, while the former may actually be a worse predictor for the test data. Therefore, to eliminate potential bias that can be introduced by using different folds, the same folds are used across models in stacking.

This method takes out-of-fold predictions which are used as input to the second level. In calculating out-of-fold predictions, rows are taken identical for all models, whereas data attributes can be considered in two ways to obtain a good prediction accuracy. One way is, base models are trained on different subsets of input features which will add randomness to the models and improves the ensemble performance. The other way is utilizing all the attributes as input features to yield the best result, as more data results in better models. In this , all attributes are considered as input features to obtain predictions. In conjunction with the input features, out-of-fold predictions play a major part in stacking. For the top level training, if the predictions were formed by fitting the model on all the training data and passed to the top level then the top level gets biased to the best of base level models. This defines the way of choosing the base level models. In base level models if one model has lower training accuracy but if it performs better on some data points and another model has high training accuracy but performs worse on some other data points then combining them will be of no use. In order to get the top model biased and have a large possibility of taking the best performance of each model, the out-of-fold predictions are created. The obtained out-of-fold predictions are given to the top level model for training and the final predictions are calculated from the model. This is depicted in Fig 4.4. along with the algorithms used in the study.

The above constructed multi-layer model is used for both prediction and multi label classification of LD in children. The following sections portrays the machine learning algorithms and multi-label classification methods used in constructing respective stacking model.

4.2 Prediction of Learning Disability

The above constructed two-layer stacking model uses three general machine learning algorithms at base level and one at the top level. The model for prediction of learning disability uses logistic regression, KNN, SVM at the base level and decision tree at the top level. This ensemble fuses the base level model algorithms and capture the predictions from three models into a prediction matrix. This prediction matrix is forwarded to train on decision tree and get the final prediction result. Logistic regression, knn and svm are individually executed and compared with the stacking ensemble.

4.3 Multi label classification for predicting the type of Learning Disability

Multi label classification techniques help in classifying the data that has multiple labels. As described in introduction binary relevance, classifier chains and label power set are individually executed and results are calculated. RAkEL, an ensemble multi label classification technique is executed to compare the predictions with individual multi level classification algorithms. Along with RAkEL, stacking ensemble model described in the above section is also implemented to have a comparative study. In this stacked ensemble model, binary relevance and classifier chains as base level algorithms along with label powerset at top level. Naive bayes, logistic regression, decision tree and SVM are the learners used in binary relevance, classifier chains, label powerset and RAkEL respectively. Random forest algorithm is used in adaption algorithms. A comparative study of binary relevance, classifier chain, adaption algorithm, RAkel and stacked ensemble model is done and results are analysed.

4.4 Algorithms Used

4.4.1 Logistic Regression

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.To squash the predicted value between 0 and 1, we use the sigmoid function.

$$g(z) = \frac{1}{1 + exp(-z)}$$

The logistic regression hypothesis is then defined as:

$$h_\beta(x) = g(\beta^T x)$$

Logistic regressions are usually fit by maximum likelihood. The cost function we want to minize is the opposite of the log-likelihood function:

$$\frac{dJ(\beta)}{d\beta} = \frac{1}{m} \sum_{i=1}^{N} x_i (h_\beta(x_i) - y_i) = 0$$

4.4.2 Decision Tree

Decision Trees are a type of supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. CART(Classification and Regression Trees) - uses Gini Index as metric and ID3 (Iterative Dichotomiser 3) - uses Entropy function and Information Gain as metrics are the couple of algorithms which are used for building a decision tree.

Entropy: Entropy, as it relates to machine learning, is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

$$H(x) = \sum_{i=1} p(x) \log_2 \frac{1}{p(x)}$$

Information Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain

$$IG(S, A) = H(S) - \sum_{i=1}^{n} p(x) * H(x)$$

4.4.3 Support Vector Machine

The target of the support vector machine calculation is to discover a hyperplane in a N-dimensional space (N-the number of highlights) that particularly orders the information focuses. In this calculation, each data is plotted as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each component being the estimation of a specific arrange. At that point, order is performed by finding the hyperplane that separate the two classes great.

4.4.4 K-Nearest Neighbours

K - Nearest Neighbours: KNN can be utilized for both characterization and relapse prescient issues. Be that as it may, it is all the more broadly utilized in grouping issues in the business. It fairs over all parameters of contemplation's. It is normally utilized for its simple of elucidation and low computation time. It uses Euclidean distance to calculate the minimum distance between the data points.

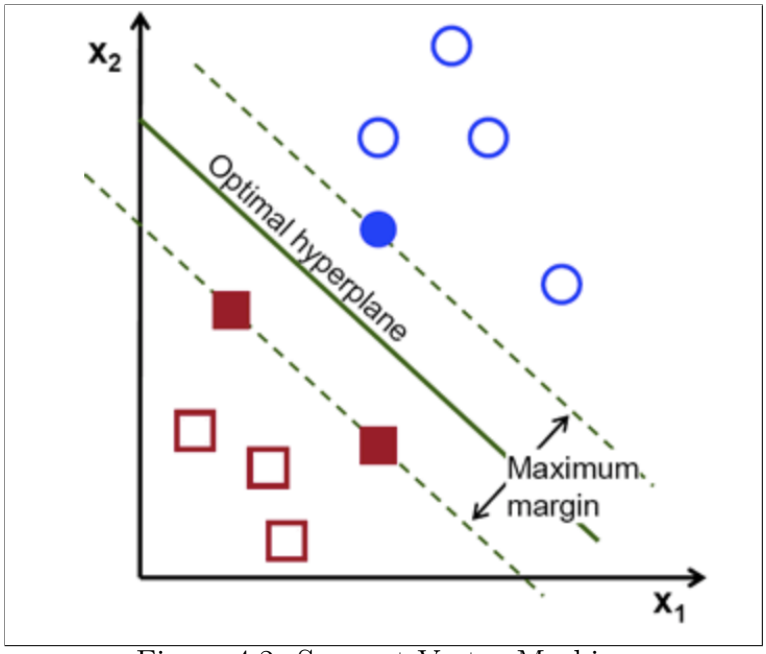

Figure 4.2: Support Vector Machine

$$Euclideandistance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The utilization of these classifier algorithms in our research is portrayed as follows: In the first level, Logistic Regression, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) algorithms are used. SVM is assumed to be a good classifier because of high classification accuracy and is best to be used when data sets are comparatively small. SVM being a good classifier taking predictions on training data at the base level is assumed to be precise. KNN, on the other hand, classifies the data points based on the similarity measure. And Logistic Regression focuses on the association of dependent variable i.e. target variable with independent variables i.e. features. It uses a sigmoid function which helps in binding the output in the range of [0,1] and helps to avoid over fitting. In the top-level Decision tree is chosen as meta-classifier since it is comparatively fast and is easy for creating rules in a classification problem. This Decision tree combines the classifiers at the low-level with its predictions and builds a complete stacked ensemble. Figure 4.3 depicts the stacked ensemble model of the study.

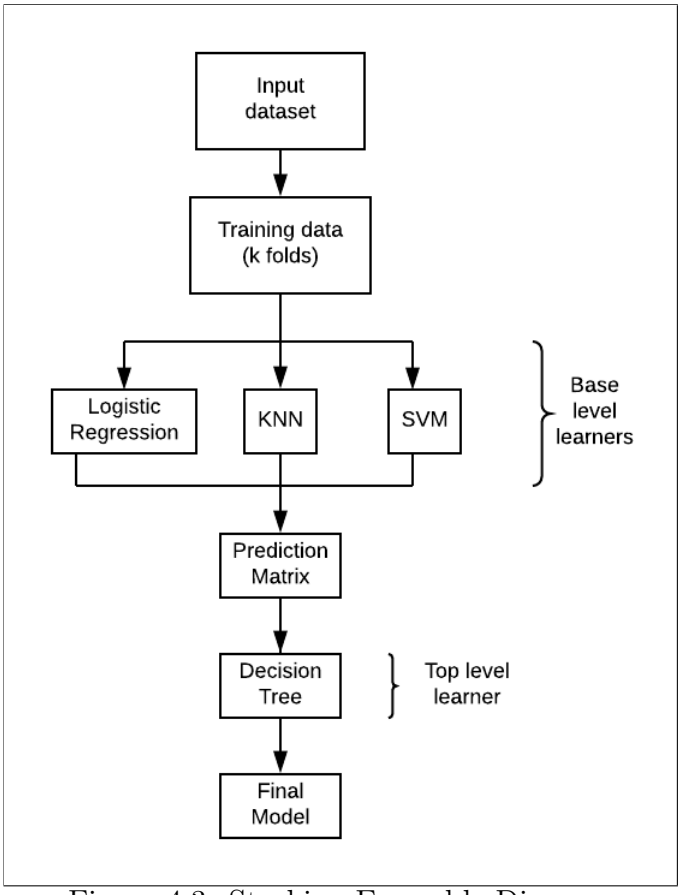

Figure 4.3: Stacking Ensemble Diagram

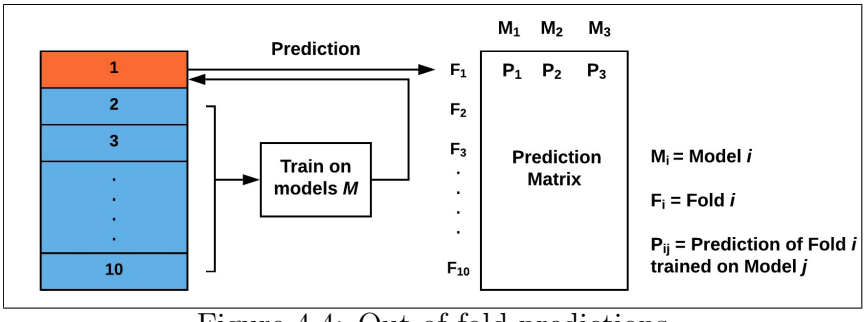

Figure 4.4: Out-of-fold predictions

4.4.5 Random Forest

Random Forests are an improvement over bagged decision trees. A problem with decision trees like CART is that they are greedy. They choose which variable to split on using a greedy algorithm that minimizes error. As such, even with Bagging, the decision trees can have a lot of structural similarities and in turn have high correlation in their predictions.

Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation. It is a simple tweak. In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited to a random sample of features of which to search.

4.4.6 Naive Bayes

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3—h), they are assumed to be conditionally independent given the target value and calculated as P(d1—h) * P(d2—H) and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

The representation for naive Bayes is probabilities.

A list of probabilities are stored to file for a learned naive Bayes model. This includes:

Class Probabilities: The probabilities of each class in the training dataset. Conditional Probabilities: The conditional probabilities of each input value given each class value.

Chapter 5

Testing and result analysis

5.1 Dataset

The experiment was conducted on LD dataset comprising 900 instances of patient records and 20 attributes. The study being related to the medical field, the data has to be real-time data based on the accurate symptoms and actions of the learning disabled children. This real-time data is obtained either through surveys or from hospitals. For this project the data was collected from near by hospital. Attributes are extracted from the unstructured data into a CSV file. The 20 attributes consist of gender, age, grade, academic performance and disabilities faced by children which include difficulty in reading, writing, spelling, copying (whether a child is able to copy the given text), decoding, fine motor skills (daily activities such as tying shoe lace), maths, attention, processing speed, auditory discrimination and memory, visual discrimination and memory, visuo-motor skills. The detailed information of attributes used in our research work are shown in Table 5.1.

Based on the values of each attribute the instances of the dataset are divided into dyslexia, dysgraphia, dyscalculia and ADHD.

5.1.1 Data Automation

Obtained data from hospitals, is an unstructured data containing all the medical diagnosis of a child and is in excel format. For using it in machine learning algorithms, it has to be structured into a csv file. For this named entity recognition, a NLP technique has been used in collecting the attributes for the project. This is just a sample implementation for automation. After implementing, attributes are collected into an excel file as mentioned in table 5.1.

Attritube	**Description**
Gender	Gender of the patient
Age	Age of the patient at the time of diagnosis
Class	Grade of the patient at the time of diagnosis
Academic	Academic performance of the patient
Reading	Whether a patient is facing difficulty in Reading
Writing	Whether a patient is facing difficulty in Writing
Spelling	Whether a patient is facing difficulty in spellings
copying	Whether a patient is facing difficulty while copying
Decoding	Whether a patient is facing difficulty in identifying already shown things
Fine Motor Skills	Whether a patient is facing difficulty in daily activities i.e, tying a shoelace, disorder of shirt buttons
Maths	Whether a patient is facing difficulty in solving maths problems
Attention	Whether a patient is attentive
HyperActive	Whether a patient shows symptoms of being Hyperactive
Impulsive	Whether a patient shows symptoms of being Impulsive
Processing Speed	Whether a patient is slow in doing and reacting to the mental task
Auditory Discrimination	Whether a patient is able to differentiate sounds
Auditory Memory	Whether a patient is able to take in the information orally and recall it
Visual Memory	Whether a patient is able to recall the images and visuals
Visual Discrimination	Whether a patient is able to differentiate objects.
Visuo-Motor	Whether a patient is able to control vision and movement together

Table 5.1: Attributes of the dataset

Named Entity Recognition

The process of detecting the named entities from the text is called as named entity recognition. A typical named entity recognition model consists of three blocks:

Noun phrase identification: This step deals with extracting all the noun phrases from a text using dependency parsing and part of speech tagging.

Phrase classification: This is the classification step in which all the extracted noun phrases are classified into respective categories. We can curate the lookup tables and dictionaries by combining information from different sources.

Entity disambiguation: Sometimes it is possible that entities are miss classified, hence creating a validation layer on top of the results is useful. Use of knowledge graphs can be exploited for this purposes.

5.1.2 Data Distribution

Multi label data is data with multiple classes or labels. There are 4 different labels namely dyslexia, dysgraphia, dyscaulcia and ADHD. This section describes on how data of total 900 records have been distributed and is depicted in the below figures as follows.

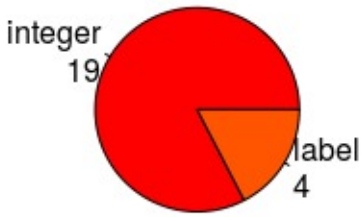

Figure 5.1: Type and number of attributes

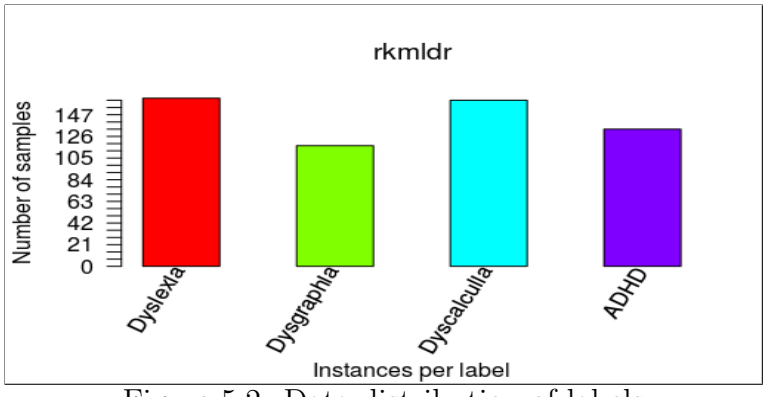

Figure 5.2: Data distribution of labels

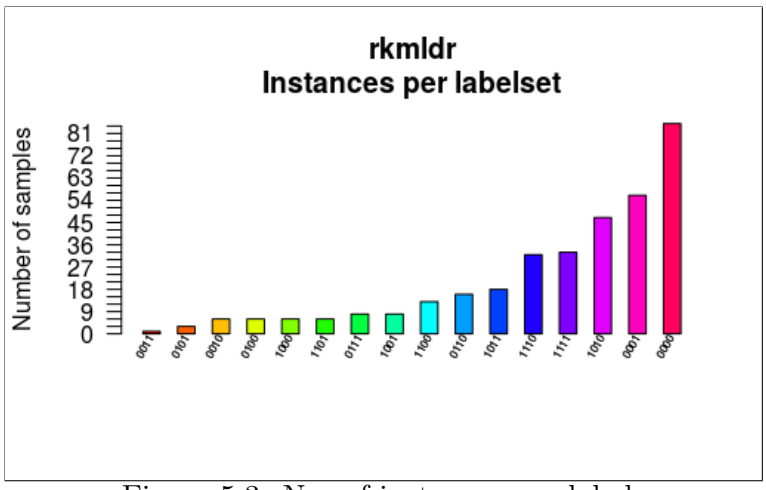

Figure 5.3: No. of instances per labels

5.2 Observations

5.2.1 Results of Prediction of Learning Disability

The built ensemble model as discussed in proposed study is used on the test data and final outcomes are considered for calculating the accuracy of the model. The evaluation metrics selected in our experimentation are accuracy, recall and receiver operating characteristic (ROC) curve. The performance of machine learning models and stacked ensemble model are compared based on the above-mentioned metrics and results are discussed. Table 5.2 and Fig.

Model	Accuracy
Logistic Regression	84.70588
KNN	89.41176
SVM	82.35294
Stacking Ensemble	90.58824

Table 5.2: Accuracy Scores

5.4. shows the accuracy rate of Logistic Regression, KNN, SVM and Stacking Ensemble model. From the accuracy graph, the high percentage value of stacking ensemble depicts that it classifies the data accurately when compared with other individual machine learning algorithms. Accuracy metric portrays how accurate the model is i.e, it calculates the correct predictions against the total number of predictions. Therefore, considering accuracy as the decision metric is not adequate. In the medical issue prediction, we focus on minimising the false negatives of a model that is capturing all the cases that have LD rather than capturing cases that are correctly predicted.

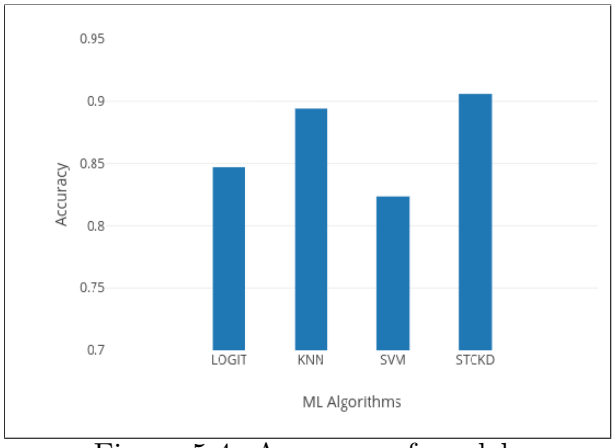

Figure 5.4: Accuracy of models

Recall is one of the performance metrics which displays information about the classifiers performance with respect to false negatives. So, we would consider the greater recall as the better performance model. The recall percentage score of individual models and the stacking model is shown in Table 5.3 and Fig. 5.5. From the figure, it is seen that recall percentage of stacking model is higher than individual algorithms.

$$Recall = \frac{1}{p}\sum_{i=1}^{p} \frac{\mid h(x_i) \cap Y_i \mid}{\mid Y_i \mid}$$

Table 5.3: Recall Scores

Model	**Recall (in percentage)**
Logistic Regression	81.81818
KNN	83.78378
SVM	77.14286
Stacking Ensemble	90.625

ROC curve is another good measure for evaluating the performance of the model at different thresholds. It is the probability curve that helps in ranking the performance of the classifier based on AUC(area under curve). The ROC curve is plotted with a true positive rate on y-axes against false positive rate on x-axes. The higher AUC value distinguishes the class variables accurately. Therefore model with higher AUC is considered as the best model. The ROC graphs and AUC values of the models are shown in Fig. 5.6. From the figure,

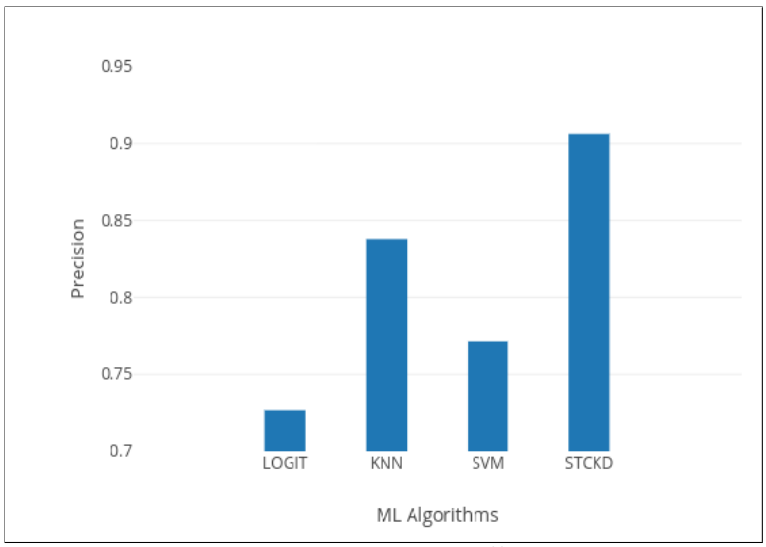

Figure 5.5: Recall

it is observed that the roc curve of Logistic Regression and SVM are lower than KNN and stacking ensemble and the AUC values of KNN and Stacking ensemble are equal.

From all the above-discussed metrics, it is shown that Stacking ensemble model enhances the stability of the model and significantly improved the prediction accuracy of the model.

5.2.2 Results of Multi Label Classification of Learning Disability

As mentioned in introduction and proposed study various multi label classification techniques are used on the dataset in classifying the types of learning disability in children. The evaluation measures for single-label are usually different than for multi-label. In multi-label classification, a miss-classification is no longer a hard wrong or right. A prediction containing a subset of the actual classes should be considered better than a prediction that contains none of them, i.e., predicting two of the three labels correctly this is better than predicting no labels at all. The evaluation metrics selected in our experimentation are accuracy, f1 score, hamming loss, subset accuracy and true positive rate. The performance of individual classifiers, stacked ensemble model are compared based on the above-mentioned metrics and results are discussed.

Table 5.4 depicts the accuracy values of the different multi label classification techniques.

Hamming-Loss is the fraction of labels that are incorrectly predicted

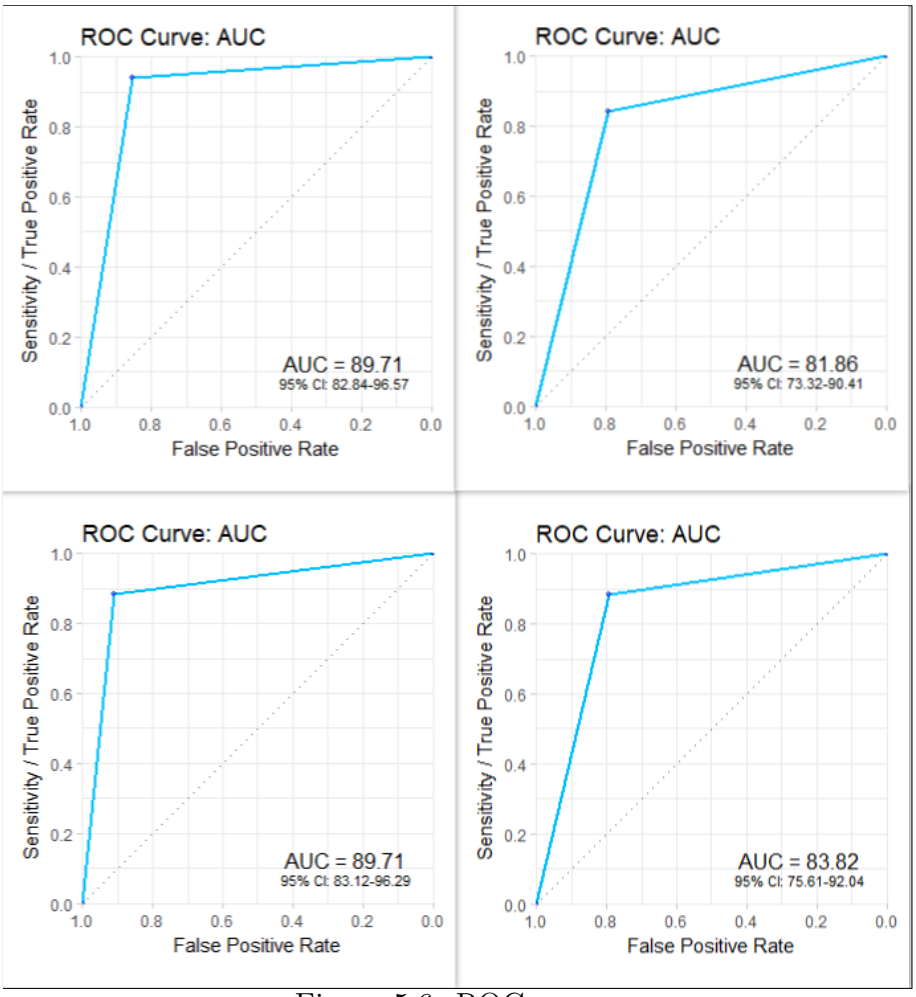

Figure 5.6: ROC curves

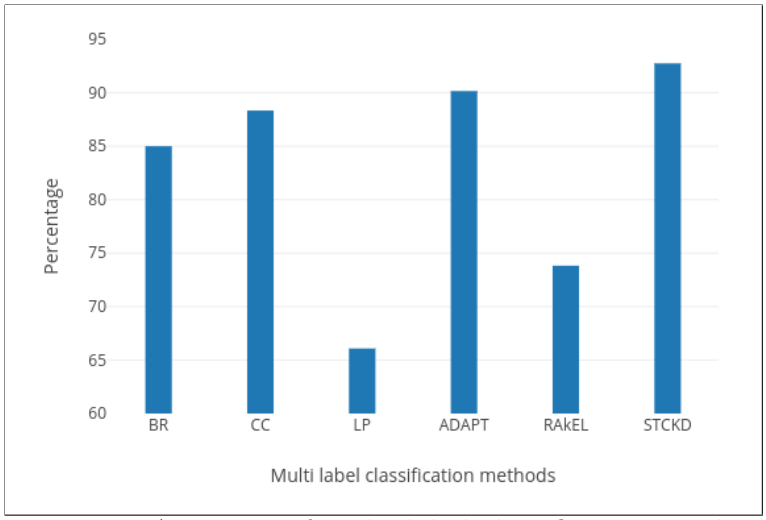

Figure 5.7: Accuracy of multi label classification methods

Models	Accuracy
Binary Relevance	0.85
Classifier Chains	0.883333
Label Powerset	0.660814
Algorithm Adaption	0.90156863
RAkEL	0.73823529
Ensemble	0.92745098

Table 5.4: Accuracy Scores of multil label classification

where as subset accuracy is the proportion of observations where the complete multi label set (all 0-1-labels) is predicted incorrectly. It is different from the subset zero-one loss. The zero-one loss considers the entire set of labels for a given sample incorrect if it does not entirely match the true set of labels. Hamming loss is more forgiving in that it penalizes only the individual labels. These metrics are different from the metrics of single label. The metric results of all the multi label classification methods, proposed stacking ensemble, adaption algorithms and RAkEL are shown in Fig 5.8

$$HammingLoss = \frac{1}{p}\sum_{i=1}^{p} \mid h(x_i)\Delta y_i \mid$$

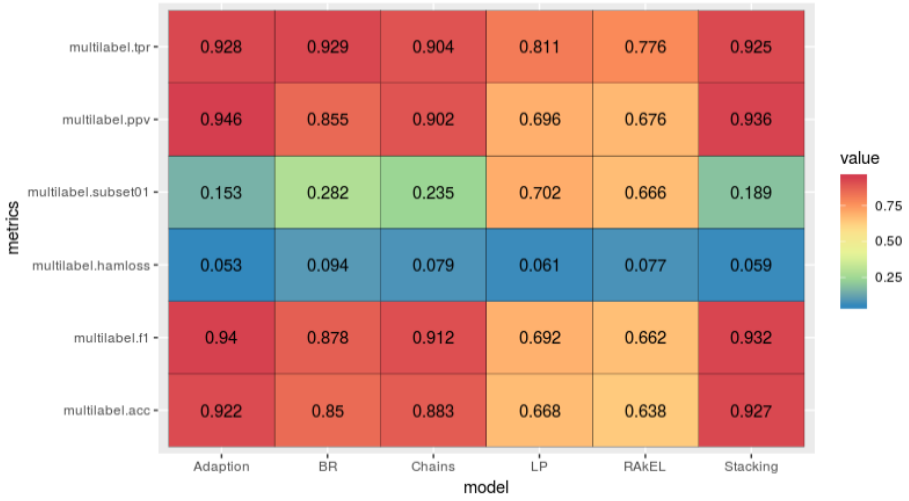

Figure 5.8: Performance of multi label classification methods

From the figure, it seems that the most accurate prediction were obtained using the algorithms that treat the labels independently, and the Binary relevance model worked slightly better than Chain model. The chained model

would work perfectly when there is a clear hierarchical or causative relationship among three pathological entities. From Label powerset, RAkEL and proposed ensemble values, it is clearly depicted that multi-layer approach produces a better model in classification of Learning disabilties in children.

Chapter 6

Conclusion

The main objective of the study is to predict learning disabilities using ensemble technique to improve the performance of predicting followed by classification of disability into dyslexia, dysgraphia, dyscaulcia and ADHD using machine learning and ensemble multi label classification algorithms. In this research, basic models decision tree, SVM, KNN, logistic regression and random forest are used.

The experimental results show that the proposed multi layer ensemble model - stacking produced a better accurate model compared to individual algorithms in finding the presence of LD in school aged children. Stacked ensemble of different multi label classification techniques like binary relevance, label powerset and classifier chains produced better accuracy than individual algorithms execution. RAkEL, an ensemble technique proved to have better accuracy compared to individual algorithms.

References

[1] Julie M. David, Kannan Balakrishnan: Significance of Classification Techniques in Prediction of Learning Disabilities in School Age Children, International Journal of Artificial Intelligence & Applications, 1(4), Oct.2010, pp 111-120.

[2] J. M. David and K. Balakrishnan (2010), Machine Learning Approach for Prediction of Learning Disabilities in School-Age Children, International Journal of Computer Applications, Volume 9, No.11, Pp. 712.

[3] Julie M. David, Kannan Balakrishnan: Prediction of Learning Disabilities in School-Age Children using SVM and Decision Tree, International Journal of Computer Science and Information Technology, 2(2), Mar-Apr. 2011, pp 829-835.

[4] Manghirmalani, P., More, D., Jain, K.: A fuzzy approach to classify learning disability. Int. J. Adv. Res. Artif. Intell. 1(2), U.S ISSN:21654069(Online), U.S ISSN : 21654050(Print) (2012).

[5] Ambili K, Afsar P :A Framework for Child Development analysis and Learning Disability Prediction using a Hybrid Naive Bayes and Decision Tree Fusion Technique NB",International Journal of Innovative Research in Science, Engineering and Technology,Vol. 5, Issue 7, July 2016

[6] Ambili k , Afsar P :A Framework for learning disability prediction in school children using Naive Bayes-Neural Network fusion technique,Journal of information,knowledge and research in computer engineering.

[7] C.V. Krishna Veni, T. Sobha Rani:Ensemble based Classification using Small Training sets : A Novel Approach

[8] Dr. Chandan Banerjee,Sayak Paul,Moinak Ghoshal:A Comparative Study of Different Ensemble Learning Techniques using Wisconsin Breast Cancer dataset

[9] Rammohan Mallipeddi and Ponnuthurai N. Suganthan: Ensemble of Constraint Handling Techniques

[10] C.A. Mitrea1,S. Carata, B. Ionescu, T. Piatrik and M. Ghenescu:Ensemble-based Learning Using Few Training Samples for Video Surveillance Scenarios

[11] Monisha Kanakaraj1,Ram Mohana Reddy Guddeti:Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques

[12] Lihua Hao, Christopher G. Healey : Effective Visualization of Temporal Ensembles

[13] B uy uk cakir, A., Bonab, H. and Can, F. A novel online stacked ensemble for multi-label stream classification. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 1063-1072). ACM, 2018.

[14] Li, R., Liu, W., Lin, Y., Zhao, H. and Zhang, C. An Ensemble Multilabel Classification for Disease Risk Prediction. Journal of health care engineering, 2017.

[15] Mahdavi-Shahri, A., Houshmand, M., Yaghoobi, M. and Jalali, M. Applying an ensemble learning method for improving multi-label classification performance. In 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS) (pp. 1-6). IEEE.

[16] Rokach, L., Schclar, A. and Itach, E. Ensemble methods for multi-label classification. Expert Systems with Applications. 41(16), pp.7507-7523, 2014.

[17] Luaces, O., D ez, J., Barranquero, J., del Coz, J.J. and Bahamonde. Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence. 1(4), pp.303-313, 2012.

[18] Boros, M. and Mars k, J. Multi-label text classification via ensemble techniques. International Journal of Computer and Communication Engineering, 1(1), pp.62-65, 2012.

[19] Tsoumakas, G., Katakis, I. and Vlahavas, I. Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering, 23(7), pp.1079-1089, 2011.

[20] Read, J., Pfahringer, B., Holmes, G. and Frank, E. Classifier chains for multi-label classification. Machine learning, 85(3), p.333, 2011.

[21] Tsoumakas, G., Katakis, I. and Vlahavas, I. A review of multi-label classification methods. In Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006) (pp. 99-109), 2006.

[22] Qing-yun Dai, Chun-ping Zhang, Hao Wu. Research of Decision Tree Classification Algorithm in Data Mining.

[23] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[24] Durgesh K. Srivastava, Lekha Bhambhu, Data Classification Using Support Vector Machine.

[25] Radhika Y Sashi M., Atmospheric Temperature Prediction using Support Vector Machines, International Journal of Computer Theory and Engineering, Vol. 1, No. 1, April 2009, 1793-8201 55-58.

[26] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001).

[27] Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions Giovanni Seni and John F. Elder Synthesis Lectures on Data Mining and Knowledge Discovery, 2010, Vol. 2, No. 1 , Pages 1-126 (https://doi.org/10.2200/S00240ED1V01Y200912DMK002).

[28] P. T. Dalvi and N. Vernekar, "Anemia detection using ensemble learning techniques and statistical models," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, pp. 1747-1751.

[29] Baneriee, C., Paul, S. and Ghoshal, M., 2017, December. A Comparative Study of Different ensemble Learning Techniques Using Wisconsin Breast Cancer Dataset. In 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE) (pp. 1-6). IEEE.

[30] P. Matikainen, R. Sukthanka, M. Hebert, Classifier ensemble Recommendation, Proceedings of workshop on Web-Scale vision and Social Media, ECCV 2012, LNCS 7583, pp. 209-218, 2012.

[31] D. Opitz, R. Maclin, Popular ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, Vol. 11, pp. 169-198, 1999.

[32] D. Opitz, R. Maclin, Popular ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, Vol. 11, pp. 169-198, 1999.

[33] Kankanala, P., Das, S. and Pahwa, A., 2014. AdaBoost +: An ensemble Learning Approach for Estimating Weather-Related Outages in Distribu- tion Systems. IEEE Transactions on Power Systems, 29(1), pp.359-367.

[34] Adeena, K. D., and R. Remya. Extraction of relevant dataset for support vector machine training: A comparison. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 222-227. IEEE, 2015.

[35] Isaac, Jackson, and Sandhya Harikumar. Logistic regression within DBMS. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 661-666. IEEE, 2016.

Appendix A

Appendix

A multi layer ensemble learning framework for Learning Disability detection in school-aged children - To be submitted on June 3rd, 2019.
A multi layer ensemble implementation for multi label classification of Learning disabilities - Work in progress