A CASCADED BI-LEVEL FEATURE SELECTION TECHNIQUE FOR PREDICTING STUDENTS' ACADEMIC PERFORMANCE

By

NAME: WOKILI ABDULLAHI
M.TECH/SICT/2018/8343

DEPARTMENT OFCOMPUTER SCIENCE

FEDERAL UNIVERSITY OFTECHNOLOGY

MINNA

NOVEMBER, 2021

A CASCADED BI-LEVEL FEATURE SELECTION TECHNIQUE FOR PREDICTING STUDENTS' ACADEMIC PERFORMANCE

By

NAME: WOKILI ABDULLAHI
M.TECH/SICT/2018/8343

A THESIS REPORT SUBMITTED TO THE POSTGRADUATE SCHOOL FEDERAL UNIVERSITY OFTECHNOLOGY, MINNA, NIGERIA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF TECHNOLOGY IN COMPUTER SCIENCE

NOVEMBER, 2021

ABSTRACT

Educational Data Mining is an important task which is used to detect and explore useful patterns applicable to student learning behavior. Features in educational data are ambiguous which leads to noisy features and the curse of dimensionality problems. These problems can be solved via feature selection. There are existing models for features selection. These models were created using either a single-level embedded, wrapper-based or filter-based methods. However single-level filter-based methods ignore feature dependencies and also ignore the interaction with the classifier. The single-level embedded and wrapper based feature selection methods interact with the classifier, they can only select the optimal subset for a particular classifier. So the features selected by them may be worse for other classifiers. Hence this research proposes a robust a cascade bi-level feature selection technique for student performance prediction that will minimize the limitations of using a single-level technique, hence improve prediction performance. The proposed cascaded bi-level feature selection technique consists of the Relief technique at first-level and the Particle Swarm Optimization (PSO) at the second-level. The proposed technique was evaluated using the Eurostat student performance dataset. In comparison with the performance of the single-level feature selection technique the proposed technique achieved an accuracy of 94.94% for Mathematics dataset which was better than the 93.67% and 92.41% achieved by the single-level PSO and Relief selectors for Mathematics dataset for the binary classification task. The proposed technique also produced better results than previous works based on Eurostat dataset. These results shows that proposed bi-level cascade can effectively predict student performance.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background of the Study

The role of education in the development of any country cannot be over emphasised. This is because of its' impacts on the social, economic and political developments in any society (Adán-Coello & Tobar, 2016). The quality of any nation is directly proportional to the quality of her education system, hence, the ongoing efforts to advance the quality of educational institutions. Academic performance of students in any educational institution is a measure of the institutions efficiency in knowledge delivery (Jembere *et al.*, 2017).

In order to increase learning and teaching in terms of teacher-created along with learner-created content, universities and colleges have begun to include collaborative learning methodologies into their conventional teaching processes (Kirsal & Dimililer, 2018). In addition, the popularity of massive open online courses has skyrocketed in recent years.

Researchers and academicians have been increasingly interested in learning outcomes, which is why scholars have been working hard to identify elements that influence good academic achievement (Ahmad *et al.,* 2015). There are different factors that affect students' performance. They include: intelligence, state of health, motivation, anxiety, suitable learning environment, adequate education infrastructures, family and parental influences, societal influences, institutional influences (David *et al.,* 2015),

In Computer Science, one of the active fields is data mining. Data mining deals with the procedure of mining valuable information from raw data (Hussain *et al.*, 2018). Data mining is critical due to the increasing volume of data and the pressing need to convert this data into valuable information. Data mining is currently utilized in diversity of fields, including banking, advertising, healthcare, architecture, politics, military, and education. It's a well-established discipline for finding meaningful patterns and relationships that allow users to retrieve knowledge and acquire more significance from data (Adejo & Connolly, 2017). With data mining, a search engine could be used to examine vast volumes of information and instantly report meaningful findings without requiring human participation (Rajagopal, 2011). The educational sector is a significant area in which data mining is gaining increasing interest. Data mining is referred to as Educational Data Mining (EDM) in the education field. EDM emphasizes that educational data systems such as course administration systems, online learning systems, registration systems, and application systems provide meaningful knowledge. This mined knowledge can help students at each stage of their studies, like primary to tertiary education (Tuaha*et al.*, 2019). Many user groups are interested in EDM, and these users use the data that EDM has found according to their vision and intent (Romero & Ventura, 2010). For example, educational data's hidden pattern can help educators develop teaching techniques, understand learners, strengthen the learning experience, and use them to boost their learning activities (Amrieh*et al.*, 2016). This secret perception will also help the administration make the necessary decisions to achieve high-quality results (Shah *et al.*, 2019). Educational information is obtained from multiple sources, such as educational institution databases, e-learning services and traditional surveys (Tuaha *et al.*, 2019). The database of today's institutions of higher learning has a wealth of information about their pupils. The amount of information available is growing all the time, yet little action is being made to gain understanding from it. Data mining is an appropriate strategy for managing data in higher education institutions in order to find new facts and knowledge about students. Machine learning, analytical, and visualization approaches are used in data mining to uncover and mine information in a way that humans can understand (Ahmad *et al*, 2015). EDM has a lot of applications, one of which is forecasting student academic achievement. In the educational environment, the analysis and estimation of student performance is an integral aspect. This prediction task foresees the importance of an unknown variable that distinguishes students with outcomes such as pass or failure, grades and marks (Imran *et al.*, 2019).

In recent years, the utilization of cognitive capacity, log activity in learning management systems (LMS), and student demographic factors has been emphasized in predicting student performance. Despite the fact that different research used machine learning approaches like Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Naive Bayes techniques, which differ from the frequently used conventional logistic regression, (Ikbal *et al* 2015) and Hoe *et al.*(2013)used demographic information and pupil score to forecast student achievement.

Romero *et al.* (2013) and Cerezo *et al.*(2016) predicted student grades via logs from web-based platforms like LMS. Login regularity, amount of online sessions, amount of initial post publish/read, percentage of follow-up posts published, amount of content pages visited, and amount of posts read were all predictive variables. Despite the prevalence of "regularity of login" in virtual learning evaluations, a few researchers focus at the quality of involvement rather than the quantity by examining the timing, volumes, and continuity of access, which gave highly effective outcome when incorporated. The most frequently used forecaster variables generated from LMS are the amount of post read, overall average duration online, regularity of accessibility to course materials, and regularity of login.

Sembiring *et al.*(2011) and Fariba (2013) utilized questionnaire approaches to obtain student inherent and behavioral factors that were not freely available in the system for anticipating student achievement. They studied how study habits, personality features, learning methodologies, and motivation variables, as well as psychiatric health, influenced students' educational excellence.

An important procedure in student academic performance prediction is the feature selection step. Automatic or manual selection of features that contribute most to a forecast variable or output is known as feature selection. Irrelevant features in a data set can decrease model correctness and cause the model to train using irrelevant attributes. Educational data are rapidly growing in volume and this large number of features in datasets which leads to a classification problem known as the curse of dimensionality (Sembiring *et al.*, 2011). With the addition of new features for learning classification algorithms, the curse of dimensionality causes an exponential growth in the dimension of the search area, making the data sparser. As the dimension of a dataset increases, the complexity of the classification model is increased, its accuracy is reduced and its computational time Increases. Feature selection or dimensionality reduction is a method adopted by researchers to decrease the dimension of feature space to increase classifiers accuracy and reduce classification complexity (Kumari & Swarnkar, 2011).

Previous works on student performance prediction made use of a single-level filter-based and wrapper-based feature selection technique to select the best features which influences a students' academic performance. However in this research a cascaded bi-level feature selection model is proposed to overcome the limitations of using a single-level feature selection technique for student academic performance prediction.

1.2 Statement of the Research Problem

Education Data Mining (EDM) arose as a result of the increasing growth of educational data, which has confronted researchers with various obstacles in developing more efficient data mining algorithms (Amoo *et al.*, 2018). In EDM the features in educational data are ambiguous which leads to the curse of dimensionality issue. This issue of curse of dimensionality and noisy features can be solved using dimensionality reduction. Dimensionality reduction can be achieved via feature selection. The purpose of attribute selection is to choose a subgroup of attributes that can effectively represent the input data while minimizing the attribute space's complexity and removing irrelevant data. There are existing models for selection of student performance features. However these models were created using either a single-level embedded, wrapper-based or filter-based methods. Filter methods are quick and independent of the classifier, but they neglect feature correlations and the classifier's interaction (Hira & Gillies, 2015). Since embedded and wrapper-based techniques interact with the classifier, they can only choose the best subset for that classifier. As a result, the features they choose may be detrimental to other classifiers (Daud *et al.*, 2017; Francis & Babu, 2019). Filter-based methods are well-suited to dealing with data with a large number of features since they have a high degree of generalization. However, because of the importance of features and the interrelationships between features, a filter-based feature selection method can only rank them. As a result, after performing filter-based selection, particle swarm optimization was utilized to optimally choose a subset of the selected features. Hence, this research work proposes a cascaded bi-level feature selection approach to overcome the drawbacks of single filter-based and wrapper-based selection techniques for student performance prediction.

1.3 Aim and Objectives

The aim of this thesis is to develop a cascaded bi-level feature selection model for predicting students’ academic performance.

The research project's objectives are as follows:

1. Develop a cascaded bi-level feature selection technique for student performance prediction.
2. Select features using the cascaded bi-level feature selection technique developed in (i).
3. Evaluate the performance of the cascaded bi-level feature selection technique in (i).

1.4 Scope of the Study

This research emphases on student academic performance prediction using cascaded bi-level filter and wrapper feature selection approach and evaluating the performance of the selected feature performance using Error-Correcting Output Code (ECOC), ensemble, Decision Tree and K-Nearest Neighbour (KNN). Also only student performance datasets were considered for evaluation of the proposed model without consideration of other types of datasets.

1.5 Significance of the Study

This research would be of benefit to educators as it will help them develop teaching techniques, understand learners, strengthen the learning experience, and use them to boost their learning activities. The secret perception from the mined data will also help the administration make the necessary decisions to achieve high-quality results. Students may be able to use the mined information to have a better picture of how well or poorly they will perform in a course and then take efforts to improve their performance.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Data Mining (DM)

Data mining (DM) is the procedure of discovering extremely important trends from big data collections. It also entails the examination of data for previously unknown or unknown relationships. It's a multidisciplinary skill set that uses statistics, data science, and artificial intelligence to analyze data and predict future outcomes. Insights from data mining are utilized in advertising, fraud protection, and scientific research(Ahmad *et al.*, 2015). The essential step in the knowledge extraction process is data mining.

2.1.1 Data mining (Knowledge Discovery)

The DM stages are carried out with the goal of identifying patterns of useful and interesting data in enormous amounts of data. The main phase of data mining are (Chamatkar & Butey, 2014; Susanto, 2019):

1. **Filtering Data:** Filtering of data is also identified as data cleaning. In DM process, data cleaning is regarded as a first step or a pre-processing step. Data cleaning is the process of cleaning data to ensure that it is free of errors and erroneous information. This stage entails locating and modifying or cleaning erroneous, redundant, unnecessary, incomplete, and "noisy" elements of a data set (Mueen *et al.*, 2016).
2. **Data Integration:** Data integration is a data preparation approach that combines data from numerous disparate data sources into a single, cohesive data storage. Data integration may contain inconsistencies in data, necessitating data cleaning (Stapel *et al.*, 2015).

3. **Selection of variables:** Data selection is the process of retrieving data from a database that is important or relevant to the analysis activity. This is the process of determining the best data kind and source, as well as data collection instruments (Hira & Gillies, 2015).
4. **Data Transformation:** This is the process of changing the structure, architecture, or contents of data. At this stage, data is processed and aggregated into suitable forms for mining by using summary or aggregate methods (Osborne, 2010). Data transformation can help to reduce skewness and the impact of outliers in the data. Centering, scaling, skewness reduction and binning are all transformation methods.
5. **Extracting knowledge:** This is a crucial procedure in which intelligent algorithms are used to extract data patterns.
6. **Pattern Interpretation and evaluation:** Using interestingness measurements, this step identifies the truly interesting patterns that constitute knowledge.
7. **Knowledge Presentation:** Graphics and knowledge representation systems are utilized to show excavated knowledge to users.

2.1.2 Kinds of Mined Data

DM is not limited to a single sort of media or information. Flat files, Data Warehouses, Relational Databases, Transaction Databases, Multimedia Databases, Time-Series Databases, Spatial Databases, and the World Wide Web are all examples of data mining applications (Chamatkar & Butey, 2014).

1. **Flat files:** A flat file, often known as a text database, is a database that stores information in plain text. Flat files, particularly at the research level, are the most popular data source for data mining methods. Flat files are simple text or binary data files with a structure that

the data mining system can recognize. Transactions, time-series data, and experimental measurements can all be found in these files.

2. **Relational Database:** A relational database is made up of a series of tables that hold either entity attribute values or attribute values from entity relationships. Tables have rows and columns, with rows representing tuples and columns representing features. It employs a structure that allows data to be identified and accessed in connection to other data in the database.
3. **Data Warehouse:** A data warehouse is a collection of data from several sources that is meant to be used as a whole under a solitary identical schema. Users can access crucial data from a variety of sources in one place using a data warehouse.
4. **Transactional Database:** This is a set of records with a date and time, an identifier, and a list of items. A transactional database also allows you to roll back transactions on data stores if they are not performed correctly.
5. **Multimedia Database:** The term "multimedia database" refers to a collection of multimedia files that are linked together. Text, photographs, graphic objects, animation sequences, audio, and video are among the primary media data kinds included in multimedia data.
6. **Spatial Database:** A spatial database is one that is designed to store and query data that signifies things in three-dimensional space. Spatial databases can store simple geometric items, such as lines, points, and polygons, in a digital format.
7. **Time series Database:** A time series database is one that is designed to store data that is time-stamped or in series. As the name implies, time series data is simply a collection of

measurements or events that have been tracked over a period of time. For instance, stock market data or activity logs.

8. **World Wide Web:** In terms of heterogeneity and versatility, the World Wide Web is unsurpassed. The data on the World Wide Web is arranged into documents that are linked together. Audio, text, video, raw data, and programs can be included in these documents.

2.1.3 Data Mining (DM) Task

The kinds of patterns that can be found are determined by the data mining activities used. Generally speaking, data mining jobs can be divided into two groups:

1. **Descriptive DM tasks:** These describe the data's overall qualities. Correlation, cross-tabulation, and frequency are all examples of this phrase. Similarities in the data and existing patterns can be found using these methods (Stattner & Collard, 2015).This type of analytics focuses on summarizing and transforming data into useful information for reporting and monitoring (Agyapong *et al.*, 2016; Olson, 2017).
2. **Predictive DM:** Predictive DM is a task that attempts to make predictions based on data inference (Agyapong *et al.*, 2016).The major purpose of this mining is to predict future outcomes rather than present behaviour. It makes use of supervised learning functions to forecast the goal value (Olson, 2017; Stattner & Collard, 2015).

2.1.4 Classification of Data Mining

There are a plethora of data mining technologies on the market. Others are more customizable and broad, while some are specialized to a certain data source or have restricted data mining skills. A variety of criteria can be used to classify data mining technologies, including the ones listed below (Deshpande & Thakare, 2010):

1. **Classification based on mining methods used:** DM solutions use and offer a variety of methodologies. Machine learning, genetic algorithms, neural networks, visualization, analytics, database, and data warehousing are among the data analysis methods used in this classification, which divides data mining methods into groups. The level of user involvement in the data analysis can be categorized as query-driven frameworks, interactive explorative frameworks, or automation technologies.
2. **Classification based on the types of knowledge mined:** This grouping separates DM systems into categories depending on the type of knowledge gained or DM capabilities like description, discrimination, correlation, categorization, and clustering.
3. **Classification based on the types of databases mined:** This categorization divides DM systems into categories based on the data type they process, such as multimedia data, time-series data, text data, and data from the World Wide Web.
4. **Classification based on the data model drawn on:** DM technologies are divided into four classes under this classification: object-oriented databases, relational databases, data warehouses, and transactional databases (Colonna, 2013).

2.2 Educational Data Mining

In many educational institutions, the quantity of data collected and stored had grown to the point where educational data analysis could no longer be done manually. EDM is a new discipline that arose from the use of data mining techniques on educational data. EDM is a subset of Learning Analytics that employs machine learning to categorize academic data sets at various levels (Pandey & Taruna, 2016; Rodrigues *et al.*, 2018; Zacharis, 2016; Zaffar*et al.*, 2018).The goal of EDM is to create and use algorithms to improve educational outcomes and to explain educational practices for future decision-making. Data mining techniques are utilized to mine knowledge from educational data and investigate the characteristics that can help to improve performance. Learning, in reality, began in the classroom and was based on behavioral, cognitive, and constructivist paradigms of the time. The ultimate purpose of EDMs is to understand how students learn and to discover aspects of learning and education that can be improved (Silva & Fonseca, 2017).

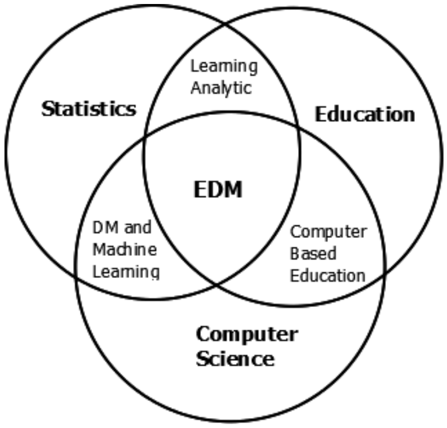

Figure.2.1 Key areas involved in Educational data mining (Manjarres *et al.*, 2018)

Learning analytics is the graph's most closely associated field area, and it can be defined as the assessment, collection, interpretation, and presentation of data about students and their surroundings with the goal of improving knowledge and learning, as well as the surroundings in which they occur. As a result, EDM may share a number of traits with the surrounding disciplines (Algarni, 2016; Manjarres *et al.*, 2018)

The EDM has risen in prominence in recent times, owing to the benefit that this discipline brings to all players participating in the learning process. This is especially true now, when people's learning habits and attendance at schools and colleges have altered considerably, not to mention the importance of digital learning materials and social media (Mohamad & Tasir, 2013). EDM can be used to classify and forecast student achievement, withdrawals, and instructor effectiveness. It can help teachers track academic success in order to improve the teaching process, as well as students choose courses and manage their education.

2.2.1 Educational Data Mining Process

EDM has four main phases, which are (Zorić, 2020):

1. **Problem Definition:** This is the first step in converting a particular problem into a data mining challenge. The project purpose and objectives, as well as the primary research topics, are developed during this period.
2. **Data Preparation:** This is the second part of the process. It can consume up to 80% of all analysis time. In data mining, data quality is a crucial issue (Blake & Mangiameli, 2011). Source data must be located, cleansed, and formatted in a pre-specified format during this step.
3. **Modeling and Evaluation:** This is the third phase of the project. The parameters are set to their optimal values in this phase, and various modeling techniques are chosen and employed.
4. **Deployment:** This is the final stage. This is the stage when the data mining results are organized and displayed in graphs and reports.

It's worth noting that data mining is an iterative process, which means it doesn't end once a solution is implemented. It could just be a fresh input for a data mining algorithm.

2.2.2 EDM Methods

EDM employs a variety of approaches, algorithms, and procedures. Classification, prediction, grouping, and association are the most common applications. Neural networks, regression analysis, decision trees, and cluster analysis are some of the most used data mining approaches (Zorić, 2020).

2.2.2.1 Classification

Classification is a DM procedure that assigns objects in a set to chosen groups. It facilitates in the examination of data and the forecasting of results. The aim of classification is to appropriately foretell the target group for every example in the data. The classifier training algorithm determines the set of parameters required for classification using pre-classified instances (Oracle, 2020). This technique is commonly used in the educational industry to classify pupils based on age, grades, gender, knowledge, academic qualifications, motivation, behavior, demographic, or regional characteristics. Examples of classification algorithms are: Naïve Bayes, Artificial Neural Network (ANN), Random forest, Decision Tree, and Support Vector Machine (SVM).

1. **Artificial Neural Network (ANN):** ANN are a class of computing systems inspired by the human central nervous system that are intended to recognize complicated forms and solve prediction problems without the use of programming (Kalejaye *et al.*, 2015). They recognize distinctive traits in the examples they process automatically. Artificial neurons are the nodes that make up neural networks. Every link can send a signal from one artificial neuron to another. A signal is a numerical value. The weights of artificial neurons and connections adjust during the learning process. Input, output, and hidden layers are the three layers that make up neurons. Signals pass from the input layer to the output layer via hidden layers, undergoing various transformations on their inputs. The ability to learn and model non-linear and complicated interactions is its most significant capability (Amoo *et al.*, 2018).

2. **Decision Tree:** A decision tree is a decision-making method that classifies data using a tree-shaped graph or model. It is a method of learning that is supervised. Each inner node represents an attribute check, each branch a test output, and each leaf node a class, which is a choice reached when all attributes have been computed. The categorization rules follow the routes from the root to the leaf (Kolo *et al.*, 2015). Their main benefit is their consistency and ease of interpretation (Feng, 2019).

2.2.2.2 Clustering

Clustering is a method of categorizing data such that items in same category are very similar while items in separate clusters are very distinct. Clustering can alternatively be defined as the transformation of a collection of abstract items into classes of related things. Clustering analysis is a technique for grouping comparable data into previously undefined clusters. It's beneficial for identifying homogeneous groups that can be utilized as input for other models during the data preprocessing step. Cluster analysis, like classification, can be used to look at the similarities and differences between students, courses, and teachers.

There are various types of clustering methods:

1. **Hierarchical Clustering:** This is a clustering algorithm that begins with a top-to-bottom hierarchy of clusters and works its way down. The clusters are then created by breaking down the data objects according to this hierarchy. This method uses either a top-down or bottom-up strategy to cluster creation, depending on the direction of advancement (Romalt & Kumar, 2020). These are the Divisive Approach (from the top down) and the Agglomerative Approach (from the bottom up). Algorithm for agglomerative clustering: Begin with independent clusters of points and combine the most alike or closest pair of

clusters at each phase. Algorithm for clustering by division: Begin with a single, all-inclusive cluster and separate it into singleton clusters of separate points at each stage.

2. **Partitioning Clustering:** Based on the qualities and similarities of the data, this clustering process divides the information into several categories. The number of clusters that must be formed for the clustering algorithms must be specified by the data analysts. The k-means and k-medoid clustering techniques are examples of this clustering.
3. **Density-based Clustering:** Density-Based is a clustering processes for recognizing unique groups in data. It is predicated on the idea that in a data space, a cluster is a continuous area of high point density that is disconnected from other clusters by continuous regions of low point density. Clusters of any shape, with no limit on cluster size, can be created using density-based methods.
4. **Fuzzy Clustering:** Fuzzy clustering extends partition clustering approaches (like k-means and medoid) by allowing an individual to be categorized partially into multiple clusters. The membership of all clusters is distributed in fuzzy clustering.

2.2.2.4 Association

Association is a data mining approach for determining the likelihood of elements in a collection occurring together. Association rules are the relationships between co-occurring things. When analyzing sales transactions, association rules are widely utilized. This type of information is useful for sales promotions, direct marketing, catalogue design, cross-sell marketing, and identifying industry trends. This strategy can be used to provide new courses or to open new institutions if certain rules are followed.

2.3 Feature Selection

The process of limiting the amount of input variables when developing a predictive model is called feature selection (Naheed *et al.,* 2020).It is preferable to reduce the quantity of input variables in order to lower the computational cost of modeling and, in some cases, to improve the model's performance. A set of features from the original array of features is picked for feature selection depending on their repetition and relevancy (Venkatesh & Anuradha, 2019).The presence of worthless information does not lead to the deformation of irrelevant and redundant features; instead, a feature is unimportant if it lacks a statistical relationship with other features. Any feature may be unimportant when used alone, but it becomes relevant when paired with other features. Filter, Wrapper, and Embedded Methods are 3 distinct types of feature selection techniques that differ in how they interact with the learning model (Meyer-Baese & Schmid, 2014).

2.3.1 Filter-based Feature Selection

Filter-based techniques choose variables regardless of the model. They are solely dependent on general characteristics such as the correlation with the variable to be predicted (Kumari & Swarnkar, 2011). The least interesting variables are suppressed via filtering procedures. Statistical metrics are used to choose features in the Filter technique. Filter-based methods are unaffected by the learning method and takes less time to compute (Romalt & Kumar, 2020). The main drawback of filter-based method is that it ignores interaction with classifiers and feature dependencies. Information gain, chi-square test, Relief, Fisher score, correlation coefficient, and variance threshold are examples of statistical metrics used to understand the value of the features (Yildirim, 2015).

2.4.2 Wrapper-based Feature Selection

The performance of the Wrapper technique is determined by the classifier. The best collection of features is picked based on the classifier's results (Romalt & Kumar, 2020). Wrapper approaches evaluate the "utility" of features based on the performance of the classifier. Because of the frequent learning procedures and cross-validation, wrapper methods are more computationally intensive than filter methods (Naheed et al., 2020; Wang et al., 2014). These wrapper methods, on the other hand, are more precise than the filter method. Advantages of Wrapper-based methods are: Simple, Interacts with classifier, feature dependencies, and good grouping accuracy. Examples of wrapped-based techniques are: Recursive feature elimination, Sequential feature selection algorithms, Particle Swarm Optimization algorithm and Genetic algorithms (Kumari & Swarnkar, 2011).

2.3.3 Embedded-based Feature Selection

In Embedded methods the feature selection technique is implemented as part of the learning process (Imani et al., 2013). Filter and wrapper methods are combined in embedded methods. It is implemented using algorithms with their own feature selection methods. A learning algorithm uses its own variable selection mechanism to simultaneously perform feature selection and classification/regression. Tree algorithms such as Random Forest and Extra Tree are the most common embedded techniques. It performs better than the filter and wrapper-based models because it makes a group decision a new characteristic is selected and the sample set is subdivided into smaller subsets in each recursive phase. As the number of child nodes in the same class grows, the features in a subset become more informative. Advantages of embedded techniques are: They takes less time to compute than wrapper methods and they are much less prone to over-fitting. This strategy, however, has the disadvantage of being particular to a learning model.

Table.2.1 Comparison of filter-based, wrapper-based and embedded feature selection techniques

	Filter	**Wrapper**	**Embedded**
Selection criteria	Features are chosen based on statistical metrics rather than a specific machine learning algorithm	Features are evaluated using a machine learning technique to find the best features.	Inserts features as the model is being built. Each iteration of the model training phase is used to choose features.
Time complexity	In terms of time complexity, it's far faster than wrapper techniques	For a data set with many features, the computation time is long.	In terms of time complexity, it falls in between the Filter and Wrapper techniques.
Over-fitting	Less prone to over-fitting	High chances of over-fitting because it involves training of machine learning models with different combination of features	Generally used to reduce over-fitting by penalizing the coefficients of a model being too large
Examples	Correlation, ANOVA, Relief, Information Gain, Mutual information and Chi-square test	Backward elimination, forward selection, and stepwise selection	Random Forest, LASSO, Ridge Regression, and Elastic Net.

2.4 Student Performance Prediction Techniques

Students' viability of progress is essential to predict student performance. The significance of predicting student performance has led researchers to become more and more interested in this field. Therefore, various researches have been published to predict students' performance.

A classification model for the prediction of student performance was built by Salal*et al.* (2019) using a dataset of 649 examples with 33 attributes obtained from 2 Portuguese high schools: Gabriel Pereira and Mousinho da Silveira High School. The dataset includes features, such as academic, demographic and social attributes of students. The classification target class ranged from 0 to 20, rendering the classification process extremely difficult as there were only 649 examples to be trained and assessed. Based on the initial class ranges, the target class was reduced to 6 categories due to this complexity. In WEKA software, the correlation assessment, gain ratio, and information gain were used as evaluation techniques, and these new target groups were used to pick attributes. After obtaining the outcome of the attribute selection algorithms' outcome, ten different attributes were selected, which were checked to influence the prediction outcome significantly. Eight classifiers, namely the Naïve Bayes, Random Tree, REP Tree, Decision Tree, Simple Logistics, One R, and Zero R, were fed with these selected classification attributes. One R was identified to have performed better with an accuracy of 76.7334% compared to the other seven classifiers with lower accuracy value. A comparative overview of a relatively large number of classifiers was provided by the study, offering an in-depth understanding of an extensive range of techniques. In this paper, each of the methods' performance was evaluated based only on accuracy without considering other performance metrics, which could say a lot about the suitability of a technique. The classification accuracy achieved was also low, unlike similar works that used the same dataset.

Iyanda *et al.* (2018) conducted a comparison between two Neural Networks (NN), (generalized regression NN and multilayer perceptron) to determine the paramount model for student academic achievement prediction founded on only the educational feature of the student. The dataset used was collected from Computer Science and Engineering Department of the Awolowo

Nigeria University of Obafemi. The data collected constitutes the academic record of learners (raw scores for each course taken) as the input variable, and the accompanying GPA as the output parameter. Using mean square error, receiver operating features, and accuracy, the two NN models' efficiency was evaluated. The generalized regression NN proved to perform better with an accuracy of 95% than the multilayer perceptron. However, without considering how demographic, social, and behavioural attributes could affect a student's output, this research used only student academic attributes for prediction.

Olalekan *et al.* (2020) adapted Bayes' theorem and ANN to construct a predictive model for students' graduation probability at a tertiary institution. Four variables were used for prediction: Unified Tertiary Matriculation Test, Number of Sessions at the high school level, Grade Points at the high school level and Entry Mode. The data used was gathered from the Computer Science School, Federal Polytechnic, Ile-Oluji, in Ondo State, Nigeria. The data were composed of forty-four examples with five attributes. The study concludes that the ANN has a 79.31% higher performance accuracy than the 77.14% obtained by the Bayes classification model. The ANN precision improved as the hidden layers increased. As compared to other previous works, the overall accuracy in this study was low because of the small size of data used. Expanding the data size would help enhance the accuracy of the classification of the model.

Zacharis (2016) utilized Model data to forecast student achievement in a course focusing on 4 learning activities: collaborative content production via wiki, email communication, content interaction assessed by files read, and self-evaluation via online quizzes. In order to forecast student performance in a blended learning environment, a model based on Multi-Layer Perceptron Neural Network was built. The proposed model was found to have an accuracy rate of 98.3%. The quantity of messages posted by students and the contributions made by students in team content creation initiatives were the most powerful predictors of course performance. However, only the accuracy was used as a performance measure for the proposed method. Also the data used was limited to just two courses in Computer engineering and Mechanical Engineering Department. Larger and more diverse samples are required for robust validation of the proposed work.

Magbag and Raga (2020) focused on building a model to predict first-year students' academic success in tertiary education. This research aimed to allow early intervention to help students stay on course and reduce non-continuance. The data utilized in this paper were obtained from three higher education institutions in Central Luzon, primarily in the cities of Angeles, San Fernando and Olongapo. The study subjects included first-year students from 8 academic departments from 2018-2019; Arts and Sciences, Engineering and Architecture, Computer Studies, Criminology, Education, Hospitality and Tourism, Business and Accountancy, Nursing and Allied Medical Sciences. The dataset was composed of 4,762 examples. The dataset was pre-processed, and missing values were deleted, leaving 3,466 available samples. Using Correlation-based Feature Selection, Gain Ratio and Information Gain for feature rating, feature selection was carried out. Using these selected features, the NN and logistic regression models were trained and evaluated. In comparison with similar works, the scale of the dataset used rendered the scheme more robust. However, the accuracy of 76% achieved in this analysis is low.

Anuradha and Velmurugan (2015) reported a data mining study of final year outcomes of undergraduate graduate degree students, which was conducted at three private colleges in Tamil Nadu, India. The fundamental goal of this project is to use the J48 decision tree, Naive Bayes (NB), k-Nearest Neighbors (K-NN), OneR, and JRip classifiers to predict student performance in end-of-semester exams. These classifiers' outputs were compared. The overall accuracy of the tested classifiers was found to be greater than 60%.Furthermore, the classification accuracy for the various classes reveals that the distinction class has the worst forecasts and the first class has the best. For the Distinction, the JRip produces the best classification accuracy. The classification of students based on attributes demonstrates that the prediction rates differ between classification methods. It also demonstrates that certain data qualities have been discovered to have an impact on the classification process. The study was able to provide insight on various classification algorithms which would aid researcher decide on the best method to use for student performance evaluation. However only 2 performances metric was used to perform the comparison. More performance metric can be used to provide a robust and more reliable evaluation among the classifiers.

The data for Hussain *et al.* (2018)study came from three separate colleges in Assam, India: Duliajan College, Doomdooma College, and Digboi College. Data on twenty-four attributes was first obtained. The name of the student's attribute was removed from the list of attributes because it has no meaning. After cleansing the data, twenty-two qualities were chosen. The J48, PART, Random Forest, and Bayes Network Classifiers were employed as classification algorithms. The Apriori technique was used to find some of the best rules in the data set. The proposed method achieved a good accuracy of roughly 99% thanks to the usage of the Apriori algorithm to determine some of the best rules to employ and the feature selection strategy.

Daud *et al.* (2017)collected scholarship holding students' data from several universities in Pakistan to explore socio-economic, educational, and demographic feature sets. To forecast whether a student will be able to complete his degree, learning analytics, discriminative, and generative classification models are used. The importance of each indicator in predicting a student's performance was assessed. Bayes Network, NB, SVM, C4.5, and Classification and Regression Tree (CART) were used in the experiment to examine the impact of each feature on predicting student achievement. Due to the use of family expenditures and students' personal information feature sets, the proposed method produced a good F-measure of up to 86.7%.

Ünal, (2020) used feature selection to forecast student success. On educational datasets, decision trees, random forests, and Naive Bayes were used to predict students' final grades. In this student two experiments was conducted. The first experiment deals with training the classifiers without feature selection. And the second experiment deals with training the classifiers after feature selection. In the second experiment wrapper feature selection technique was used to select the most relevant feature set, while the irrelevant features were removed. The second experimentation produced an improved accuracy due to the applied feature selection than the first experiment without feature selection. For instance the accuracy of Naïve Bayes improved from 67.80% in the first experiment to 74.88% in the second experiment. The EuroStat dataset from secondary education of two Portuguese schools were used. This issue with the feature selection technique used in this study is that they are classifier dependent. That is a set of features selected by a particular classifier and works well for that classifiers, those not mean those set of features will also perform well for other classifiers/models.

Salal et al., (2019) presented a model for student performance classification based on the Eurostat Portuguese data set consisting of 33 attributes and 649 instances. Nine classifiers namely: ZeroR, Naïve Bayes, Random Forest, Random Tree, Decision Tree (J48), REPTree, Simple Logistic, JRip, and OneR were utilized in this study. In this study feature selection was performance using filter-based technique. All the nine classifiers had performance improvement when trained with the selected features. For instance the decision tree classifier with an accuracy of 67.79% when trained with all the feature attained 76.27% when trained with the selected features. This shows that student’s attributes affect the student performance. Filter-based feature selection strategies overlook feature dependencies and the interaction with the classifier, which is a flaw in the suggested system.

A new prediction algorithm to determine students' progress in academia using a hybrid (classification and clustering) Francis and Babu (2019) proposed a data mining technique. The analysis used information from X-API education obtained from the kaggle repository consisting of 16 attributes with 480 instances. The dataset characteristics are demographic, academic, behavioural, and additional attributes (parent school satisfaction, student absentee days and parent response survey). Using classifiers such as SVM, Naïve Bayes, Decision tree, and NN, feature selection experiments were performed. The selection of attributes was based on the accuracy provided by each classifier after the demographic, academic, behavioural and extra attributes were trained separately. Compared to using behavioural characteristics alone, additional features alone, educational features alone or demographic features alone the academic + behavioural + extra features provided a higher classification accuracy. These selected features were used as input for K-mean clustering and the majority vote approach. When applied to the dataset’s academic, behavioural, and additional features, the proposed hybrid approach achieved an accuracy of 75.47%. However, related works using the X-API education dataset achieved greater accuracy of approximately 82% compared to this study.

Adejo and Connolly (2018) set out to empirically study and liken the usage of diverse data sources, various classifiers, and ensembles of classifiers in forecasting student academic achievement. The study contrasted the performance and competence of ensemble approaches that used several data sources with base classifiers that only used a single data source in their research. A total of 141 students registered at the University of the West of Scotland had their information taken from the organization's databases and gathered via a questionnaire. The study used 3existingDM classifiers, namely decision tree, ANN, and SVM, to model 3 data sources: student record system, learning management system, and survey. Furthermore, the ensembles of these basis classifiers were utilized to predict student achievement, and the 7 distinct models produced were compared using 6 different assessment measures. The results revealed that combining numerous data sources with heterogeneous ensemble methodologies to predict student performance is very efficient and accurate, as well as assisting in the right identification of students at risk of abrasion. However, when compared to other relevant study, this paper has a low accuracy of 81.67%.

The study by Mala *et al.* (2018) tries to forecast student accomplishment based on the importance of subjects that will be examined on the final national exam. As a method of foresting, extreme learning was applied. The extreme learning approach works on the same principles as the ANN method in general. There are three layers in total: input, concealed, and output. Extreme learning generates strong generalization performance by randomly assigning the input parameters. This study achieved a low RMSE of 0.314 by employing a 20-20-1 network topology. The data set employed in this work has a restriction in that it only contains one year of academic records for students. Using more than one year of academic records for training and testing would increase the model's robustness.

ANN was utilized in a study by Lau *et al.* (2019) to assess and forecast the students' CGPA via data from their socioeconomic upbringing and entrance examination grades of undergraduate students from a Chinese institution. To assess the effectiveness of ANN, calculations of Mean Square Error (MSE), regression analysis, error histogram, and confusion matrix are used to determine that ANN's performance in preventing over-fitting concerns is suitable. Overall, the ANN has an AUC value of 0.86 and a prediction accuracy of 84.8%.When it comes to classifying students based on gender, the suggested ANN method performs badly due to high False Negative rates, which is probable due to the large imbalance proportion of the two different types of sample.

Abu-Zohair (2019) suggested a method for demonstrating the feasibility of training and modeling a small data set size and developing a prediction model with a plausible accuracy rate using a short dataset size. Using visualization and clustering methods, this study looks at the feasibility of identifying the main indicators in the tiny data set that was used to create the prediction model. Multiple machine learning techniques were used to analyze the best indications in order to find the most accurate model. A British university in Dubai provided the data set for this study. The results demonstrated the capacity of the clustering algorithm to discover key indicators in small datasets among the techniques chosen. Machine learning techniques like Linear Discriminant Analysis (LDA), SVM, and Multiple Perceptron ANN require numeric types of characteristics, hence data encoding was used to convert all data types to numerical data types. The proposed system had a 79% accuracy rate. The study's weakness is that it solely used students' administrative records to create classification models, ignoring other variables such as attendance and instructor course delivery that could have an impact on students' learning results.

Table 2. 2 Summary of Review of Related Works

S/No	Authors Name	Problems Addressed	Techniques used	List of Features Used	Source of Data	Performance Evaluation Parameters	Weakness of the Study	Strength of the Study
1	Anuradha and Velmurugan (2015)	A comparative examination of classification algorithms in the prediction of students' performance was offered.	J48 decision tree, Naive Bayes Classifier, k-Nearest Neighbors, OneR and JRip	Demographic and pre-collegiate	Three private colleges in Tamil Nadu state of India	True Positive Rate, Precision	Only 2 performance metric was used to perform the comparison. However more performance metric can be used to provide a robust and more reliable evaluation.	The study was able to provide insight on various classification algorithms which would aid researcher decide on the best method to use for student performance evaluation.
2	Ahmad *et al.* (2015)	To forecast first-year undergraduate students' academic success inComputer Science Department	Decision Tree, Naïve Bayes, and Rule Based classification	Demographic, Academic	Faculty of Informatics and Computing, Universiti Sultan ZainalAbidin , Terengganu, Malaysia.	Accuracy	The modest amount of the data in this study is a restriction due to incomplete and missing values in the obtained data.	There was a correlation between the independent and dependent variables in this study, and this pattern can be utilized to forecast student achievement
3	Kalejaye *et al.* (2015)	To tackle the problem of academic failure by seeking ways to make the process more effective, efficient and reliable based on prediction of students' academic performance in a University System	ANN	Academic	Department of Computer and Information Science, Tai Solarin University of Education in Ogun State	Accuracy	The dataset used was small and limited to a specific department which makes the model less robust	This study has a 91.7% prediction accuracy due to the use of a feed forward ANN that regulates the network synaptic weights and neuron biases to lessen the mean square error between the actual and predicted outputs.
4	Adán-Coello and Tobar (2016)	Student performance prediction	Matrix Factorization, Biased Matrix Factorization, user-kNN,	Academic	Assistant system (an online tutoring system created in 2004 using	Accuracy, Precision, Recall	The study achieved very low recall and precision for incorrect answers	The system was able to achieve high recall and precision for correct answers particularly for the

			item-kNN, random baseline methods, global average baseline methods, Slope One and Bipolar Slope On		8th grade Massachusetts Comprehensive Assessment System test items from 1998 to 2007). Pittsburgh Science of Learning Center (PSLC) Data Shop			assistments_2009_2010 dataset.
5	Amrieh *et al.* (2016)	Using Ensemble Methods to Mine Educational Data to forecast Student Academic achievement	ANN, Naïve Bayesian, Decision tree and Ensemble	Demographic, Academic, Behavioral	Learning Management System (LMS) called Kalboard 360	Accuracy, Precision, Recall and F-measure	The precision of a student's predictive model based on behavioral characteristics When compared to the results of deleting these features, there was a 22.1% improvement, although these features were not investigated to see why they effect prediction accuracy more than the other features.	The implementation of the ensemble technique resulted in an accurate evaluation of the features that may have an impact on the students' performance level, as well as an improvement of over 25.8% in the accuracy of the student's prediction model.
6	Mueen *et al.* (2016)	To forecast and analyze students' academic achievement based on their academic record and participation in forums.	Naïve Bayes, Neural Network, and Decision Tree	Demographic, Academic and Forum	Learning Management System (LMS)	Accuracy, Precision, Recall, Specificity	The data used was limited to only two undergraduate courses over a period of one year. Which makes the proposed system not to be robust	Unbalanced data is an issue in student performance prediction; nevertheless, the SMOTE technique was employed to solve the problem and offer an unbiased conclusion. The data set was also studied to uncover factors that lead students to lose their academic standing owing to bad academic

								performance, and it was discovered that poor academic performance was caused by a lack of engagement in an online discussion forum.
7	Zacharias (2016)	To forecast student success in a blended learning setting	Artificial Neural Network (Multilayer Perceptron)	Academic	A learning Management Course called Moodle of Computer engineering and mechanical Engineering Department of a technology school not mentioned	Accuracy	The data used was limited to just two courses in Computer engineering and mechanical Engineering Department, however larger and more diverse samples are required for robust validation of the proposed work	The proposed model had a high accuracy rate of 98.3%. The quantity of messages posted by students and the contributions made by students in team content creation initiatives were the most powerful indicators of course performance, according to this study.
8	Badr *et al.* (2016)	Students' performance in a programming course can be predicted based on their grades in other areas.	Classification Based on Associations	Academic	Mathematics students who graduated from King Saud University (KSU) between 2008 and 2014	Accuracy	The proposed model achieved a very low accuracy of the 67.3%	The model was able to forecast students' achievement in programming course founded on their performance in English and mathematics subjects. This shows that the model can be adopted for other science courses
9	Singh and Kaur (2016)	To create a model for predicting a student's GPA based on their socioeconomic circumstances and previous academic success.	REPTree and J48	Social, Academic	Department of Computer Engineering, Punjabi University, Patiala	Accuracy, True positive rate, recall, precision	The achieved accuracy of 67.37% is low as compared to related works	This study is considers the third-semester GPA of a student instead of the eight-semester considered by most literatures and this has given this study an edge as it observed that some students drop out after the first year and some students change their stream after second

								semester hence this study creates an early prediction of weak students in academics which will help the authorities to make the necessary decisions for improving students' performance.
10	Umer *et al.* (2017)	In learning analytics, process mining can be used to predict student academic success.	logistic regression, Naïve Bayes, random forest and KNN	Academic, Demographic, Behavior	Coursera for course "Principles of Economics "offered in Summer 2014.	F-Score and Area Under Curve (AUC)	The missing values and small size of the data are the study's limitations.	The application of process mining to enrich the features is the significance of this study.
11	Jembere *et al.* (2017)	To forecast a student's mark for a module based on the student's previous performance in similar courses.	Singular Value Decomposition (Matrix Factorization technique)	Academic	College of Agriculture, Engineering and Science at University of KwaZulu-Natal	Root Mean Square Error (RMSE)	There were only 501 students in the dataset. This is substantially less data than is often utilized in recommender systems. The number of students was insufficient to identify latent variables that could explain variation in student grades concretely.	The study achieved a low RMSE as the matrix factorization method used is robust with sparse data which makes it suitable for this problem domain.
12	Ermiyas and Gobena (2017)		ANN, Naive Bayesian and SVM	Academic	Wolkite university registries office for college of computing and informatics	Accuracy and Execution time	Based on the case study used only one department out of 32 departments was selected for the study. Inclusion of other courses in the model, on the other hand, may be able to provide new perspectives and assist the university to obtain a better knowledge of students' academic achievement.	The study was able to evaluate the 3 techniques not just based on the accuracy but also based on the execution time which serves as a good performance metric to for evaluating the techniques. Also the study was able to attain a high accuracy with a low execution time.
13	Daud *et al.*	To investigate	SVM, C4.5,	Socio-	Different	Precision,	In addition, just a tiny	Since this proposed

	(2017)	feature sets used for students' academic performance prediction	Classification and Regression Tree (CART), Bayes Network, Naive Bayes	economic, academic and demographic	universities located in Pakistan	Recall and F-measure	amount of data was employed in this investigation. Increasing the quantity of datasets, on the other hand, can improve the suggested system's capability.	strategy exploited family spending and student personal information, it provided a good F-measure of up to 86.7%.
14	Salvi *et al.* (2018)	Prediction of student final period grade	Fuzzy Logic	Academic	Not Mentioned	Not Mentioned	This study does not specify the source of the data or the number of attributes it used. Furthermore, no performance metric was utilized to evaluate the suggested method's performance, making the system's performance unverifiable.	Most earlier strategies relied solely on past numeric data for prediction, resulting in intricate predicting procedures with difficult-to-understand findings. This paper presents a method for predicting a student's final period grade based on widely available and clearly interpretable qualities connected to the student's previous academic records and aspects of regular study behavior.
15	Adejo and Connolly (2018)	To evaluate and compare the efficacy of various data sources, classifiers, and ensembles of classifiers in forecasting student academic achievement.	Decision tree, ANN, ensemble and SVM	Demographic, Psychological, Academic	University of the West of Scotland	precision, recall, F-measures, Classification error and Root Mean Square Error (RMSE).	Three features were utilized in this study, however more features or combination of features can be used which might improve the system accuracy. As a low accuracy of 81.67% was achieved in this work as compared to other related work	The findings suggest that combining numerous data sources with heterogeneous ensemble methodologies to predict student performance is very efficient and accurate, as well as assisting in the proper identification of students at risk of attrition.
16	Amoo *et al.* (2018)	Predicting and analyzing secondary school pupils' academic achievement.	Feed-Forward neural network	Cognitive and Psychological	Methodist Grammar school, Emmanuel College,	Accuracy	Despite the fact that ANN has a high level of prediction accuracy in nonlinear occurrences, the model does not	The suggested model has a 90% accuracy rate, demonstrating its potential efficacy as a prediction model, a

					Tafseer model college and Community grammar school all in Ibadan North Local government area of Oyo state		simply allow for the identification of how predictor variables are related to each other in the justification of academic outcomes.	clustering instrument, and a selection criterion for individuals seeking admission to the university.
17	Iyanda *et al.* (2018)	Identification of the best ANN (Multi-layer Perceptron and Generalized Regression) model for forecasting students' academic achievement based on a single measure.	Multilayer Perceptron and Generalized Regression ANN	Academic	Computer Science and Engineering Department of ObafemiAwolowo University, Nigeria	Mean Square Error, Accuracy and Receiver Operating Characteristics (ROC)	Although the number of instances employed was limited, increasing the number of datasets can help neural networks better grasp the complicated behavior of the system and adjust the learning parameters to produce better results.	The methodologies used have various advantages, such as generalisation, efficiency, and simplicity, which make them excellent for forecasting academic achievement in kids. The technique's effectiveness was demonstrated when the Generalized Regression Neural Network obtained a 95% accuracy rate.
18	Hussain *et al.* (2018)	Students' academic achievement is evaluated based on personal and academic data.	J48, PART, Random Forest and Bayes Network	socio-economic, demographic, academic	Duliajan College, Doomdooma College and Digboi College of Assam, India.	Sensitivity, Precision, F-score, Accuracy, Mean Absolute Error, Relative Absolute Error, Root Relative Squared Error and Root Mean Square Error	Because the types of attributes used were limited and the examples used were limited, data may be expanded to include some of the students' extracurricular activities and technical skills, which would increase the system's robustness.	The proposed method achieved a good accuracy of roughly 99% thanks to the usage of the Apriori algorithm to determine some of the best rules to employ and the feature selection strategy.
19	Mala *et al.* (2018)	Predicting student accomplishment	Extreme Learning	Academic	SMAN 1 BatuanSumene	Root Mean Square Error	The dataset utilized for training and testing is	The Extreme Learning Machine used generates

		based on the importance of subjects that will be examined on the final national test.	Machine		p, Indonesian	(RMSE) and Execution time	only one year's worth of academic records of students. Using more than one year's worth of academic records might increase the model's robustness.	good generalization performance.
20	Salal *et al.* (2019)	Comprehensive analysis of student performance dataset via a Classification model.	Naive Bayes, Decision Tree (J48), JRip, OneR, Simple Logistic, Random Forest, Random Tree, REPTree, and ZeroR.	Academic, Demographic, Social	2 Portuguese secondary school namely: Gabriel Pereira and Mousinho da Silveira	Accuracy	The performance of each of the techniques were evaluated based only on accuracy without consideration of other performance metric which could say a lot on a technique suitability.	The study was able to provide a comparative analysis of quite a good number of classifiers which give a better insight on a broad range of techniques
21	Almasri *et al.* (2019)	Statistical analytic approaches were used to examine selected attributes and their influence on performance value, as well as to forecast student performance using an ensemble meta-based tree.	Ensemble Meta-Based Tree	Academic and Demographic	Unspecified Registration office.	Accuracy, Precision, Recall, F-measure and Area Under the Curve	Only a few features were employed; it is advised that more features be included, such as investigating how students' use of social media affects their performance.	The proposed ensemble bagging technique used shown a significant improvement of 98.5% accuracy as compared to other existing works
22	Francis and Babu (2019)	Prediction of student performance in academia	K-means, SVM, Naives Bayes, Decision Tree, Neural Network	Demographic, Academic, Behaviour, Extra	Various higher educational institutions in Kerala, India	precision, recall, f-score and accuracy	The student data set had a limited number of features to consider, and the accuracy of 75.5% is low	The utilized hybrid model based on a combination of clustering and classification improved the prediction accuracy of the system
23	Lau *et al.* (2019)	Evaluation and prediction of the students' CGPA	Artificial Neural Network (ANN)	socio-economic, Academic	undergraduate students from a Chinese university	Mean Square Error (MSE), regression analysis, and	The suggested Artificial Neural Network (ANN) system works badly in classifying students by	The challenge with the modeled Artificial Neural Network (ANN) method is that it is susceptible to

						error histogram	gender, with significant False Negative rates as a result, which is likely owing to the high imbalance ratio of two types of sample data (students' gender) used.	over-fitting and time-consuming, but this study was able to overcome these drawbacks by reducing the sphere of competence for both traditional statistics and ANN analysis.
24	VeeraManickam *et al.* (2019)	Prediction of final semester marks of students based on map-reduced cluster architecture	Cumulative dragonfly based neural network	Academic	Different colleges whose names were not mentioned	Mean Square Error, Root Mean Square Error	To improve prediction performance, this approach could be improved by using hybrid optimization methods with deep learning.	The proposed model was able to attain a better performance result as compared with some existing works
25	Sekeroglu *et al.* (2019)	Using machine learning methods, predict and classify student performance.	Back propagation (BP), Support Vector Regression (SVR) and Long-Short Term Memory (LSTM)	Academic, socio-economic, demographic,	Not Mentioned	Mean Square Error (MSE), R2 Score and Explained Variance (EV) Score and accuracy	The classification accuracy is low as compared to other related works.	This study performed both students' performance prediction and student's performance classification and performance of the prediction and classification models were all evaluate separately in order to get a better view of the system.
26	Shah *et al.* (2019)	To create models that can forecast a student's performance and grades while also taking into account other important personality traits such as hobbies, qualities, and beliefs that influence a student's lifestyle.	Decision Tree, SVM, Random Forest, Logistic Regression, Gradient Boosting, XGBoosting, AdaBoosting, ANN, Recurrent Neural Networks	Demographic, social, academic	UCI Machine Learning Repository	Accuracy	The study did not present enough performance metric for evaluation of the nine (9) techniques as only accuracy was considered without consideration of other significant evaluation metric.	The research provided valuable insight into the performance of nine (9) machine learning algorithms in respect to the challenge of predicting student performance. This insight revealed that the Gradient Boosting method improves accuracy by up to 93.8% since it employs hyper parameters to fine-tune performance.

27	Imran *et al.* (2019)	To create a supervised learning decision tree classifier-based student performance forecasting models.	J48, NNge and MLP	Academic, demographic, social	UCI Machine Learning Repository	Accuracy, Precision, Recall and F-Measure	Just three classifiers were evaluated using the 10 fold cross validation technique without consideration of other cross validation techniques. Also to increase robustness more techniques should be considered for evaluation.	The role of data preprocessing and algorithm fine-tuning tasks in resolving data quality concerns in student performance prediction challenges is demonstrated in this work. This research also looked into the topic of class disparity.
28	Abu-Zohair (2019)	Evaluation of the possibility of predicting students' performance by modelling small dataset size	KNN, Multilayer Perceptron, Naïve Bayes, SVM and Linear Discriminant Analysis	Academic	British University in Dubai	Accuracy and Cohen's kappa	To develop the categorization models, this study solely looked at students' administrative records, neglecting other characteristics like attendance and teacher course delivery that could affect students' learning results.	Developed a machine learning classification method with an acceptable and considerable accuracy rate for classifying student's dissertation project grades.
29	Olalekan *et al.* (2020)	Performance Analysis Of Two Machine Learning Approaches For Graduating Student Prediction	Naïve Bayes and Artificial Neural Network	Academic	Department of Computer Science, Federal Polytechnic, Ile-Oluji in Ondo State, Nigeria	True Positive Rate, False Positive Rate, Precision ,Recall, F-Measure	The data set of 44 instances was used which small and the accuracy of 79.3% obtained is low as compared to previous literatures.	The proposed system is appropriate for predicting student performance since datasets utilized may contain missing data, and this suggested ANN model has the capacity to learn from instances and apply them when a related event occurs, it can perform several tasks in parallel without affecting performance of the system, and it can detect faults and generate output with missing data.
30	Magbag and Jr (2020)	Prediction Of Academic	Logistic Regression	Demographic and	3 Higher Education	Accuracy, Precision,	Due to factors such as the uniqueness of the	The study used large and more robust dataset when

Performance Of Senior High School Graduates	and Artificial Neural Network	Academic	Institutions in the cities of Angeles, San Fernando and Olongapo in Central Luzon,	Recall, F-score and Area under the curve	data collected from each Higher Education Institutions, results may not be generalizable to all institutions. And also an accuracy of 76% was achieved which shows that the proposed model is not as efficient previous works	compared with related works

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Research Design

This chapter provides an overview of the investigation methods used in the study. It contains information on the data set, such as a description of the data and its sources. The research methodology adopted for the goal of this study, as well as the reasons for it, are discussed. A thorough examination of data encoding, feature selection, and data categorization is provided. Finally, the assessment measures that were used to verify the proposed system are explained. Figure 3.1 shows a schematic illustrating each of the methods for predicting student achievement.

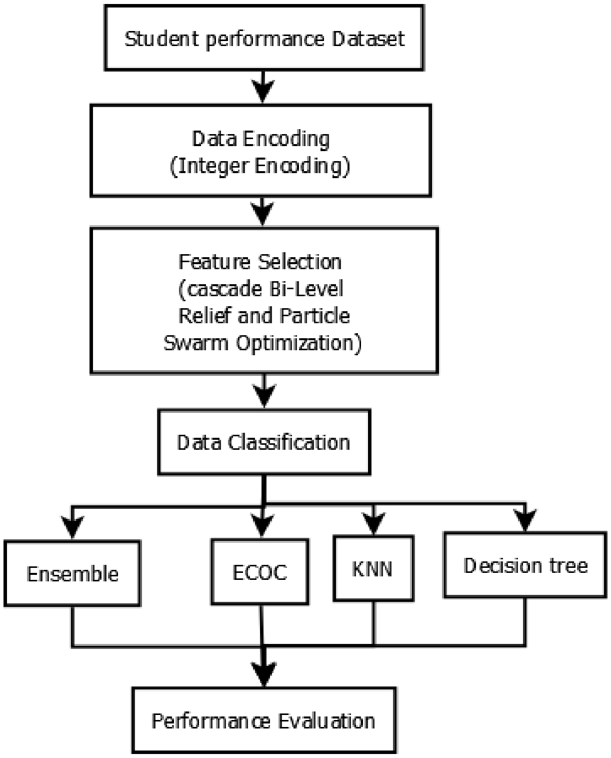

Figure.3.1 Proposed System

3.2 Dataset

The EuroStat known as the Student Performance Data sets were used to carry out the research. The EuroStat datasets was data gotten from UCI repository (Cortez, 2008).

3.2.1 EuroStat (Student Performance Data Set)

The EuroStat, also known as the Student Performance Data set in the UCI repository, was obtained from different public schools in Portugal's Alentejo province during the 2005-2006 academic year. The data set consists of secondary school accomplishment statistics from two Portuguese schools. The data was acquired through school reports and surveys and includes student grades, social, demographic, and school-related characteristics. There are two datasets available, one for mathematics and the other for Portuguese language performance. The mathematics data set contains 395 instances and 33 attributes, 32 of which are predictors and one of which is the target (attribute 33). There are 649 instances in the Portuguese data set, each having 33 properties. The two datasets were modeled using binary/five-level classification tasks by Cortez and Silva (2008). Table 3.1 lists the characteristics and descriptions of the Mathematics and Portuguese datasets.

Table.3.1 Features of Student performance dataset(Cortez & Silva, 2008)

Attributes	Description
School	student’s school (binary: Gabriel Pereira or Mousinho da Silveira)
Sex	student’s sex (binary: female or male)
Age	student’s age (numeric: from 15 to 22)
Address	student’s home address type (binary: urban or rural)
Famsize	family size (binary: ‘LE3’ (≤ 3) or ‘GT3’ (> 3))
Pstatus	parent’s cohabitation status (binary: living together or apart)
Medu	mother’s education (numeric: numeric: 0-none, 1-primary education(4th grade), 2-‘5th to 9th grade’, 3- secondary education, 4-higher education)
Fedu	father’s education (numeric: 0-none, 1-primary education(4th grade), 2-‘5th to 9th grade’, 3- secondary education, 4-higher education)
Mjob	mother’s job (nominal: ‘teacher’, ‘health care related’, ‘civil service’, ‘at_hand’ or ‘other’)
Fjob	father’s job (nominal: ‘teacher’, ‘health care related’, ‘civil service’, ‘at_hand’ or ‘other’)
Reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
Guardian	student’s guardian (nominal: mother, father or other)
Traveltime	home to school travel time (numeric:1(< 15 min.), 2(15 to 30 min)., 3(30 min. to 1 hour) or 4(> 1 hour))
Studytime	weekly study time (numeric: 1(< 2 hours), 2(2 to 5 hours), 3(5 to 10 hours) or 4(> 10 hours))
Failures	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
Schoolsup	extra educational school support (binary: yes or no)
Famsup	family educational support (binary: yes or no)
Paid	extra paid classes (binary: yes or no)
Activities	extra-curricular activities (binary: yes or no)
Nursery	attended nursery school (binary: yes or no)
Higher	wants to take higher education (binary: yes or no)
Internet	Internet access at home (binary: yes or no)
Romantic	with a romantic relationship (binary: yes or no)
Famrel	quality of family relationships (numeric: from 1 –very bad to 5 – excellent)
Freetime	free time after school (numeric: from 1 – very low to 5 – very high)
Goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Health	current health status (numeric: from 1 – very bad to 5 – very good)
Absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)

G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, Output Target)

3.3 Data Preprocessing

As in several countries, the overall assessment in the original data is in the scale of 0–20, with 0 being the worst and 20 being the highest. The data had to be converted to categories according to a scoring policy since the students' final score is in the form of integers, and the expected class should be in the form of categorical values. In this study two different grading systems was used: binary grading and five-level grading. The final grade was first categorized into five categories. The Erasmus framework is used to describe these ranges. Table 3.2 shows that the scale 0–9 corresponds to grade F, which is the lowest grade and corresponds to the mark "fail." The remaining class labels (10–11, 12–13, 14–15, and 16–20) correspond to D (sufficient), C (satisfactory), B (good), and A (excellent) respectively. Secondly, the final grade was categorized into two (binary) categories: fail and pass. In Table 3.3, the range of 0–9 corresponds to F, and it means “fail”; the range of 10–20 refers to A, B, C, and D, and it means “pass.”

Table 3. 2 Five-level grading categories

1	**2**	**3**	**4**	**5**
Excellent	Good	Satisfactory	Sufficient	Fail
16-20	14-15	12-13	10-11	0-9
A	B	C	D	F

Table 3. 3 Binary-level grading categories

0	**1**
Fail	Pass
0-9	10-20
F	A, B, C, D

3.4 Data Encoding

A machine learning model's performance is determined not only by the model and hyper-parameters, but also by how various types of variables are analyzed and fed into the model. Most machine learning models only take numerical variables, therefore preprocessing categorical variables becomes necessary. In order for the model to understand and extract relevant data, these categorical data must be converted to integers (Potdar*et al.*, 2017).

There are both numeric variables and categorical variables in the dataset used. Categories and strings are the most common forms of categorical variables, which have a finite number of possible values to choose from.In this phase, the categorical data types of attributes were converted to numeric attributes. Data encoding was done because specific machine learning algorithms such as Naïve Bayes, support vector machine and Ensemble need numeric attribute types to work. In dealing with numeric data types, machine learning models have also proven to be efficient.

In this study, the integer encoding approach was used. Each string attribute is mapped to an integer value via integer encoding. For the categorical (string) class, integer encoding was employed since integer values have a natural obvious connection between them, which machine learning algorithms may be able to grasp and exploit. And there is an order link between the categorical qualities (Seger, 2018). Gender representation after integer encoding is seen in Table 3.4.

Table 3. 4Integer encoding for Gender Attribute

Male	1
Female	2

3.5 Feature Selection

The process of selecting relevant features as a subset of original features is known as feature selection. One of the most essential and widely used approaches in data preprocessing for data mining is feature selection (Wang *et al.*, 2014). Relevant traits are frequently unknown a priori in real-world settings. As a result, feature selection is essential for predicting student performance by identifying and removing irrelevant and noisy characteristics (Kumari & Swarnkar, 2011). This work introduced a novel cascade bi-level feature selection method for the categorization of student performance data, which employed Relief (RF) filtering and Particle Swarm Optimization (PSO) techniques.

There are two levels to the proposed technique. The Relief approach was used to choose 20 sets of features based on their shared information in the first level (level 1). A new feature subset was created using the 20 sets of features that were chosen. The new 20 feature subset is used as input to the PSO at the second level (level 2), and an optimized feature subset is chosen. The flowchart of the proposed cascaded feature selection technique is shown in figure 3.2. While, figure 3.3 illustrates the proposed feature selection scheme.

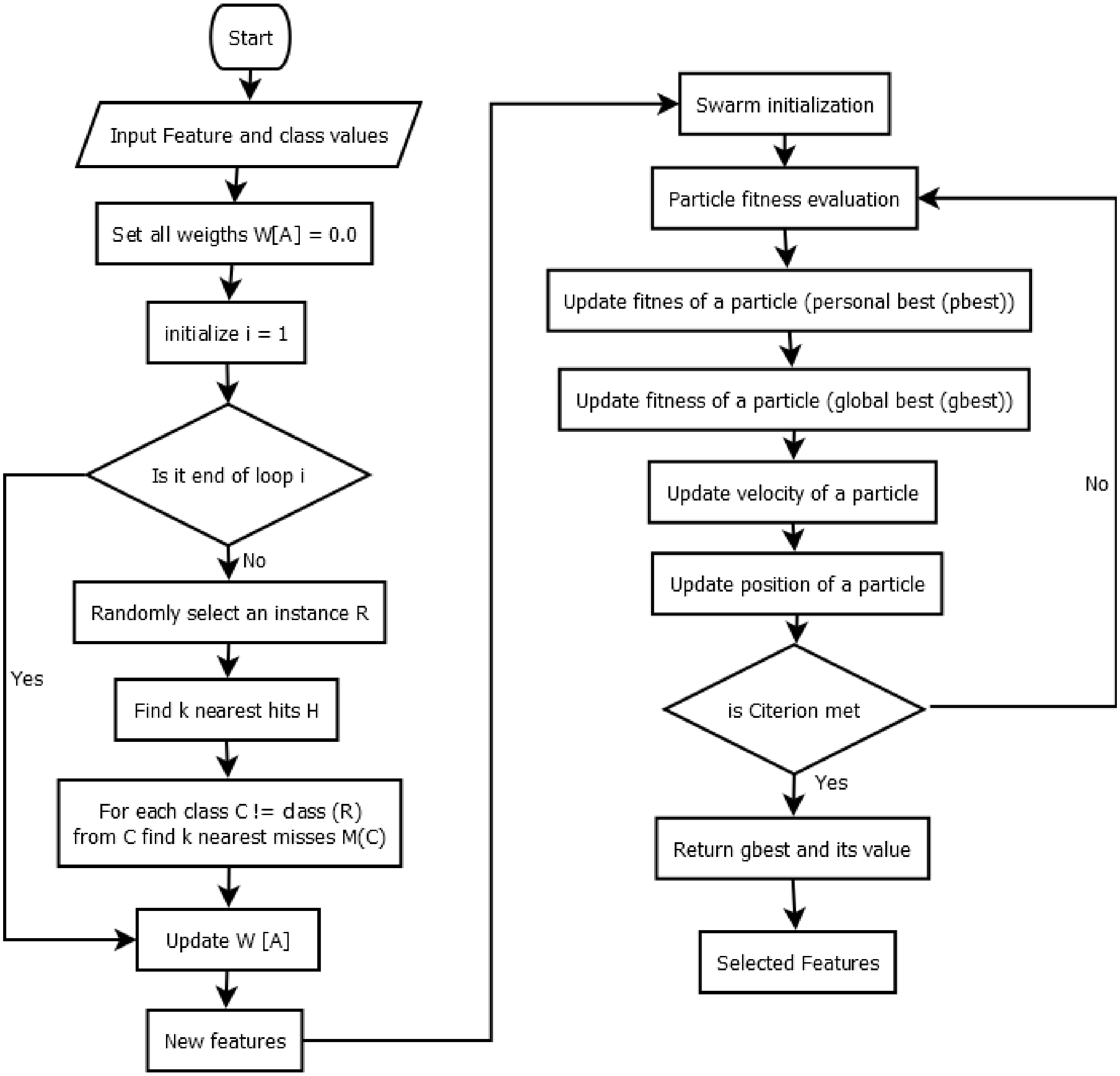
Start
Input Feature and class values
Set all weigths W[A] = 0.0
initialize i = 1
Is it end of loop i
No
Yes
Randomly select an instance R
Find k nearest hits H
For each class C != dass (R) from C find k nearest misses M(C)
Update W [A]
New features
Swarm initialization
Particle fitness evaluation
Update fitnes of a particle (personal best (pbest))
Update fitness of a particle (global best (gbest))
Update velocity of a particle
Update position of a particle
is Citerion met
No
Yes
Return gbest and its value
Selected Features

Figure.3.2 Flowchart of the feature selection technique

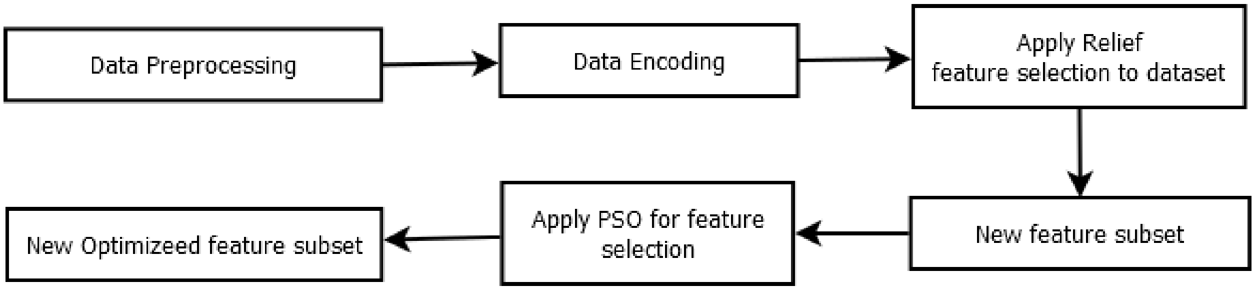
Data Preprocessing
Data Encoding
Apply Relief feature selection to dataset
New feature subset
Apply PSO for feature selection
New Optimizeed feature subset

Figure.3.3 Proposed feature selection

3.5.1 Relief (RF)

Relief is a feature selection algorithm that uses a filter-method approach that is particularly sensitive to feature interactions. Relief generates a proxy statistic for each attribute that can be utilized to assess feature attribute "quality" or "importance" to the target definition (Durgabai & Ravi Bhushan, 2014).). These attribute weights (W [A] = weight of attribute 'A') or feature 'scores' can range from -1 (worst) to +1(best).

The following is how the weight of an attribute is modified iteratively. After a sample is chosen from the results, the nearest adjacent instance that belongs to the same group (nearest hit) and the nearest nearby instance that belongs to the opposite group (nearest miss) are identified. When the value of an attribute changes along with its class, the attribute is weighted with the presumption that the attribute change is the cause of the class change. However, if the attribute value is changed without the class being altered, the attribute is de-weighted because the attribute change has no impact on the class (Rosario, 2015). This process for adjusting the attribute's weight is carried out for a random sample collection or for any sample in the data. The weight changes are then averaged, resulting in a final weight that falls between [-1, 1]. The approximate attribute weight of Relief has a probabilistic interpretation. It's equivalent to the difference between two conditional probabilities, that is the likelihood of the attribute's value changing dependent on the nearest miss and nearest hit (Urbanowicz, *et al.,* 2018).The advantages of using the Relief method is that it is computational fast even when there is big amount of data. Time complexity is not a problem because a consistent number of trials is completed. As a result, the relief technique may complete faster than other filter-based approaches that require all of the data to be considered (Urbanowicz, *et al.,* 2018).

Relief filter-based feature selection algorithm identifies relevant features based on their relationship with the dependent variable. However, Relief technique ignores feature dependencies by considering the relationship between the classifier and each feature autonomously. The particle swarm optimization is used to optimally choose a subset from the selected features to enable feature dependencies and classifier interaction.

3.5.2 Particle Swarm Optimization (PSO)

PSO is a computer approach for addressing issues in which a candidate solution is repeatedly improved in terms of a quality indicator (Talukder, 2011). It resolves a problem by creating a population of potential answers, known as particles, and shifting them around in the search area using a basic mathematical formula depending on their velocity and position (Sengupta *et al.*, 2019). Consider the global optimum of an m-dimensional function in equation 3.1.

$$G(x_1, x_2, x_3, x_4, \dots, x_m) = G(X) \tag{3.1}$$

Where x_i is the search variable, which reflects the supplied function's set of free variables? The goal is to locate a value x^*that is either a maximum or a minimum in the search space for the function $G(x^*)$.

The PSO technique is a multiply agent simultaneous search technique that keeps track of a swarm of particles, each of which represents a possible answer in the swarm. Every particle moves across a multi-dimensional search area, adjusting its position based on its own and surrounding experiences (Poli*et al.*, 2007). If $x_i^s$is the position vector of particle I in the multidimensional search space at time step s, then the locations of each particle in the search space are updated using equation 3.2.

$$x_i^{s+1} = x_i^s + v_i^{s+1} \ with \ x_i^0 \sim U(x_{min}, x_{max}) \tag{3.2}$$

Where $v_i^s$is the velocity vector of particle I, which drives the optimization process and represents both the individual and social experience knowledge of all particles? The uniform distribution $U(x_{min}, x_{max})$has $x_{min}$and x_{max} as its minimum and maximum values, respectively.

The velocity of the particle $i$up dated using equation 3.3.

$$v_i^{s+1} = w * v_i^s + c_1 * r_{1i} * (p_i - x_i^s) + c_2 * r_{2i} * \left(p_g - x_i^s\right) \qquad (3.3)$$

Where $s$denotes the sth iteration in the process, w is inertia weight andc_1 and c_2 are acceleration constants. $r_{1i}$and$r_{2i}$are random values uniformly distributed in [0,1]. $p_i$and$p_g$represent the elements of $pbest$ and $gbest$ respectively. The flowchart for selection of features using PSO is displayed in figure.3.4

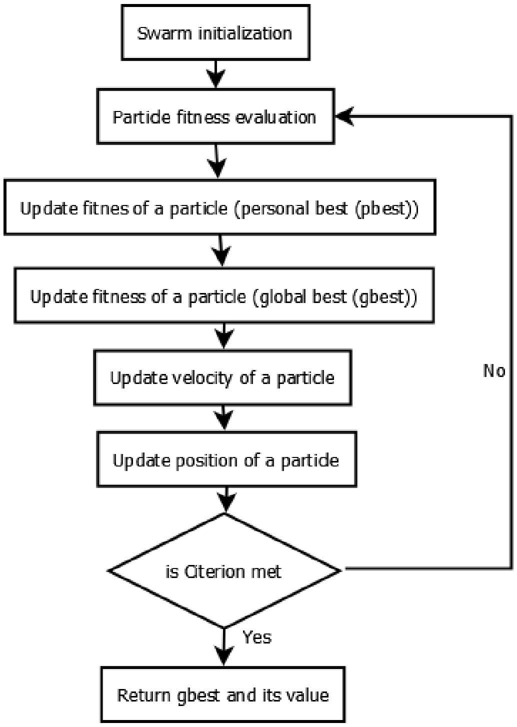
Swarm initialization
Partide fitness evaluation
Update fitnes of a particle (personal best (pbest))
Update fitness of a particle (global best (gbest))
Update velocity of a particle
Update position of a partide
is Citerion met
No
Yes
Return gbest and its value

Figure.3.4Flow chat for PSO feature selection

As a result, in a PSO approach, all particles are launched at random and evaluated to determine fitness, as well as the personal best (best value of each particle) and global best (best value of particle in the entire swarm). After that, a loop is started in order to discover the best option. The personal and global bests are used to update the velocity of the particles in the loop, and then the current velocity is used to update the position of each particle. The loop is terminated by a defined ending criterion (Sahu & Mishra, 2012).

3.6 Data Classification

Machine learning capability lies in its ability to generalize by correctly classifying unknown information based on models developed using the training dataset. The extracted optimized subset features were used for training of the classifiers for student performance prediction. The classifiers were also trained with the original features without selection. In this work four machine learning classification models were used for training and classification, namely, Error-Correcting Output Codes (ECOC), Decision Tree (DT), Ensemble, and K-Nearest Neighbor.

Each of the four classifiers was trained to classify students with the selected features and with the whole features without selection. 80% of the data was used for training, and the remaining 20% was used to test the trained models. These five classification model used for evaluation of the proposed model are presented below.

3.6.1 Error-Correcting Output Codes (ECOC)

Machine learning models are built for binary classification problems, such as Support Vector Machine (SVM) and logistic regression. As such, these binary algorithms either need to be updated or not used at all for multiclass classification problems. The ECOC technique is a tool that allows the issue of multiclass classification to be interpreted as multiple problems of binary type, enabling the direct use of native binary classification models (Armano *et al.*, 2013). The ECOC enables the encoding of an infinite number of binary classification problems for each class (Dietterich & Bakiri, 1995). ECOC designs are independent of the classifier depending on the implementation. ECOC has error-correcting properties and has shown that the learning algorithm's bias and variance can be decreased (Escalera *et al.*, 2010).

The ECOC architecture has two fundamental processes: coding and decoding. The creation of a code matrix$\in \{-1, 0, 1\}^{c^X n}$ with crows and n columns, where c and n signify the number of classes and dichotomizers, respectively, is the key to the coding process. The code word $C_i$for the i^{th}class($i = 1, 2, \dots, c$) appears in the i^{th}row of *M*. In the meantime, each column of *M* reflects the dichotomizer's partition of classes. When training the dichotomizers, classes coded with 1 and -1 are regarded as positives and negatives, accordingly, whereas classes coded with 0 are omitted from the training set.The outputs of these n trained dichotomizers for the test sample are given as a vector $V = \{v_1, v_2, \dots, v_n\}$ in the decoding process, and compared with the codeword of each class to find the closest one to determine the test sample's class label. In equation 3.4, the distance between the vector *V* and each codeword $C_i$is calculated.

$$D(V, C_i) = \frac{1}{2}\sum_{j=1}^{n} L(V(j). C_i(j)) \tag{3.4}$$

Where$L(.)$ signifies the loss function which is reliant on the type of dichotomizer.

3.6.2 Decision Tree (DT)

A DT is a simple and commonly used predictive modeling technique. DT is a type of supervised learning where, according to a particular parameter, the data is continually split (Patel & Singh, 2015). The decision tree employs a tree-like framework to progress from observations about an item (symbolized by the branches) to inferences about the item's target value (described in the leaves) (Kolo *et al.*, 2015). Regression and classification problems can be solved using the DT algorithm. DT is easy to understand and view. It does not require normalization of data and preparation of data; it needs less effort. The decision to do strategic splits has a significant effect on a tree's precision (Olaniyi *et al.*, 2018). Entropy, information gain and reduction invariance are techniques used in determining which attribute to the position at the root or the different levels of the tree.

The entropy of processed data is a measure of its randomness. The higher the entropy, the more difficult it is to draw any judgments from the information. A branch with an entropy of zero, for example, is chosen as the root node, and further division is required for a branch with an entropy greater than zero (Olaniyi *et al.*, 2018). In equation.3.5 entropy for a single attribute is expressed.

$$E(S) = \sum_{i=1}^{n} -p_i log_2\, p_i \qquad (3.5)$$

Where S denotes the present state, $p_i$is the likelihood of an event i of state S.

Information Gain (IG) is a statistical property that tests how well the training examples are segregated according to their target classification by a given attribute. In equation 3.6, information gain is expressed mathematically.

$$IG = Entropy(before) - \sum_{j=1}^{N} Entropy\ (j, after) \qquad (3.6)$$

Where “*before*” refers to the dataset prior to the split, *N* refers to the number of subsets formed by the division, and *(j, after)* refers to subset *j* following the division.

Reduction invariant is a technique for solving regression problems. To choose the optimal split, this technique uses the usual variance formula. The split with the lowest variance is chosen as the criterion for dividing the population. The usual variance formula employed in this technique is stated in equation 3.7.

$$variance = \frac{\sum(X-\mu)^2}{n} \quad (3.7)$$

Where μ is the mean of the values and X is the actual value and n is the number of values.

3.6.3 K-Nearest Neighbour (KNN)

When solving classification and regression issues, the K-Nearest Neighbors (KNN) technique is used (Zhang, 2016).The KNN algorithm takes into account the proximity of linked objects. An item is grouped in KNN by the majority vote of its KNN, with an item allocated to the most mutual class of its KNNs (Arade & Patil, 2017; Kataria & Singh, 2013). KNN does not need a training phase. KNN, however, suffers from the curse of dimensionality, and it is vulnerable to outliers. The Euclidean distance is a commonly used similarity measure in KNN (Gu *et al.*, 2019). The Euclidean distance is the linear distance between two points in Euclidean space. The Euclidean distance is expressed in equation 3.8.

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (3.8)$$

Where p, q are two points in Euclidean n-space, $q_i \; and \; p_i$ are the Euclidean vectors, starting from the origin of the space and n is the n-space.

3.6.4 Ensemble Classifier

When opposed to a single classifier, an ensemble learning model integrates predictions from several models with a two-fold goal: the first goal is to maximize prediction accuracy (Amrieh *et al.*, 2016). Due to the use of several advanced classifiers, the second benefit is increased critical generalizability. As a result, an ensemble can find answers to problems that a single prediction model would have. An ensemble can choose a group of hypotheses from a much wider hypothesis space and combine their predictions into a single forecast (Adejo & Connolly, 2018).

Classifiers in the ensemble learning model are combined into meta-classifiers via voting or weighted voting of their forecast for the final estimates (Almasri *et al.*, 2019).

3.7 Performance Metrics

In this study, the accuracy, Precision, Recall, and F-score performance measures were used to evaluate the proposed method. This measure is explained below.

3.7.1 Accuracy

The rate of correct classifications is referred to as accuracy. Accuracy is calculated in equation 3.9.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True negative}}{\text{True Positive} + \text{True negative} + \text{False Positive} + \text{False negative}} \tag{3.9}$$

3.7.2 Recall

Sensitivity is another term for recall. The amount of correct positive predictions that could have been made from all positive predictions is calculated by recall. The recall is calculated using the formula in equation 3.10.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}} \tag{3.10}$$

3.7.3 Precision

Precision is a metric used to calculate how many positive predictions are accurately made. The number of true positive elements is derived by dividing the total number of true positives by the total number of false positives. The formula in equation 3.11 is used to define precision.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \tag{3.11}$$

3.7.4 F-Score

The f-score of a model is defined as the harmonic average of precision and recall. F-Score is represented in equation 3.12.

$$\mathrm{F-Score} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{3.12}$$

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION

4.1 Results

In this work two classification task was carried out. These classification task are: binary-level grading classification and the five-level grading classification task. The Mathematics and the Portuguese dataset was used. The mathematics data set consist of 395 instances and 33 attributes while the Portuguese data set consist of 649 instances with 33 attributes. The datasets were divided in the ratio of 4:1 for training and testing (80% for training and 20% for testing).

4.1.1 Binary-level Grading Classification

The binary classification deals with classification using the two classes which are pass and fail. The original label as mention in chapter 3 consists of 0-20 labels or grades. Where 0 represents the lowest rating and 20 represents the highest. Because the students' final grades are in the form of integers, the projected class should be in the form of categorical values, so the data must be translated into categories using a grading scheme. In the binary classification the integer labels were categorized into two classes where Fail (0) represents grade 0-9 and Pass (1) represents grade 10-20. Using the binary labels the four classifiers (ECOC, Ensemble, KNN and Decision Tree) were trained and tested using the original features (no feature selection done), sub-features using relief feature selector, sub-features using PSO feature selector and sub-features using the cascade bi-level feature selector. The classification result of the four classifiers before and after performing relief feature selection is shown Table 4.1.

Table.4.1 Binary-level Classification Performance Before and After Relief Feature Selection

Accuracy (%) Before and After Relief Feature Selection								
Feature Selection	Mathematics				Portuguese			
	ECOC	Ensemble	KNN	DT	ECOC	Ensemble	KNN	DT
Before	89.87	91.14	70..89	87.34	92.31	92.31	89.23	91.54
After	91.14	92.41	79.75	89.87	93.08	93.08	91.54	92.31
Selected Features	G2, G1, Sex, Paid, Failures, Activities, Romantic, Famsup, Studytime, Higher, Mjob, Pstatus, Dalc, Medu, Guardian, Goout, Walc, Absences, Age, School				School, G2, G1, Activities, Sex, Address, Famsup, Failures, Nursery, Reason, Romantic, Higher, Medu, Famrel, Schoolsup, Fedu, Internet, Goout, Studytime, Health			

Table 4.1 shows the accuracy of ECOC, Ensemble, KNN and Decision Tree (DT) for both Mathematics and Portuguese data sets using all of the 32 original features and using just the Relief selected 20 feature subsets. The best performance for Mathematics before relief feature selection is 91.14% which was obtained by the ensemble classifier. After performing relief feature selection the accuracy of the ensemble classifier moved from 91.14% to 92.41% which shows an improvement. For the Portuguese data set before relief feature selection the ensemble and ECOC classifiers achieved the best performance with an accuracy of 92.31% each. However the ECOC and Ensemble classifier accuracy attained better accuracy of 93.08% when trained with the Relief selected feature subset. The twenty Relief selected feature subsets are presented in the last row of Table 4.1. It can also be seen from the results in Table 4.1 that all of the classifiers performed better when trained with Relief selected feature sets for both the Mathematics and Portuguese data set. Table 4.2 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the Relief selected sub-features.

Table.4.2 Binary–level Classification Results for Relief Selected Features

Relief Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	91.14	85.71	87.50	84.00	93.08	82.35	80.77	84.00
Ensemble	92.41	86.35	79.71	95.00	93.08	81.63	76.92	86.96
KNN	79.75	66.67	66.67	66.67	91.54	77.55	73.08	82.61
DT	89.87	82.61	79.17	86.36	92.31	79.17	82.61	86.36
Selected Features	G2, G1, Sex, Paid, Failures, Activities, Romantic, Famsup, Stidytime, Higher, Mjob, Pstatus, Dalc, Medu, Guardian, Goout, Walc, Absences, Age, School				School, G2, G1, Activities, Sex, Address, Famsup, Failures, Nursery, Reason, Romantic, Higher, Medu, Famrel, Schoolsup, Fedu, Internet, Goout, Studytime, Health			

To properly evaluate the performance of ECOC, Ensemble, DT and KNN classifiers when trained with Relief selected feature subsets for both Mathematics and Portuguese data set, their precision, recall, f-score and accuracy are presented in Table 4.2. The ensemble classifier performed best with an accuracy of 92.41% and f-score of 86.35% for Mathematics dataset. While ECOC performed best with an accuracy of 93.08% and f-score of 82.35% for Portuguese dataset. Table 4.3 shows the performance of the four classification models when trained with the initial 32 features of the Mathematics and Portuguese data set and when trained with the PSO selected feature subsets.

Table.4.3 Binary-level Classification Before and After PSO Feature Selection

Accuracy (%) Before and After PSO Feature Selection								
Feature Selection	**Mathematics**				**Portuguese**			
	ECOC	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**

Before	89.87	91.14	70.89	87.34	92.31	92.31	89.23	91.54
After	92.41	93.67	92.41	92.41	93.85	94.62	91.42	93.85
Selected Features	School, Age, Famsize, Medu, Fjob, Guardian, Failures, Famsup, Paid, Activities, Nursery, Internet, Romantic, Freetime, G1, G2				Age, Address, Famsize, Fjob, Reason, Traveltime, Studytime, Failures, Famsup, Paid, Freetime, Goout, G1, G2			

The PSO feature selector selected 16 features out of the initial 32 features for Mathematics dataset. For the Portuguese data set PSO selected 14 sub-features out of the initial 32 features. Table 4.3 shows the performance of ECOC, Ensemble, DT and KNN classifiers before and after applying PSO feature selection on Mathematics and Portuguese datasets. The performance of ECOC improved from 89.87% before feature selection to 92.41% after apply PSO feature selection for Mathematics dataset. The performance of Ensemble increased from 91.14% before feature selection to 93.67% after apply PSO feature selection for Mathematics dataset. Also the performance of KNN and DT improved from 70.89% and 87.34% respectively to 92.41% each for Mathematics dataset. For the Portuguese dataset the training with the PSO selected feature set produced better accuracies than using the classifiers with all the features. This improvement is seen in the ECOC classifier which had an increase from 92.31% to 93.85%. The improvement is also seen in Ensemble (from 92.31% to 94.62%), KNN (from 89.31% to 91.42%) and DT (from 91.54% to 93.85%) performances. Table 4.4 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the PSO selected sub-features.

Table.4.4 Binary–level Classification for PSO Selected Features

PSO Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	92.41	84.21	76.19	94.12	93.85	78.95	75.00	83.33
Ensemble	93.67	87.90	85.71	90.00	94.62	82.05	80.00	84.21
KNN	92.41	84.21	76.19	94.12	91.42	70.59	60.00	85.71
DT	92.41	84.21	76.19	94.12	93.85	78.95	75.00	83.33
Selected features	School, Age , Famsize, Medu, Fjob, Guardian, Failures, Famsup, Paid, Activities, Nursery, Internet, Romantic, Freetime, G1, G2				Age, Address, Famsize, Fjob, Reason, Traveltime, Studytime, Failures, Famsup, Paid, Freetime, Goout, G1, G2			

In Table 4.4 Ensemble classifier obtained the best performance for both Mathematics and Portuguese data sets with an accuracy of 93.67% and an f-score of 87.90% for Mathematics data set and an accuracy of 94.62% and an f-score of 82.05% for Portuguese data set. ECOC, KNN and DT performance equally when trained with PSO selected Mathematics sub-features. However for the Portuguese data set KNN performed least with an accuracy of 91.42% and f-score of 70.59%. Table 4.5 is a comparative result of the student’s prediction models when trained with all the features in the Mathematics and Portuguese data set and when trained with the cascaded bi-level selected feature subsets.

Table.4.5 Binary-level Classification Before and After Cascaded Bi-level Feature Selection

Accuracy (%) Before and After Cascaded Bi-level Feature Selection								
Feature Selection	**Mathematics**				**Portuguese**			
	ECOC	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Before	89.87	91.14	70..89	87.34	92.31	92.31	89.23	91.54
After	93.67	94.94	92.89	92.89	95.38	96.15	93.85	93.85
Selected Features	G2, G1, Sex, Activities,Famsup, Studytime, Mjob, Medu, Guardian, Goout, Walc				G2, G1,Nursery, Reason, Romantic, Higher, Schoolsup, Goout			

The cascaded bi-level selected features are shown in the last row of Table 4.5. For Mathematics data set only 11 features where selected as the optimal features which have higher impact on the student's grade. For the Portuguese data set 8 features where selected as the optimal features which have higher impact on the student's grade. From this selected features it can be seen that G2, G1, and Goout is a common selected feature in both Mathematics and Portuguese data sets. The results in Table 4.5 shows that cascaded bi-level feature selector selected optimal features as there was an increase in performance after applying the cascaded bi-level feature selection technique. Table 4.6 is comprehensive performance report of ECOC, Ensemble, DT and KNN when trained with cascaded bi-level selected feature subsets.

Table.4.6 Binary–level Classification for Cascaded Bi-level Selected Features

Cascaded Bi-level Selected Features Classification Results								
	Mathematics				**Portuguese**			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	93.67	89.36	87.50	91.30	95.38	78.95	75.00	83.33
Ensemble	94.94	91.30	87.50	95.45	96.15	82.05	80.00	84.21

KNN	92.89	84.21	87.50	80.77	93.85	70.59	60.00	85.71
DT	92.89	86.76	83.33	90.91	93.85	78.95	75.00	83.33
Selected Features	G2, G1, Sex, Activities, Famsup, Studytime, Mjob, Medu, Guardian, Goout, Walc				G2, G1,Nursery, Reason, Romantic, Higher, Schoolsup, Goout			

In Table 4.6 the Ensemble classifier produced the best performance for both the Mathematics and Portuguese data sets. Ensemble classifier got an accuracy of 94.94%, f-score of 91.30%, and precision of 87.50% and recall of 95.45% for Mathematics dataset. For the Portuguese data set Ensemble classifier obtained an accuracy of 96.15%, f-score of 82.05%, and precision of 80.0% and recall of 84.21%. Table 4.7 is a comparison of the performance based on accuracy of the Relief, PSO and Cascaded bi-level feature selection techniques.

Table.4.7 Comparison of Feature Selection Techniques

	Accuracy (%)							
	Mathematics				**Portuguese**			
Feature Selection	**ECOC**	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Relief	91.14	92.41	79.75	89.87	93.08	93.08	91.54	92.31
PSO	92.41	93.67	92.41	92.41	93.85	94.62	91.42	93.85
Cascaded Bi-level	**93.67**	**94.94**	**92.89**	**92.89**	**95.38**	**96.15**	**93.85**	**93.85**

In Table 4.7 the highest classification accuracy obtained for classification using the Relief selected features for the Mathematics data set is 92.41%. The PSO selected sub-features obtained a classification accuracy of up to 93.67% for the Mathematics data set. The proposed cascaded bi-level obtained a classification accuracy of 94.94% for the Mathematics data set. For the Portuguese data set the highest classification accuracy obtained for classification using the Relief selected features is 93.08% by Ensemble and ECOC classifiers. The PSO selected sub-features obtained a classification accuracy of up to 94.62% for the Portuguese data set. The proposed cascaded bi-level obtained a classification accuracy of 96.15% for the Portuguese data set. In conclusion the proposed technique selected the best sub-features that achieved higher classification accuracy than the sub-features selected by a single-level relief or PSO selector.

Figure 4.1 is a visual representation of the obtained results in Table 4.7 for Mathematics data set using Relief selected feature set, PSO selected feature set and cascaded bi-level selected feature sets.

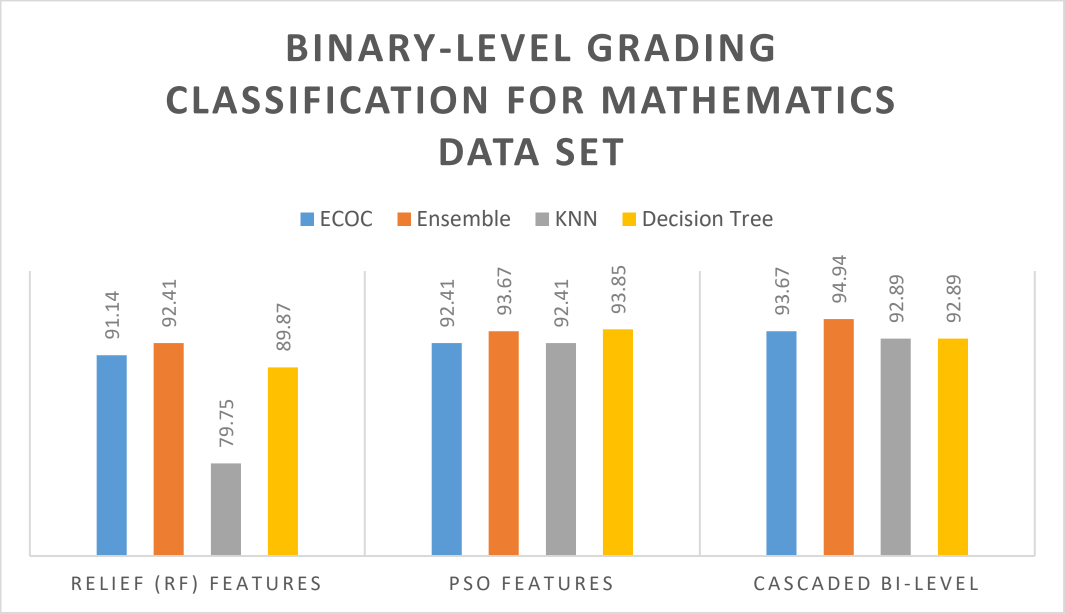

Figure 4. 1 Classification performance for binary-level grading for Mathematics data set

Figure 4.2 is a visual representation of the obtained results in Table 4.7 for Portuguese data set using Relief selected feature set, PSO selected feature set and cascaded bi-level selected feature sets.

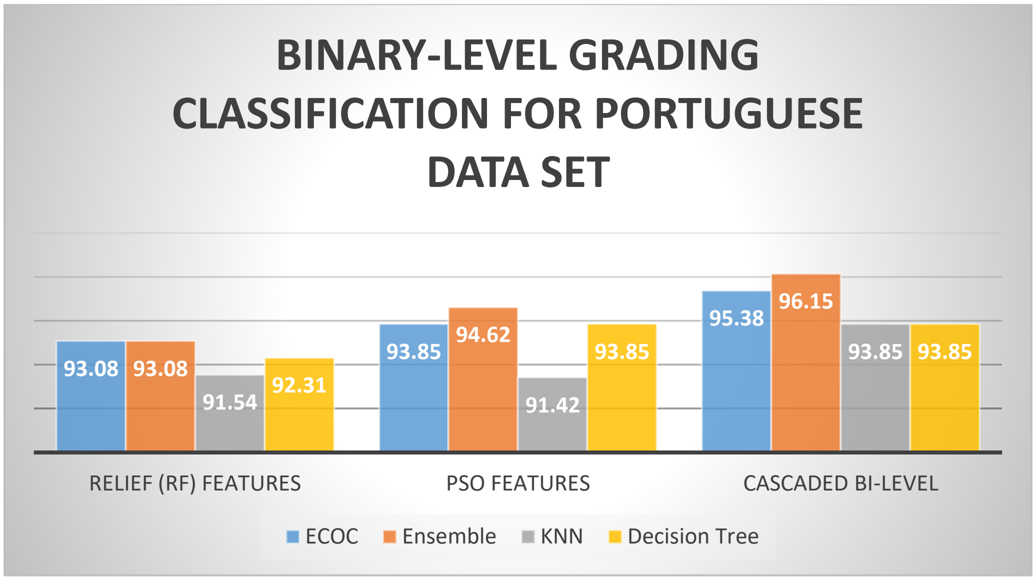

Figure 4. 2 Classification performance for binary-level grading for Portuguese data set

Table 4.8 presents a comparison of the performance of the proposed technique with related work that used the EuroStat dataset with respect to binary classification. The results obtained showed that the proposed technique achieved a higher student prediction accuracy than related work.

Table 4. 8 Comparison of binary classification performance with related work

		Mathematics	**Portuguese**
Techniques	Dataset	Highest Obtained Accuracy (%)	Highest Obtained Accuracy (%)
Ünal (2020)	EuroStat	93.67	93.22
Shah *et al.*, (2019)	EuroStat	93.80	
Cascaded Bi-level	EuroStat	**94.94**	**96.15**

4.1.2 Five-Level Grading Classification

The five-level grading classification deals with classification using the five classes which are excellent (5), good (4), satisfactory (3), sufficient (2) and fail (1). The original label of 0-20 labels or grades were categorized into the aforementioned five classes.

Using the five-level grading the four classifiers (ECOC, Ensemble, KNN and Decision Tree) were trained and tested using the original features (no feature selection done), sub-features using relief feature selector, sub-features using PSO feature selector and sub-features using a cascade bi-level feature selector. In Table 4.9, the accuracy rates for ECOC, Ensemble, KNN and DT were equated before and after the feature selection procedure using Relief technique for the Mathematics and Portuguese dataset for five-level grade version.

Table.4.9 Five-level Classification Before and After Relief Feature Selection

Accuracy (%) Before and After Relief Feature Selection								
Feature Selection	**Mathematics**				**Portuguese**			
	ECOC	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Before	72.42	72.68	69.62	64.56	72.31	74.62	70.77	66.15
After	75.95	79.75	75.95	70.89	76.15	76.92	74.62	72.31
Selected Features	G2, G1, Sex, Medu, Walc, Studytime, Address, Paid, Schoolsup, Mjob, Failures, Higher, Pstatus, Dalc, school, Freetime, Age, Famsup, Internet, Absences				G2, G1, School, Activities, sex, Studytime, Higher, Medu, Failures, Schoolsup, Nursery, Health, Famsup, Goout, Pstatus, Address, Fedu, Internet, Reason, Walc			

The accuracy rate of the ECOC algorithm has grown from 72.42% to 75.95% with the selected attributes in Table 4.9 for the Mathematics data set. The Ensemble algorithm's accuracy rate climbed from 72.68% to 79.75%. The accuracy rate for the KNN algorithm has grown from 69.62% to 75.95%. Finally, the accuracy rate for DT improved from 64.56% to 70.89%. With the selected attributes, the ECOC algorithm's accuracy rate increased from 72.31% to 76.15% for the Portuguese data set. The Ensemble algorithm's accuracy rate went from 74.62% to 76.92%. The accuracy rate for the KNN algorithm has grown from 70.77% to 74.62%. Finally, the accuracy rate for DT increased from 66.15% to 72.31%.

Table.4.10 Five–level Classification for Relief Selected Features

Relief Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	75.95	88.52	84.38	93.10	93.08	82.35	80.77	84.00
Ensemble	79.75	91.80	87.50	85.25	93.08	81.63	76.92	86.96
KNN	75.95	87.50	87.50	87.50	91.54	77.55	73.08	82.61
DT	70.89	85.25	81.25	89.66	92.31	79.17	82.61	86.36
Selected Features	G2, G1, Sex, Medu, Walc, Studytime, Address, Paid, Schoolsup, Mjob, Failures, Higher, Pstatus, Dalc, school, Freetime, Age, Famsup, Internet, Absences				G2, G1, School, Activities, sex, Studytime, Higher, Medu, Failures, Schoolsup, Nursery, Health, Famsup, Goout, Pstatus, Address, Fedu, Internet, Reason, Walc			

To properly evaluate the performance of ECOC, Ensemble, DT and KNN classifiers when trained with Relief selected feature subsets for both Mathematics and Portuguese data set for the five-level grading version, their precision, recall, f-score and accuracy are presented in Table 4.10. The ensemble classifier performed best with an accuracy of 79.75% and f-score of 91.80% for Mathematics dataset. While ECOC performed best with an accuracy of 93.08% and f-score of 82.35% for Portuguese dataset.

Table 4.11 shows the performance of the four classification models for the five-level grading when trained with the initial 32 features of the Mathematics and Portuguese data set and when trained with the PSO selected feature subsets.

Table.4.11 Five-level Classification Before and After PSO Feature Selection

Accuracy (%) Before and After PSO Feature Selection								
Feature Selection	Mathematics				Portuguese			
	ECOC	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**

Before	72.42	72.68	69.62	64.56	72.31	74.62	70.77	66.15
After	75.95	78.48	74.68	72.15	77.69	78.64	76.92	71.77
Selected Features	Sex, Age, Famsize, Medu, Failures, Schoolsup, Famsup, Paid, Activities, Nursery, Internet, Romantic, Famrel, Freetime, G1, G2				Sex, Medu, Failures, Schoolsup, Paid, Activities, Internet, Famrel, Freetime, Goout, Health, G2, G1			

The accuracy rates for the Portuguese and Mathematics data sets for the five-level grading version were compared in Table 4.11 before and after the PSO attribute selection process. The accuracy of the classifiers in these data sets was greatly improved by attribute selection. For both the Portuguese and Mathematics data sets, the ensemble approach produced the best leap. Using the PSO selected feature sets increased the accuracy rate from 72.68% to 78.48% for the Mathematics data set and the ensemble accuracy increased from 74.62% to 78.64% for the Portuguese dataset. Using the PSO selected features for training and testing increased the accuracy of ECOC, KNN and DT significantly.

Table.4.12 Five–level Classification for PSO Selected Features

PSO Selected Features Classification Results								
	Mathematics				**Portuguese**			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	75.95	78.26	75.00	81.82	77.69	77.78	77.78	77.78
Ensemble	78.48	80.85	79.17	82.26	78.64	78.28	69.23	90.00
KNN	74.68	80.00	75.00	85.71	76.92	76.92	76.92	76.92
DT	72.15	78.26	75.00	81.82	71.77	62.52	76.92	52.63
Selected Features	Sex, Age, Famsize, Medu, Failures, Schoolsup, Famsup, Paid, Activities, Nursery, Internet, Romantic, Famrel, Freetime, G1, G2				Sex, Medu, Failures, Schoolsup, Paid, Activities, Internet, Famrel, Freetime, Goout, Health, G2, G1			

In Table 4.12 Ensemble classifier obtained the best performance for both Mathematics and Portuguese data sets with an accuracy of 78.26% and an f-score of 80.85% for Mathematics data set and an accuracy of 78.64% and an f-score of 78.28% for Portuguese data set. For both Mathematics and Portuguese data set DT performed least with an accuracy of 72.15% and f-score of 78.26% for Mathematics data set and accuracy of 71.77% and f-score of 62.52% for Portuguese data set. Table 4.13 is a comparative result of the student's prediction models when trained with all the features in the Mathematics and Portuguese data set and when trained with the cascaded bi-level selected feature subsets.

Table.4.13 Five-level Classification Before and After Cascaded Bi-Level Feature Selection

	Accuracy (%) Before and After Cascaded Bi-Level Feature Selection							
Feature	**Mathematics**				**Portuguese**			
Selection	**ECOC**	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Before	72.42	72.68	69.62	64.56	72.31	74.62	70.77	66.15
After	83.54	84.81	81.01	73.42	83.08	83.85	77.38	72.67
Selected Features	G2, G1, Walc, Studytime, Address, Paid, Schoolsup, Failures, Dalc, Internet				G2, G1, sex, Famsup, Pstatus, Address			

The outcomes of the cascaded bi-level feature selection approach are compared in Table 4.13 before and after administration to the Portuguese and Mathematics five-level grading versions. There was a significant improvement in precision. For both the Portuguese and Mathematics data sets, Ensemble classifier produced the best results. The Ensemble approach produced the best leap in the Mathematics data set. Ensemble result has risen from 72.68% to 84.81% which is a 12.13% increase. For Portuguese data set the best jump was experienced by the ECOC method. ECOC result has risen from 72.31% to 83.08% which is a 10.77% increase. Table 4.14 is comprehensive performance report of ECOC, Ensemble, DT and KNN when trained with cascaded bi-level selected feature subsets for the five-level grading version of Portuguese and Mathematics.

Table.4.14 Five–level Classification for Cascaded Bi-level Selected Features

Cascaded Bi-level Selected Features Classification Results								
Classifiers	**Mathematics**				**Portuguese**			
	Accuracy (%)	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	83.54	90.00	84.38	96.43	83.08	88.24	83.33	93.75
Ensemble	84.81	92.31	93.75	90.91	83.85	87.50	77.78	100
KNN	81.01	88.14	81.25	96.30	77.38	74.29	72.22	76.47
DT	73.42	81.97	78.13	86.21	72.67	76.19	88.89	66.77
Selected Features	G2, G1, Walc, Studytime, Address, Paid, Schoolsup, Failures, Dalc, Internet				G2, G1, sex, Famsup, Pstatus, Address			

In Table 4.14 the Ensemble classifier produced the best performance for both the Mathematics and Portuguese data sets. Ensemble classifier got an accuracy of 84.81%, f-score of 92.31%, and precision of 93.75% and recall of 90.91% for Mathematics dataset. For the Portuguese data set the Ensemble classifier obtained an accuracy of 83.85%, f-score of 87.50%, and precision of 77.78% and recall of 100%. Table 4.15 is a comparison of the performance based on accuracy of the Relief, PSO and Cascaded bi-level feature selection techniques.

Table.4.15 Comparison of Feature Selection Techniques

Accuracy (%)								
	Mathematics				**Portuguese**			
Feature Selection	**ECOC**	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Relief	75.95	79.75	75.95	70.89	76.15	76.92	74.62	72.31
PSO	75.95	78.48	74.68	72.15	77.69	78.64	76.92	71.77

Cascaded Bi-level	**83.54**	**84.81**	**81.01**	**73.42**	**83.08**	**83.85**	**77.38**	**72.67**

In Table 4.15 the highest classification accuracy which was obtained by Ensemble classifier using the Relief selected features for the Mathematics data set is 79.75%. The PSO selected sub-features obtained a classification accuracy of up to 78.48% from Ensemble classifier for the Mathematics data set. The proposed cascaded bi-level obtained the highest accuracy of 84.81% when compared with Relief and PSO performance for the Mathematics data set. For the Portuguese data set the highest classification accuracy obtained for classification using the Relief selected features is 76.92% by Ensemble classifier. The PSO selected sub-features obtained a classification accuracy of up to 78.64% for the Portuguese data set using the Ensemble classifier. The proposed cascaded bi-level obtained highest accuracy of 83.85% when compared with Relief and PSO performance for the Portuguese data set. In conclusion the proposed technique selected the best sub-features that achieved a higher classification accuracy than the sub-features selected by a single-level relief or PSO selector.

Figure 4.3 is a visual representation of the obtained results in Table 4.15 for Mathematics data set using Relief selected feature set, PSO selected feature set and cascaded bi-level selected feature sets. Figure 4.4 is a visual representation of the obtained results in Table 4.15 for Portuguese data set using Relief selected feature set, PSO selected feature set and cascaded bi-level selected feature sets.

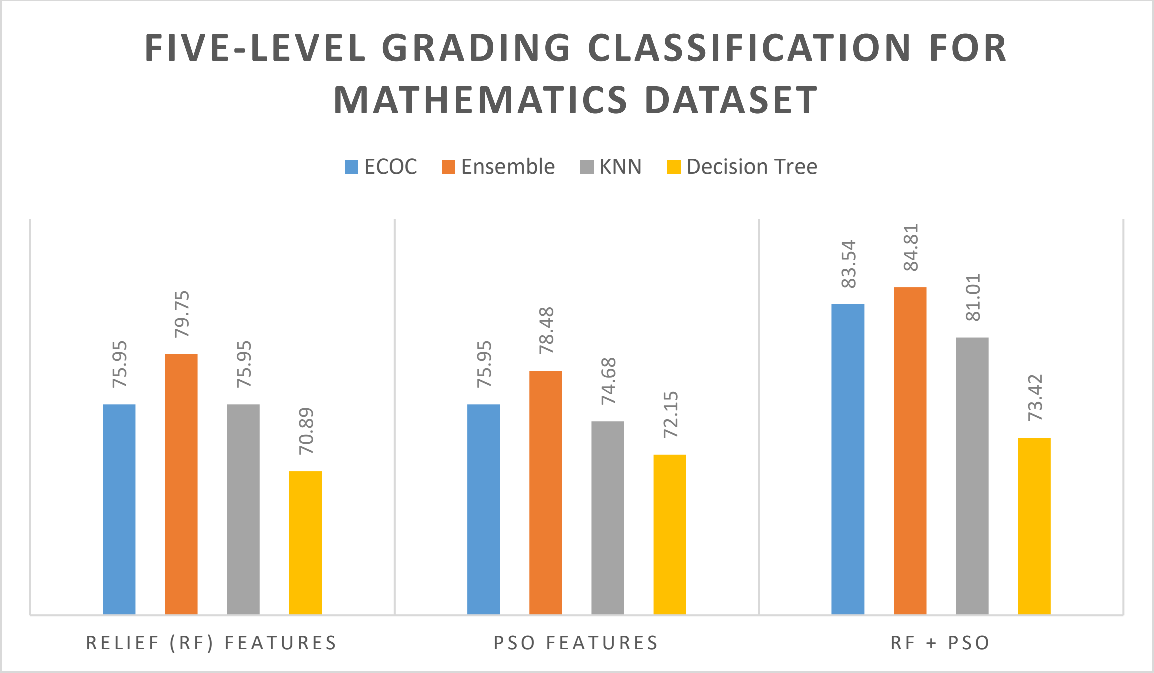

Figure.4.3 Classification performance for five-level grading for Mathematics Dataset

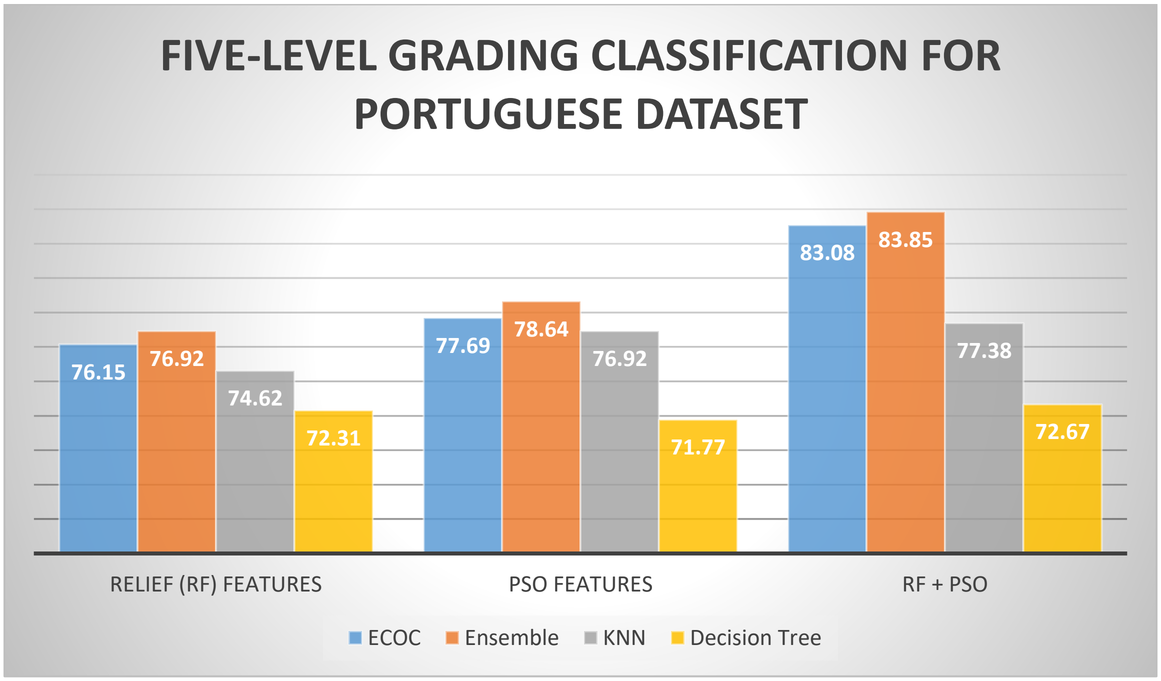

Figure.4.4 Classification performance for five-level grading for Portuguese Dataset

Table 4.17 presents a comparison of the performance of the proposed technique with related works that used the EuroStat dataset with respect to five-level grading classification. The results obtained showed that the proposed technique achieved higher student prediction accuracy than related works based on Portuguese and Mathematics data set.

Table.4.16 Comparison of five-level grading performance with related work

		Mathematics	**Portuguese**
Techniques	**Dataset**	**Highest Obtained Accuracy (%)**	**Highest Obtained Accuracy (%)**
Salal*et al.*, (2019)	EuroStat		76.73
Ünal, (2020)	EuroStat	79.49	77.20
Proposed Technique	EuroStat	**84.81**	**83.85**

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study developed a cascade bi-level feature selection technique for predicting students' academic performance. The first objective of developing a cascade bi-level feature selection technique for improved prediction accuracy was achieved using Relief filter-based algorithm and Particle Swarm Optimization (PSO) algorithm. First the relief algorithm was used to select features based on their relevance to the target class. These selected features were feed as input to the PSO. The PSO then optimally selects the subset of the selected features based on the particle fitness. The features that influence students' performance were selected using the cascaded bi-level feature selection technique achieved in the first objectives. These selected features were analyzed using Error-Correcting Output Code (ECOC), ensemble, Decision Tree and K-Nearest Neighbour (KNN) machine learning models, thus achieving the second objective.

The cascaded bi-level feature selection technique was evaluated against single-level feature selection techniques and against related works which achieved the third objective. The accuracy performance metric was used to perform this assessment. The proposed cascaded bi-level feature selection technique obtained an accuracy of 94.94% for Mathematics data set and 96.15% for Portuguese data set using the binary-level grading version of the data set. The cascaded bi-level feature selection technique also obtained accuracy 84.81% for Mathematics data set and 83.85% for Portuguese data set using the five-level grading version of the data set. The results indicate the effectiveness of the cascaded bi-level feature selection technique in achieving an improved student performance prediction as it selects the best sub-features.

5.2 Contribution to Knowledge

This study contributed to knowledge in the following ways: firstly, the study developed an effective cascaded bi-level approach for feature selection. Secondly, the study established the efficiency of bi-level feature selectors in selecting the optimal features for student academic performance prediction.

5.3Recommendations

This study utilized Relief a filter-based technique and Particle swarm optimization a wrapper technique for feature selection. For future work other filter and wrapper-based feature selection techniques can be utilized, which can provide an insight on which filter and/or wrapper-based selection techniques produces better results when combined. In this study, the bi-level selection approach was considered. It is recommended that further research should explore multiple-level techniques for feature selection.

REFERENCES

Abu-Zohair, L. M. (2019). Prediction of Student’s performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1), 27. https://doi.org/10.1186/s41239-019-0160-3

Adán-Coello, J. M., &Tobar, C. M. (2016). Using Collaborative Filtering Algorithms for Predicting Student Performance. In A. Kő& E. Francesconi (Eds.), *Electronic Government and the Information Systems Perspective* (Vol. 9831, pp. 206–218). Springer International Publishing. https://doi.org/10.1007/978-3-319-44159-7_15

Adejo, O., & Connolly, T. (2017). An Integrated System Framework for Predicting Students’ Academic Performance in Higher Educational Institutions. *International Journal of Computer Science and Information Technology*, *9*(3), 149–157. https://doi.org/10.5121/ijcsit.2017.93013

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61–75. https://doi.org/10.1108/JARHE-09-2017-0113

Agyapong, K. B., Hayfron-Acquah, D. J. B., & Michael, D. (2016). An Overview of Data Mining Models (Descriptive and Predictive).*International Journal of Software & Hardware Research in Engineering.4*(5), 53-61.

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students’ academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*, 6415–6426. https://doi.org/10.12988/ams.2015.53289

Algarni, A. (2016). Data Mining in Education. *International Journal of Advanced Computer Science and Applications*, *7*(6), 456–461.

Almasri, A., Celebi, E., &Alkhawaldeh, R. S. (2019). EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance. *Scientific Programming*, *2019*, 1–13. https://doi.org/10.1155/2019/3610248

Amoo, M. A., Alaba, O. B., & Usman, O. L. (2018). Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach. *International Journal of Science and Technology Education Research*, *9*(1), 1–8. https://doi.org/10.5897/IJSTER2017.0415

Amrieh, E. A., Hamtini, T., &Aljarah, I. (2016). Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, *9*(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Anuradha, C., &Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, *8*(15). https://doi.org/10.17485/ijst/2015/v8i15/74555

Arade, S. P., &Patil, J. K. (2017). Comparative Study of Diabetic Retinopathy Using K-NN and Bayesian Classifier. *International Journal of Innovations in Engineering Research and Technology*,*4*(5), 55–61.

Armano, G., Chira, C., &Hatami, N. (2013). Error-Correcting Output Codes for Multi-Label Text Categorization.*International Journal of Refrigeration*. 12 (4), 26-37.

Badr, G., Algobail, A., Almutairi, H., &Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. *Procedia Computer Science*, *82*, 80–89. https://doi.org/10.1016/j.procs.2016.04.012

Bilal Zorić, A. (2020). Benefits of Educational Data Mining. *Journal of International Business Research and Marketing*, *6*(1), 12–16. https://doi.org/10.18775/jibrm.1849-8558.2015.61.3002

Blake, R. H., &Mangiameli, P. (2011). The effects and interactions of data quality and problem complexity on data mining.*Journal of Data and Information Quality*2(2), 1-28.https://doi.org/10.1145/1891879.1891881

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., &Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42–54. https://doi.org/10.1016/j.compedu.2016.02.006

Chamatkar, A. J., &Butey, D. P. K. (2014). Importance of Data Mining with Different Types of Data Applications and Challenging Areas.*International Journal of Engineering Research and Applications. 4*(5), 38-41.

Colonna, L. (2013). A Taxonomy and Classification of Data Mining.*Science and Technology Law Review*.6(4), 309-312.https://doi.org/10.4018/978-1-7998-2460-2.ch026

Cortez, P. (2008). Student Performance Data Set. *UCI Repository*.Retrieved August March 12, 2021, from https://archive.ics.uci.edu/ml/datasets/student+performance

Cortez, P., & Silva, A. (2008). Using data mining to Predict Secondary School Student Performance.*In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto.* 9, 5-12.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., &Alowibdi, J. S. (2017). Predicting Student Performance using Advanced Learning Analytics. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 415–421. https://doi.org/10.1145/3041021.3054164

David Kolo, K., A. Adepoju, S., &KoloAlhassan, J. (2015). A Decision Tree Approach for Predicting Students Academic Performance. *International Journal of Education and Management Engineering*, *5*(5), 12–19. https://doi.org/10.5815/ijeme.2015.05.02

Deshpande, S. P., &Thakare, V. M. (2010). Data Mining System and Applications: A Review. *International Journal of Distributed and Parallel Systems*, *1*(1), 32–44. https://doi.org/10.5121/ijdps.2010.1103

Dietterich, T. G., &Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, *2*, 263–286. https://doi.org/10.1613/jair.105

Durgabai, R. P. L., & Ravi Bhushan, Y. (2014). Feature Selection using Relief Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 8215–8218. https://doi.org/10.17148/IJARCCE.2014.31031

Ermiyas, B. B., &Gobena, F. A. (2017). Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(2), 46–50. https://doi.org/10.23956/ijarcsse/V7I2/01219

Escalera, S., Pujol, O., Radeva, P., &Ivanova, P. (2010). Error-Correcting Ouput Codes Library. *Journal of Machine Learning Research*.11 (2),1-4.

Fariba, T. B. (2013). Academic Performance of Virtual Students based on their Personality Traits, Learning Styles and Psychological Well Being: A Prediction. *Procedia - Social and Behavioral Sciences*, *84*, 112–116. https://doi.org/10.1016/j.sbspro.2013.06.519

Feng, J. (2019).Predicting Students' Academic Performance with Decision Tree and Neural Network. *University of Central Florida*. 44.

Francis, B. K., &Babu, S. S. (2019). Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *Journal of Medical Systems*, *43*(6), 162. https://doi.org/10.1007/s10916-019-1295-4

Gu, X., Akoglu, L., &Rinaldo, A. (2019). Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. *33rd Conference on Neural Information Processing Systems*, 1-11.

Hira, Z. M., &Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, *2015*, 1–13. https://doi.org/10.1155/2015/198363

Hoe, A. C. K., Ahmad, M. S., Hooi, T. C., Shanmugam, M., Gunasekaran, S. S., Cob, Z. C., &Ramasamy, A. (2013). Analyzing students records to identify patterns of students' performance. *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, 544–547. https://doi.org/10.1109/ICRIIS.2013.6716767

Hussain, S., AbdulazizDahan, N., Ba-Alwi, F. M., &Ribata, N. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Ikbal, S., Tamhane, A., Sengupta, B., Chetlur, M., Ghosh, S., & Appleton, J. (2015). On early prediction of risks in academic performance for students. *IBM Journal of Research and Development*, *59*(6), 5:1-5:14. https://doi.org/10.1147/JRD.2015.2458631

Imani, M. B., Keyvanpour, M. R., &Azmi, R. (2013). A Novel Embedded Feature Selection Method: A Comparative Study in the Application of Text Categorization.*Applied Artificial Intelligence*, *27*(5), 408–427. https://doi.org/10.1080/08839514.2013.774211

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(14), 92. https://doi.org/10.3991/ijet.v14i14.10310

Iyanda, A. R., D. Ninan, O., O. Ajayi, A., & G. Anyabolu, O. (2018). Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models. *International Journal of Modern Education and Computer Science*, *10*(6), 1–9. https://doi.org/10.5815/ijmecs.2018.06.01

Jembere, E., Rawatlal, R., &Pillay, A. W. (2017). Matrix Factorisation for Predicting Student Performance. *2017 7th World Engineering Education Forum (WEEF)*, 513–518. https://doi.org/10.1109/WEEF.2017.8467150

Kalejaye, B. A., Folorunso, O., & Usman, O. L. (2015). Predicting Students' Grade Scores Using Training Functions of Artificial Neural Network. *Journal of Natural Sciences, Engineering and Technology*, 14 (2), 1-23.

Kataria, A., & Singh, M. D. (2013). A Review of Data Classification Using K-Nearest Neighbour Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 354–360.

Kirsal Ever, Y., &Dimililer, K. (2018). The effectiveness of a new classification system in higher education as a new e-learning tool. *Quality & Quantity*, *52*(S1), 573–582. https://doi.org/10.1007/s11135-017-0636-y

Kumari, B., &Swarnkar, T. (2011). Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review.*International Journal of Computer Science and Information Technologies,2(3),*1048-1053.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 982. https://doi.org/10.1007/s42452-019-0884-7

Magbag, A., & Jr, R. R. (2020). Prediction of College Academic Performance Of Senior High School Graduates Using Classification Techniques.*International Journal of Scientific & Technology Research 9*(4), 2104-2109.

Mala Sari Rochman, E., Rachmad, A., &Damayanti, F. (2018). Predicting the Final result of Student National Test with Extreme Learning Machine. *PancaranPendidikan*, *7*(2). https://doi.org/10.25037/pancaran.v7i1.159

Manjarres, A. V., Sandoval, L. G. M., &Suárez, M. J. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*, 2(2)*, 33*. http://greav.ub.edu/der/

Meyer-Baese, A., &Schmid, V. (2014). Feature Selection and Extraction. In *Pattern Recognition and Signal Analysis in Medical Imaging*, 21–69. Elsevier. https://doi.org/10.1016/B978-0-12-409545-8.00002-9

Mohamad, S. K., &Tasir, Z. (2013). Educational Data Mining: A Review. *Procedia - Social and Behavioral Sciences*, *97*, 320–324. https://doi.org/10.1016/j.sbspro.2013.10.240

Mueen, A., Manzoor, U., & Zafar, B. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36–42. https://doi.org/10.5815/ijmecs.2016.11.05

Naheed, N., Shaheen, M., Ali Khan, S., Alawairdhi, M., &Attique Khan, M. (2020). Importance of Features Selection, Attributes Selection,Challenges and Future Directions for Medical Imaging Data:A Review. *Computer Modeling in Engineering & Sciences*, *125*(1), 315–344. https://doi.org/10.32604/cmes.2020.011380

Olalekan, A. M., Egwuche, O. S., &Olatunji, S. O. (2020). Performance Evaluation of Machine Learning Techniques for Prediction of Graduating Students In Tertiary Institution. *2020*

International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), 1–7. https://doi.org/10.1109/ICMCECS47690.2020.240888

Olaniyi, A. S., Kayode, S. Y., Abiola, H. M., Tosin, S.-I. T., &Babatunde, A. N. (2018). Student's Performance Analysis Using Decision Tree Algorithms. *International Journal of Computational Engineering Research*, *08*(9), 8.

Olson, D. L. (2017). *Descriptive Data Mining*. In: Descriptive Data Mining. Computational Risk Management. Springer, Singapore. https://doi.org/10.1007/978-981-10-3340-7

Oracle. (2020). Data Mining Concepts. *Oracle Help Center*. Retrieved March30, 2021, fromhttps://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *North Carolina State University*, *15*(12), 10.

Pandey, M., &Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366. https://doi.org/10.1016/j.pisc.2016.04.076

Patel, N., & Singh, D. (2015). An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor. *International Journal of Computer Applications*, *111*(10), 22–26. https://doi.org/10.5120/19575-1376

Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization: An overview. *Swarm Intelligence, 1*(1), 33–57. https://doi.org/10.1007/s11721-007-0002-0

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications, 175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

Rajagopal, S. (2011). *Customer Data Clustering using Data Mining Technique*.International Journal of Database Management Systems,*3*(4), 1-11. https://doi.org/10.5121/ijdms.2011.3401

Rodrigues, M. W., Isotani, S., &Zárate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics, 35*(6), 1701–1717. https://doi.org/10.1016/j.tele.2018.04.015

Romalt, A. A., & Kumar, M. S. (2020). An Analysis of Feature Selection Methods, Clustering and Classification used in Heart Disease Prediction-A Machine Learning Approach. *Journal of Critical Reviews, 7*(06), 138–142. https://doi.org/10.31838/jcr.07.06.27

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68,* 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rosario, S. F. (2015). RELIEF: Feature Selection Approach. *International Journal of Innovative Research and Development, 4*(11), 219-224.

Sahu, B., & Mishra, D. (2012). A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data. *Procedia Engineering, 38,* 27–31. https://doi.org/10.1016/j.proeng.2012.06.005

Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). *Educational Data Mining: Student Performance Prediction in Academic*. *8*(4), 6.

Salvi, A., Medha, K., Shaikh, S., &Kokani, S. (2018). Prediction and Evaluation of Students Academic Performance using Fuzzy Logic. *International Research Journal of Engineering and Technology*, *5*(2). www.irjet.net

Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing* [Bachelor Degree]. Kth Royal Institute Of Technology School Of Electrical Engineering And Computer Science.

Sekeroglu, B., Dimililer, K., &Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. *Proceedings of the 2019 8th International Conference on Educational and Information Technology - ICEIT 2019*, 7–11. https://doi.org/10.1145/3318396.3318419

Sembiring, R. W., Zain, J. M., &Embong, A. (2011). Dimension Reduction of Health Data Clustering. *International Journal on New Computer Architectures and Their Applications*1(4) 1041-1050.

Sengupta, S., Basak, S., & Ii, R. A. P. (2019). *Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives*. Machine Learning and Knowledge Extraction, 1(1), 157–191.https://doi.org/10.3390/make1010010

Shah, M. B., Kaistha, M., & Gupta, Y. (2019). Student Performance Assessment and Prediction System using Machine Learning. *2019 4th International Conference on Information Systems and Computer Networks (ISCON),* 386–390. https://doi.org/10.1109/ISCON47742.2019.9036250

Silva, C., & Fonseca, J. (2017). Educational Data Mining: A Literature Review. In Á. Rocha, M. Serrhini, & C. Felgueiras (Eds.), *Europe and MENA Cooperation Advances in Information and Communication Technologies* (Vol. 520, pp. 87–94). Springer International Publishing. https://doi.org/10.1007/978-3-319-46568-5_9

Singh, W., & Kaur, P. (2016). Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance. *International Journal of Advanced Research in Computer Science*, *7*(6), 31–36.

Stapel, M., Zheng, Z., &Pinkwart, N. (2015). An Ensemble Method to Predict Student Performance in an Online Math Learning Environment. *Proceedings of the 9th International Conference on Educational Data Mining*, 231-238.

Stattner, E., & Collard, M. (2015). Descriptive Modeling of Social Networks. *Procedia Computer Science*, *52*, 226–233. https://doi.org/10.1016/j.procs.2015.05.505

Susanto, A. (2019). Functions, Processes, Stages and Application of Data Mining.*International Journal of Scientific and Technology Research8*(7), 135-140.

Talukder, S. (2011). *Mathematical Modelling and Applications of Particle Swarm Optimization* [Master's Thesis]. Blekinge Institute of Technology, 65.

Tuaha, S., Siddiqui, I. F., & Ali Arain, Q. (2019). Analyzing Students' Academic Performance through Educational Data Mining. *3C Tecnología_Glosas de InnovaciónAplicadas a La Pyme*, 402–421. https://doi.org/10.17993/3ctecno.2019.specialissue2.402-421

Umer, R., Susnjak, T., Mathrani, A., &Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning*, *10*(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022

Ünal, F. (2020). Data Mining for Student Performance Prediction in Education. *IntechOpen*, 12. http://dx.doi.org/10.5772/intechopen.91449

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, *85*, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2018). Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining. *ArXiv:1711.08477 [Cs]*. http://arxiv.org/abs/1711.08477

VeeraManickam, M. R. M., Mohanapriya, M., Pandey, B. K., Akhade, S., Kale, S. A., Patil, R., &Vigneshwar, M. (2019). Map-Reduce framework based cluster architecture for academic student’s performance prediction using cumulative dragonfly based neural network. *Cluster Computing*, *22*(S1), 1259–1275. https://doi.org/10.1007/s10586-017-1553-5

Venkatesh, B., &Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/cait-2019-0001

Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). An Improved Feature Selection Based on Effective Range for Classification. *The Scientific World Journal*, *2014*, 1–8. https://doi.org/10.1155/2014/972125

Yildirim, P. (2015). Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *International Journal of Machine Learning and Computing*, *5*(4), 258–263. https://doi.org/10.7763/IJMLC.2015.V5.517

Zacharis, N. (2016). Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks. *International Journal of Artificial Intelligence & Applications*, *7*(5), 17–29. https://doi.org/10.5121/ijaia.2016.7502

Zaffar, M., Ahmed, M., Savita, K. S., &Sajjad, S. (2018). A Study of Feature Selection Algorithms for Predicting Students Academic Performance. *International Journal of Advanced Computer Science and Applications*, *9*(5). https://doi.org/10.14569/IJACSA.2018.090569

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218–218. https://doi.org/10.21037/atm.2016.03.37

APPENDIX A

Source Code

```
% Filter Feature Selection

%Model for Relief Filter-based selection
function model = jffs(type,feat,label,opts)
switch type
case'rf'     ; fun = @jReliefF;
end
tic;
model = fun(feat,label,opts);
% Computational time
t = toc;

model.t = t;
fprintf('\n Processing Time (s): %f % \n',t); fprintf('\n');
end

% Relief Filter-based selection Function
function RF = jReliefF(feat,label,opts)
% Parameter
K = 5;

ifisfield(opts,'Nf'), num_feat = opts.Nf; end
ifisfield(opts,'K'), K = opts.K; end

% Convert format to categorical
label         = categorical(label);
% Relief-F Algorithm
[idx, weight] = relieff(feat,label,K);
% Select features based on ranking
Sf            = idx(1:num_feat);
sFeat         = feat(:, Sf);
% Store results
RF.sf = Sf;
RF.ff = sFeat;
RF.nf = num_feat;
RF.f  = feat;
RF.l  = label;
RF.s  = weight;
end

%---Input-----------------------------------------------------------
```

```
% feat    : Feature vector matrix (Instances x Features)
% label   : Label matrix (Instances x 1)
% opts    : Parameter settings
% opts.Nf : Number of selected features


%---Output-----------------------------------------------------------
% FS    : Feature selection model
% FS.sf : Index of selected features
% FS.ff : Selected features
% FS.nf : Number of selected features
% FS.s  : Weight or score


%Relief-F
opts.K  = 3;     % number of nearest neighbors
opts.Nf = 20;    % select 20 features

% Perform feature selection
FS     = jffs('rf',Mathsintegerdataset,Mathscategorylabel,opts);

% Define index of selected features
sf_idx = FS.sf;

% Accuracy
kfold  = 5;
Acc    = mSVM(Mathsintegerdataset(:,sf_idx),Mathscategorylabel,kfold);


%-----------------------------------------------------------------%
% Particle Swarm Optimization (PSO) source %
%-----------------------------------------------------------------%


% Fitness Function
%-----------------------------------------------------------------%

function cost = jFitnessFunction(feat,label,X,HO)
if sum(X == 1) == 0
cost = 1;
else
cost = jwrapperKNN(feat(:, X == 1),label,HO);
end
end
```

```
function error = jwrapperKNN(sFeat,label,HO)
%---// Parameter setting for k-value of KNN //
k = 5;

xtrain = sFeat(HO.training == 1,:);
ytrain = label(HO.training == 1);
xvalid = sFeat(HO.test == 1,:);
yvalid = label(HO.test == 1);

Model     = fitcknn(xtrain,ytrain,'NumNeighbors',k);
pred      = predict(Model,xvalid);
num_valid = length(yvalid);
correct   = 0;
fori = 1:num_valid
ifisequal(yvalid(i),pred(i))
correct = correct + 1;
end
end
Acc   = correct / num_valid;
error = 1 - Acc;
end

%PSO Function
%---------------------------------------------------------------------%

function [sFeat,Sf,Nf,curve]=jPSO(feat,label,N,max_Iter,c1,c2,w,HO)
% Parameters
lb    = 0;
ub    = 1;
thres = 0.5;

% Objective function
fun = @jFitnessFunction;
% Number of dimensions
dim = size(feat,2);
% Initial
X   = zeros(N,dim);
V   = zeros(N,dim);
fori = 1:N
for d = 1:dim
X(i,d) = lb + (ub - lb) * rand();
end
end
% Fitness
fit  =zeros(1,N);
fitG = inf;
```

```
fori = 1:N
fit(i) = fun(feat,label,(X(i,:) >thres),HO);
% Gbest update
if fit(i) <fitG
Xgb  = X(i,:);
fitG = fit(i);
end
end
% PBest
Xpb  = X;
fitP = fit;
% Pre
curve = inf;
t = 1;
% Iterations
while t <= max_Iter
fori = 1:N
for d = 1:dim
    r1 = rand();
    r2 = rand();
% Velocity update
    V(i,d) = w * V(i,d) + c1 * r1 * (Xpb(i,d) - X(i,d)) + ...
      c2 * r2 * (Xgb(d) - X(i,d));
% Position update
X(i,d) = X(i,d) + V(i,d);
end
% Boundary
   XB = X(i,:); XB(XB >ub) = ub; XB(XB <lb) = lb;
X(i,:) = XB;
% Fitness
fit(i) = fun(feat,label,(X(i,:) >thres),HO);
% Pbest update
if fit(i) <fitP(i)
Xpb(i,:) = X(i,:);
fitP(i)  = fit(i);
end
% Gbest update
iffitP(i) <fitG
Xgb  =Xpb(i,:);
fitG = fitP(i);
end
end
curve(t) = fitG;
fprintf('\nIteration %d GBest (PSO)= %f',t,curve(t))
  t = t + 1;
end
```

```
% Select features based on selected index
Pos   = 1:dim;
Sf    = Pos((Xgb>thres) == 1);
sFeat = feat(:,Sf);
Nf    = length(Sf);
end

%Main Class
%---Inputs-----------------------------------------------------------
% feat     : feature vector
% label    : label vector
% N        : Number of particles
% max_Iter : Maximum number of iterations
% c1       : Cognitive factor
% c2       : Social factor
% w        : Inertia weight

%---Outputs----------------------------------------------------------
% sFeat    : Selected features
% Sf       : Selected feature index
% Nf       : Number of selected features
% curve    : Convergence curve
%--------------------------------------------------------------------


% Set 20% data as validation set
ho = 0.2;
% Hold-out method
HO = cvpartition(Mathscategorylabel,'HoldOut',ho);

% Parameter setting
N        = 10;
max_Iter = 100;
c1       = 2;     % cognitive factor
c2       = 2;     % social factor
w        = 1;     % inertia weight

% Particle Swarm Optimization
[sFeat,Sf,Nf,curve] = jPSO(FS.ff,Mathscategorylabel,N,max_Iter,c1,c2,w,HO);

% Plot convergence curve
plot(1:max_Iter,curve);
xlabel('Number of iterations');
ylabel('Fitness Value');
title('PSO'); grid on;
```