**TABLE OF CONTENTS**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CHAPTER** | | | **TITLE** | | | | **PAGE** |
|  | | | **ABSTRACT** | | | | i |
|  | | | **LIST OF TABLES** | | | | iii |
|  | | | **LIST OF FIGURES** | | | | iv |
|  | | | **LIST OF ABBREVIATIONS** | | | | vi |
|  | **1.0** |  | **INTRODUCTION** | | | |  |
|  |  |  | 1.1 | OVERVIEW | | | 1 |
|  |  |  | 1.2 | DATA MINING | | | 1 |
|  |  |  | 1.3 | EDUCATIONAL DATA MINING (EDM) | | | 3 |
|  |  |  |  | 1.3.1 | Brief History of EDM | | 6 |
|  |  |  |  | 1.3.2 | Lifecycle/ Process involved in EDM | | 7 |
|  |  |  |  | 1.3.3 | An EDM Example | | 9 |
|  |  |  |  | 1.3.4 | EDM in HE | | 11 |
|  |  |  | 1.4 | APPLICATIONS OF EDM IN HIGHER EDUCATION | | | 13 |
|  |  |  |  | 1.4.1 | Prediction the student’s performance | | 15 |
|  |  |  |  | 1.4.2 | Detection Behavior of students | | 15 |
|  |  |  |  | 1.4.3 | Profiling and Clustering students | | 16 |
|  |  |  |  | 1.4.4 | Social networking analysis | | 16 |
|  |  |  | 1.5 | EDM IN DECISION SUPPORT SYSTEM | | | 17 |
|  |  |  |  | 1.5.1 | Providing reports | | 17 |
|  |  |  |  | 1.5.2 | Raising alarms for stake holders | | 18 |
|  |  |  |  | 1.5.3 | Educational Scheduling process | | 18 |
|  |  |  |  | 1.5.4 | Generation of courseware | | 19 |
|  |  |  |  | 1.5.5 | Concept Map Creation | | 19 |
|  |  |  |  | 1.5.6 | Providing recommendations | | 19 |
|  |  |  | 1.6 | IMPACT OF DIFFERENT FACTORS ON STUDENT’S PERFORMANCE | | | 20 |
|  |  |  | 1.7 | METHODS INVOLVED IN EDM PROCESS | | | 22 |
|  |  |  |  | 1.7.1 | Feature Selection | | 22 |
|  |  |  |  | 1.7.2 | Clustering Process | | 24 |
|  |  |  |  | 1.7.3 | Classification Process | | 25 |
|  |  |  | 1.8 | MACHINE LEARNING | | | 27 |
|  |  |  | 1.9 | METAHEURISITCS | | | 29 |
|  |  |  | 1.10 | MOTIVATION AND RESEARCH CONTRIBUTIONS | | | 31 |
|  |  |  | 1.11 | THESIS ORGANIZATION | | | 32 |
|  | **2.0** |  | **LITERATURE SURVEY** | | | |  |
|  |  |  | 2.1 | OVERVIEW | | | 34 |
|  |  |  | 2.2 | REVIEW OF EXISTING EDM TECHNIQUES FOR HE STUDENTS | | | 34 |
|  |  |  | 2.3 | REVIEW OF EXISTING EDM TECHNIQUES FOR STUDENT PERFORMANCE ASSESSMENT | | | 36 |
|  |  |  |  | 2.3.1 | Other Methods | | 39 |
|  |  |  | 2.4 | REVIEW OF FS BASED EDM TECHNIQUES | | | 42 |
|  |  |  | 2.5 | REVIEW OF CLUSTERING BASED EDM TECHNIQUES | | | 45 |
|  |  |  | 2.6 | RESEARCH GAP | | | 48 |
|  |  |  | 2.7 | RESEARCH OBJECTVES | | | 48 |
|  | **3.0** |  | **MATERIALS AND METHODS** | | | |  |
|  |  |  | 3.1 | PREDICTIVE ANALYSIS OF STUDENTS ACADEMIC PERFORMANCE USING MACHINE LEARNING MODELS | | | 50 |
|  |  |  | 3.2 | PROPOSED HLVQ ALGORITHM FOR STUDENT’S PERFORMANCE AND EMPLOYABILITY ASSESSMENT | | | 50 |
|  |  |  |  | 3.2.1 | Preprocessing | | 51 |
|  |  |  |  | 3.2.2 | Linear Vector Quantization (LVQ) | | 52 |
|  |  |  |  | 3.2.3 | AdaBoost | | 54 |
|  |  |  |  | 3.2.4 | Working Principle of HLVQ Model | | 55 |
|  |  |  |  | 3.2.5 | Pseudo Code of HLVQ Model | | 56 |
|  |  |  | 3.3 | ACO-LR BASED FEATURE SELECTION AND CLASSIFICATION MODEL | | | 58 |
|  |  |  |  | 3.3.1 | Preprocessing | | 59 |
|  |  |  |  | 3.3.2 | ACO based Feature Selection | | 59 |
|  |  |  |  |  | 3.3.2.1 | Need of Feature Selection | 59 |
|  |  |  |  |  | 3.3.2.2 | Overview of ACO algorithm | 60 |
|  |  |  |  |  | 3.3.2.3 | Process involved in ACO-FS algorithm | 60 |
|  |  |  |  | 3.3.3 | LR based Classification | | 64 |
|  |  |  |  | 3.3.4 | Pseudo code of ACO-FS algorithm | | 66 |
|  |  |  | 3.4 | AN ENSEMBLE OF DENSITY BASED CLUSTERING WITH OPITMAL MULTILAYER PERCEPTRON (DBC-OMLP) MODEL | | | 67 |
|  |  |  |  | 3.4.1 | Preprocessing | | 67 |
|  |  |  |  | 3.4.2 | DBC based Clustering | | 68 |
|  |  |  |  |  | 3.4.2.1 | Overview of DBC | 68 |
|  |  |  |  |  | 3.4.2.2 | DBC based Clustering Process | 69 |
|  |  |  |  |  | 3.4.2.3 | Pseudocode of DBC model | 71 |
|  |  |  |  | 3.4.3 | OMLP based Classification | | 72 |
|  |  |  |  |  | 3.4.3.1 | Overview of MLP | 72 |
|  |  |  |  |  | 3.4.3.2 | Overview of BBO | 74 |
|  |  |  |  |  | 3.4.3.3 | FSCBBO algorithm for training MLP | 75 |
|  |  |  |  | 3.4.4 | Pseudocode of OMLP Model | | 77 |
|  | **4.0** |  | **RESULT AND DISCUSSION** | | | |  |
|  |  |  | 4.1 | OVERVIEW | | | 80 |
|  |  |  | 4.2 | PERFORMANCE METRICS | | | 80 |
|  |  |  | 4.3 | PERFORMANCE EVALUATION OF HLVQ MODEL | | | 81 |
|  |  |  |  | 4.3.1 | Implementation Data | | 81 |
|  |  |  |  | 4.3.2 | Predictive Results analysis of Graduated Students | | 81 |
|  |  |  |  | 4.3.3 | Predictive Results analysis of Placed Students | | 83 |
|  |  |  |  | 4.3.4 | Classification results analysis of LVQ and HLVQ models | | 83 |
|  |  |  |  | 4.3.5 | Complexity analysis of LVQ and HLVQ models | | 84 |
|  |  |  | 4.4 | PERFORMANCE EVALUATION OF ACO-LR MODEL | | | 85 |
|  |  |  |  | 4.4.1 | Dataset used | | 85 |
|  |  |  |  | 4.4.2 | Results analysis of ACO-LR based Feature Selection Process | | 88 |
|  |  |  |  | 4.4.3 | Results analysis of ACO-LR based Classification Process | | 90 |
|  |  |  | 4.5 | PERFORMANCE EVALUATION OF DBC-OMLP MODEL | | | 95 |
|  |  |  |  | 4.5.1 | Dataset used | | 95 |
|  |  |  |  | 4.5.2 | Results analysis of DBC-OMLP based Clustering Process | | 95 |
|  |  |  |  | 4.5.3 | Results analysis of DBC-OMLP based Classification Process | | 97 |
|  |  |  | 4.6 | CONCLUSION | | | 102 |
|  | **5** |  | **SUMMARY** | | | | **103** |
|  | **6** |  | **FUTURE DIRECTIONS** | | | | **105** |
|  | **7** |  | **REFERENCES** | | | | **106** |
|  |  |  |  | | | |  |
| **ANNEXURE** | | |  | | | |  |
|  | **1A** |  | **LIST OF PUBLICATIONS** | | | |  |
|  | **2B** |  | **REPRINTS OF PUBLICATIONS** | | | |  |

# ABSTRACT

**Educational Data Mining** (EDM) is an emerging discipline, related to the application of data mining (DM), machine learning (ML), and statistics to the details produced from the educational environment. It is used for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to understand the students effectively. Higher educational sectors are also started to use the significance of EDM models on the learning task and output for moving towards the university to the next level. Particularly, EDM assists educational institutions and educational plan makers in offering solutions to improve the effectiveness and quality of teaching and learning. The assessment of student’s performance in educational institutions indicates the level of efforts taken by educational institutions should be done for enhancing the poor or average learner. The significance of applying EDM models, make use of the previous data of the students for predicting the unseen or upcoming performance of the students. This idea has attracted several researchers to develop classification models to predict the unknown labels of future instances.

Several research people and educational institutions started to get attracted to the domain of predicting the performance of the student to classify the educational level of student performance. Though the educational sector uses several techniques for extracting useful information on the features of students undertake the learning process, it is needed to develop a student performance assessment model to assist the students as well as faculties to improve their performance to the next stage. This research work majorly concentrates to examine and identify useful rules and patterns to motivate the students for handling their education as well as carrier in a good manner, also to improve the and functions of academics to supervise the policies for student's benefit. Here, EDM is employed to identify the effective data which helps in applying the active learning in technical applications. Initially a review of different DM, machine learning (ML), and meta heuristic models to assess the performance of the students, then hybridization of linear vector quantization (HLVQ) model for predicting the educational results and employability chances of the students is designed and developed an Ant Colony Optimization (ACO) with feature subset selection and logistic regression (LR) model for classifying educational DM .Finally ensemble of density based clustering (DBC) with optimal multilayer perceptron (OMLP) based classification model, abbreviated as DBC-OMLP for effective EDM.

In order to validate the experimental results analysis of the proposed models, a series of simulations take place on the benchmark Student Performance assessment dataset. The results are validated interms of different measures namely precision, recall, accuracy, F-score, and kappa. The attained comparative results analysis stated the superior performance of the proposed models over the compared methods under different aspects. Therefore, the proposed models can be employed as an appropriate tool for mining educational data to achieve effective performance assessment of the students.

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE** | **TITLE** | **PAGE** |
| 3.1 | Data Transformed Values in the dataset | 52 |
| 4.1 | Sample visualization of the training dataset | 82 |
| 4.2 | Details of the dataset | 85 |
| 4.3 | Class Variable Distribution (0-20) | 86 |
| 4.4 | Analysis of FS results | 89 |
| 4.5 | Confusion matrix of Proposed ACO-LR with Existing Methods | 91 |
| 4.6 | Comparative analysis of diverse models under several measures | 92 |
| 4.7 | Details of the clustered dataset | 96 |
| 4.8 | Confusion Matrix of DBC-OMLP model with Existing Methods | 98 |
| 4.9 | Performance of Proposed Method | 99 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE** | **TITLE** | **PAGE** |
| 1.1 | An illustration of DM use in HE | 5 |
| 1.2 | Workflow of EDM | 8 |
| 1.3 | Main areas involved in Educational data mining | 11 |
| 1.4 | Classification of different EDM Applications | 14 |
| 1.5 | Process involved in data classification | 26 |
| 1.6 | Types of machine learning | 28 |
| 1.7 | Overview of metaheuristics | 30 |
| 3.1 | Working Process of HLVQ based Predictive model | 51 |
| 3.2 | Block diagram of ACO-LR model | 58 |
| 3.3 | Flowchart of ACO-FS technique | 63 |
| 3.4 | Process involved in LR based classification | 64 |
| 3.5 | Sample Representation of LR | 65 |
| 3.6 | Overall process of DBC-OMLP model | 68 |
| 3.7 | Flowchart of DBC technique | 70 |
| 3.8 | An MLP with one hidden layer | 74 |
| 4.1 | Predictive analysis of the students to be graduated | 82 |
| 4.2 | % of students predicted to be placed | 83 |
| 4.3 | Performance Analysis of LVQ and HLVQ | 84 |
| 4.4 | Time and Space Complexities of LVQ and HLVQ | 85 |
| 4.5 | List of features | 87 |
| 4.6 | Attributes Description | 88 |
| 4.7 | Feature selection results in ACO-FS model | 89 |
| 4.8 | Confusion Matrix of LR and ACO-LR | 90 |
| 4.9 | Confusion Matrix analysis of different models | 91 |
| 4.10 | Precision and recall analysis | 93 |
| 4.11 | F-measure and kappa analysis | 93 |
| 4.12 | Accuracy analysis of various methods | 94 |
| 4.13 | Visualization of Clustering Results using DBC | 96 |
| 4.14 | Confusion Matrix of OMLP and DBC-OMLP | 97 |
| 4.15 | Confusion matrix | 98 |
| 4.16 | Precision and recall analysis of diverse models | 100 |
| 4.17 | F-measure and Kappa analysis of diverse models | 100 |
| 4.18 | Accuracy analysis of diverse models | 101 |

**LIST OF ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| **ABBREVIATION** |  | **EXPANSION** |
| AAAI | - | Association for the Advancement of Artificial Intelligence |
| ACO | - | Ant Colony Optimization |
| AI | - | Artificial Intelligence |
| ANN | - | Artificial Neural Network |
| BBO | - | Biogeography-Based Optimization |
| BP | - | Back Propagation |
| CAL | - | Computer-Assisted Learning |
| CAR | - | Coursework Assessment Ratio |
| CBT | - | Computer-Based Training |
| CGPA | - | Cumulative Grade Point Average |
| CMS | - | Content Management System |
| CO | - | Combinatorial Optimization |
| CRISP | - | Cross-Industry Standard Process for Data Mining |
| CS | - | Cuckoo Search |
| DA | - | Discriminant Analysis |
| DBC | - | Density Based Clustering |
| DBN | - | Deep Belief Network |
| DM | - | Data Mining |
| DSS | - | Decision Support System |
| DT | - | Decision Tree |
| EA | - | Evolutionary Algorithm |
| EDC | - | Educational Data Clustering |
| EDM | - | Educational Data Mining |
| EM | - | Expectation-Maximisation |
| **ABBREVIATION** |  | **EXPANSION** |
| FA | - | Firefly Algorithm |
| FCBF | - | Fast Correlation-Based Filter |
| FCM | - | Fuzzy C-Means |
| FN | - | False Negative |
| FP | - | False Positive |
| FS | - | Feature Selection |
| FSC | - | Fitness Scaling Chaotic |
| FSCBBO | - | Fitness Scaling Chaotic Biogeography-Based Optimization |
| GA | - | Genetic Algorithm |
| GPA | - | Grade Point Average |
| HE | - | Higher Education |
| HEI | - | Higher Education Institution |
| HLVQ | - | Hybridization of Linear Vector Quantization |
| ITS | - | Intelligent Tutoring System |
| KNIME | - | Konstanz Information Miner |
| KNN | - | K-Nearest Neighbor |
| LA | - | Learning Analytics |
| LMS | - | Learning Management System |
| LR | - | Logistic Regression |
| LVQ | - | Linear Vector Quantization |
| ML | - | Machine Learning |
| MLP | - | Multilayer Perception |
| MOOC | - | Massive Open Online Courses |
| NB | - | Naïve Bayes |
| NN | - | Neural Network |
| **ABBREVIATION** |  | **EXPANSION** |
| OMLP | - | Optimal Multilayer Perceptron |
| PCA | - | Principal Component Analysis |
| PPL | - | Placement Performance Level |
| PSC | - | Particle Swarm Classification |
| PSO | - | Particle Swarm Optimization |
| RA | - | Rocchio Algorithm |
| RAR | - | Relational Association Rule |
| RecPart | - | Recursive Partitioning |
| REPTree | - | Reduced Error Pruning Decision Tree |
| RF | - | Random Forest |
| RR | - | Robust Regression |
| RS | - | Recommender System |
| SA | - | Simulated Annealing |
| SAP | - | Student Academic Performance |
| SI | - | Swarm Intelligence |
| SIV | - | Suitability Index Variable |
| SMO | - | Sequential Minimal Optimization |
| SMOTE | - | Synthetic Minority Oversampling Technique |
| SNA | - | Social Network Analysis |
| SOM | - | Self-Organizing Maps |
| SPL | - | Student Performance Level |
| SVM | - | Support Vector Machine |
| SVR | - | Support Vector Regression |
| TN | - | True Negative |
| TP | - | True Positive |
| **ABBREVIATION** |  | **EXPANSION** |
| TS | - | Tabu Search |
| TSP | - | Travelling Salesman Problem |
| UL | - | Unsupervised Learning |
| VLE | - | Virtual Learning Environment |
| WebCT | - | WEB-Based Course Development Tools |
| WEKA | - | Waikato Environment for Knowledge Analysis |
| WWW | - | World Wide Web |