International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

Weighted Residual Target Proximity Kernel Pursuit Regression based Students Admission Prediction for Higher Education

P. Christy Jeyakumar [1]
[1] Ph.D. Research Scholar,
Department of Computer Applications,
Dr. M.G.R. Educational and Research Institute, Chennai

Dr. S. Ismail Kalilulah [2]
[2] Associate Professor, Department of Computer Science and Engineering,
Dr. M.G.R. Educational and Research Institute, Chennai

Dr. V. N. Rajavarman [3]
[3] Professor, Department of Computer Science and Engineering,
Dr. M.G.R. Educational and Research Institute, Chennai

ABSTRACT

Education plays a significant role in providing individuals with the knowledge, skills, and tools needed for personal as well as academic growth. Due to the increasing number of higher education graduates, student admissions process is essential for selecting qualified candidates for admission in a universities or colleges. An admissions system with suitable and reliable criteria is important to select students who performing well academically as well as other activities at institutions. Therefore, each university or college needs to use the best possible techniques for analyzing the history of a student's academic performance and other extracurricular activities before admitting them. Education Data Mining (EDM) involves the application of data mining techniques to large educational databases with the aim of discovering useful information. Several machine learning techniques have been developed in this area, but there are issues related to time efficiency and errors in prediction of admissibility for higher education. To address the aforementioned challenge, a novel technique named Weighted Residual Target Proximity Kernel Pursuit Regression (WRTPKPR) has been developed. This technique aims for the accurate prediction of graduate admissions with minimal error by mapping the course to the students based on their interests and CGPA secured. The proposed WRTPKPR technique includes three major phases namely data acquisition, preprocessing, and feature selection for accurate predictive analytics. The WRTPKPR technique initiates by collecting information from the dataset during the data acquisition phase. Following data acquisition, the WRTPKPR technique undergoes data preprocessing to transform the input data into a suitable format for accurately predicting whether the student is admissible or not. Two key processes are conducted in the data preprocessing phase, namely, missing data imputation and outlier data detection. In the initial step, the Horvitz–Thompson Weighted imputation method is applied to generate missing data points based on other known data points in the dataset. In the second step, an outlier detection method based on the maximum normalized residual test is employed to identify data points that significantly deviate from the rest of the data point in the dataset. With the preprocessed dataset, the target feature selection process is conducted by applying Kernel Cook's Proximity Projection Pursuit Regression. Based on the selected target features, accurate admission predictions are made for higher education graduates with minimal time consumption. Experimental evaluation considers factors such as admission prediction accuracy, precision, recall, F1-score and admission prediction time. The results demonstrate that the proposed WRTPKPR technique achieves efficient performance outcomes, including higher accuracy, precision, with minimized time.

***Keywords*:** Education Data Mining (EDM), Students admission prediction, data preprocessing, Horvitz–Thompson Weighted imputation, maximum normalized residual test, Kernel Cook's Proximity Projection Pursuit Regression

1. INTRODUCTION

Education is an inclusive concept that involves the acquisition of knowledge, skills, values, and attitudes. It is a fundamental aspect of human development and plays a crucial role in the shaping individuals and societies. Every year many college students apply for post graduate admission to various universities or colleges. With advancements in the education system and increasing competitiveness, accurately predicting students' admission chances has become a vital process. Educational institutions attempt to select the best-qualified students who have demonstrated superior academic achievements. Various machine learning methods have been developed to predict students' admission probabilities.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

An integration of Edited Nearest Neighbor (ENN) and Borderline Support Vector (SVM) based SMOTE was developed in [1] for predicting undergraduate admissions. However, the designed methodology did not effectively predict the outcome of a bachelor's admission test at a university when applied to a large dataset. Decision trees were developed in [2] to extract the decision rules-based admission process prediction effectively for specific departments. However, an efficient machine learning algorithm was not applied to large educational institute datasets to achieve better accuracy and minimal time.

The model, which combines the PNN and SVM stacking ensemble was developed in [3], aimed to predict the admission of students for higher education. However, it faced challenges in creating a feature selection model to minimize the time complexity of the admission prediction process. The Artificial Neural Network technique was developed [4] to support decision-making in university admission systems. However, efficient data mining techniques were not applied for accurate university admission prediction.

Artificial Neural network (ANN) was developed in [5] for the effective prediction of admissibility of candidates for post-secondary education. However, the accuracy performance in predicting the admissibility of candidates was not improved. Machine Learning model was developed in [16] to measure their chances of getting admission into a particular university with better accuracy.

A Random Forest Regression model was developed in [7] for predicting university admission. However, hybrid feature selection algorithms were not employed for admission prediction. Additionally, a dataset with a substantial size and diverse features was not utilized to address scalability issues. An optimization hybrid Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) was developed in [8] for the probability prediction of graduate admission. However, it failed to apply an effective prediction model for postgraduate student admissions.

A linear regression model was developed in [9] to predict the relationship between admission chances and score values. However, it failed for constructing a more advanced model to enhance the data processing and methods. Machine learning models were introduced in [10] to detect admissions, focusing on enhancing the contribution of the decision-making process. However, these models faced limitations in selecting feature assessments to improve the accuracy of admission predictions.

1.1 Key contributions of the paper

The key contributions of the WRTPKPR technique is summarized as follows,

- To enhance the accuracy of predicting the admissions, the WRTPKPR technique is designed based on preprocessing and feature selection.
- To reduce the prediction time for student admissions, the WRTPKPR technique employs the Horvitz–Thompson Weighted imputation to handle missing data. Additionally, the technique utilizes the maximum normalized residual statistical method to effectively identify and eliminate outliers.
- The WRTPKPR technique utilizes Kernel Cook's Proximity Projection Pursuit Regression to select relevant features, thereby improving prediction accuracy through the application of a Laplace kernel function.
- Finally, extensive experiments were conducted to evaluate the performance of our WRTPKPR technique in comparison to other existing approaches.

1.2 Organization of the paper

The paper is organized into five sections as follows: Section 2 reviews related works. In Section 3, the WRTPKPR technique is described with clear diagram. Section 4 presents experimental settings and provides a description of the dataset. A comparative analysis of different metrics is presented in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORKS

Efficient machine learning approaches were developed in [11] for predicting admission outcomes. However, these approaches failed to focus the collection of larger datasets and the exploration of deep learning techniques to enhance the university admission process.

Supervised machine learning algorithm was introduced in [12] for Graduate Admission Prediction based on exploratory data analysis. In [13], various machine learning methods were performed for predicting admissions. But the methods failed to improve the prediction accuracy. An Artificial Neural Network (ANN) was developed in [14] to predict the percentage chance of student admittance by utilizing various test attributes. However, it did not incorporate additional features such as internship details, work experience, or other relevant parameters for making admission decisions. Machine learning-based decision-making method was developed in [15] for detecting the postgraduate admissions. A Logistic Regression method was introduced in [16] with the aim of detecting the student university admission.

Naïve Bayes and Logistic Regression classification models were developed in [17] for the purpose of detecting graduate admission. A Naïve Bayes Classification and Kernel Density Estimations (KDE) method were introduced in [18] to detect the student's admission based on students' examination score and other significant features. But it failed to investigate the accuracy provided by using different

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

predictive models for getting admission into universities or institutions.

Fine-tuned random forest classifiers were introduced in [19] for detecting new student performances based on admission data attributes. However, feature extraction techniques were not applied during the classification process to enhance detection results. A 2-stage architecture with a deep neural network (DNN) was developed in [20] for extracting information from textual data.

3. PROPOSAL METHODOLOGY

The admission process plays a pivotal role in shaping the future of aspiring students. Educational institutions face the complex task of selecting the most appropriate candidates for higher education. In recent days, machine learning has emerged as a dominant tool to enhance decision-making processes in various domains, especially in education. By leveraging large datasets and advanced algorithms, machine learning methods are used to analyze complex patterns and make accurate predictions. Therefore, a novel machine learning technique called WRTPKPR is introduced in this section for selecting the most appropriate candidates in organizations during the admission process.

The diagram presented in figure 1 illustrates the architecture of the proposed WRTPKPR technique aimed at improving the accuracy of predicting student's admission in educational data mining. The WRTPKPR technique consists of three essential processes: data acquisition, preprocessing, and feature selection.

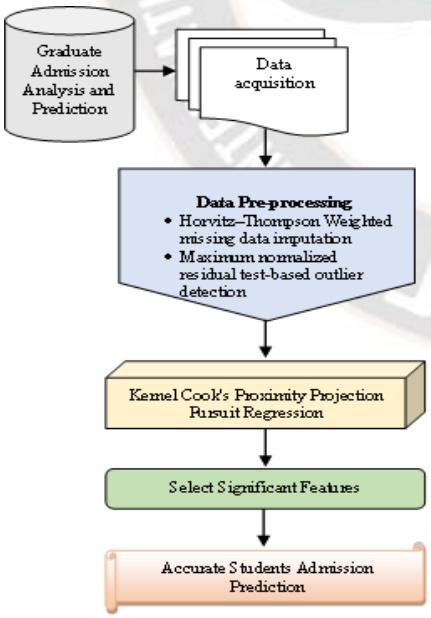

Figure 1: Architecture of the proposed WRTPKPR technique

In the data acquisition phase, student data is collected from the dataset to predict the probability of success in postgraduate admissions. Subsequently, preprocessing is conducted to organize the raw data, including handling missing data imputation and detecting outliers. Finally, Kernel Cook's Proximity Projection Pursuit Regression is employed to select significant features for precise prediction of student admissions.

3.1 Data acquisition phase

In the WRTPKPR technique, data acquisition involves the process of gathering raw data from the dataset to predict the probability of success in postgraduate admissions. This phase serves as the initial step in the overall methodology and is fundamental for obtaining the necessary information for accurate predictions. The Graduate Admission Analysis and Prediction dataset [21] is used for collecting the data. This dataset includes Graduate Admissions data i.e. GRE Score, TOEFL Score, University Rating, Statement of Purpose, Letter of Recommendation Strength, Undergraduate CGPA, and Chance of Admit. These Graduate Admissions information are used for predictive analytics.

3.2 Data preprocessing

In the WRTPKPR technique, data preprocessing is a significant step that involves cleaning and organizing the raw data to ensure its quality and suitability for further analysis. The main objectives of data preprocessing are to handle missing data, identify and address outliers for effective feature selection and model training. The two processes in the preprocessing techniques are explained in below subsections.

3.2.1 Horvitz-Thompson Weighted missing data imputation

Handling missing data is a fundamental step in the data preprocessing phase. Missing data refers to the absence of data for certain features in a given dataset. The proposed WRTPKPR technique utilizes the Horvitz–Thompson Weighted imputation technique for finding the missing values with known other values of the particular feature.

The Horvitz–Thompson estimator is a statistical method used for detecting the missing values in the given dataset. Let us consider the raw input dataset 'DS' and it formulated in the form of input matrix as follows,

$$M = \begin{bmatrix} f_1 & f_2 & \dots & f_m \\ Dp_{11} & Dp_{12} & \dots & Dp_{1n} \\ Dp_{21} & Dp_{22} & \dots & Dp_{2n} \\ \vdots & \vdots & \dots & \vdots \\ Dp_{m1} & Dp_{m2} & \dots & Dp_{mn} \end{bmatrix} \quad (1)$$

Where, M 'denotes an input matrix, $f_1, f_2, \dots f_m$ indicates a number of features in the column of the matrix, the row 'n' represents the data samples or instances or student

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

data which contains the number of data points $Dp_1, Dp_2, \dots Dp_n$.

After formulating the input matrix, the missing value is computed by applying a Horvitz–Thompson estimator. The proposed estimator is used to compute the mean by summing up the product of the inverse weights and the independent data points in the sample and then dividing by the total population size.

Let us consider the number of independent data points $Dp_1, Dp_2, \dots Dp_n$. Therefore, the missing data point is identified as follows,

$$\mu_e = \frac{1}{N} \sum_{i=1}^{n} Dp_i * \varphi^{-1} \quad (2)$$

Where, μ_e denotes the mean value i.e. imputed value, N denotes a number of data points or total population size, Dp indicates a data points, φ^{-1} indicates an inverse weight which is the reciprocal of the inclusion probability. In this way, all the missing values are handled in the dataset.

3.2.2 Maximum normalized residual test-based outlier data detection

Outlier's data detection is another preprocessing step to identify data points that deviate significantly from the other data points in the given dataset. This outlier's data affect the performance of machine learning models during the prediction. Therefore, the proposed technique uses the maximum normalized residual test for detecting the outlier data in the given dataset. It is a statistical method used for detecting the outliers in a normally distributed data point.

Let us consider the number of normally distributed data points $Dp_1, Dp_2, \dots Dp_n$ in the given dataset.

The maximum normalized residual test statistic 'RT' is defined as follows,

$$RT = \frac{arg\max \beta}{\sigma} \quad (3)$$

$$\beta = \frac{1}{n} |Dp_i - M|^2 \quad (4)$$

$$M = \frac{\sum_{i=1}^{n} Dp_i}{n} \quad (5)$$

Where, β denotes a maximum absolute difference between any individual data point (Dp_i) and the sample mean 'M', σ indicates a deviation, n denotes a number of data points, $\frac{1}{n}|Dp_i - M|^2$ indicates Prevosti's distance measure, $arg\ max$ indicates an argument of maximum function. The absolute difference between data point is greater than the maximum allowable deviation, then it is said to be an outlier data point. Otherwise, the data point is said to be a normal. This data points are removed from the dataset. Therefore, the preprocessing step of the proposed technique improves the accurate student's admission prediction with minimum time consumption. The algorithm of the preprocessing is given below,

Algorithm 1: Data Preprocessing

Input: Datasets 'DS', data points '$Dp = \{Dp_1, Dp_2, \dots, Dp_n\}$', Features '$f = \{f_1, f_2, \dots, f_m\}$'

Output: Preprocessed dataset

Begin
Step 1: Collect the number of features '$f = \{f_1, f_2, \dots, f_m\}$', data points $Dp_1, Dp_2, \dots, Dp_n$ from the dataset
Step 2: **Formulate the** input matrix 'M' using (1)
Step 3: **For** each feature 'f' and data points 'Dp'
Step 4: Perform missing data imputation using (2)
Step 5: **End for**
Step 6: **For** each data point Dp_i
Step 7: Perform **maximum** normalized residual test statistic using (3) (4) (5)
Step 8: Detect **o**utlier data points
Step 9: **End for**
Step 10: **Return** (preprocessed dataset)
End

The above algorithm 1 describes the step by step process of data preprocessing. The process begins by collecting number of features and data points from datasets. After the data acquisition, missing data handling process is performed by applying a Horvitz–Thompson Weighted imputation technique. Following imputation process, outlier data detection is performed by using Maximum normalized residual test. This helps for identifying observations that deviate significantly from the overall dataset. Finally, the preprocessed results are obtained.

3.3 Kernel Cook's Proximity Projection Pursuit Regression based feature selection

The second process of the proposed WRTPKPR technique is the target feature selection process is carried out by applying Kernel Cook's Proximity Projection Pursuit Regression. Projection Pursuit Regression is a machine learning technique used to analyze the relationship between features and find the minimal residual (i.e., variance) between them. The Kernel Cook's Proximity is a statistical method applied in regression to identify the variance between features in the dataset through the least-squares analysis. As a result, the features with the minimum variance are selected for predictive analytics, while other highly deviated features are removed. Based on the selected target features, accurate admission predictions are made for higher education graduates with minimal time consumption.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

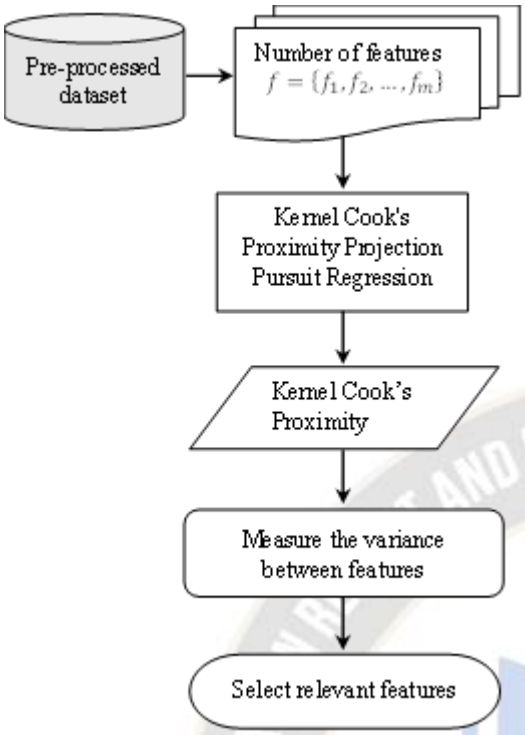

Figure 2: Flow process of Kernel Cook's Proximity Projection Pursuit Regression

Figure 2 flow process of the Kernel Cook's Proximity Projection Pursuit Regression based feature selection for accurate student admission prediction. Let us consider the number of features $\{f_1, f_2, \dots, f_m\}$ taken from the dataset. Then applying a projection pursuit regression for measuring the relationship and finds the minimal residual (i.e., variance) between them.

Cook's distance is a commonly used estimate of the distance of every feature when performing a least-squares regression analysis. The distance is measured based on Laplace kernel as follows,

$$CD = \frac{\exp |f_m - f_j|}{N\vartheta^2} \quad (6)$$

Where, CD indicates a Cook's distance which is measured difference between two features $|f_m - f_j|$, N represents the number of independent variables or features included in the model, ϑ^2 denotes a mean squared error of the model fit,

$$\vartheta = \sum (CD_{obs} - CD_{prd})^2 \quad (7)$$

Where, average of the least squared differences between the observed 'CD_{obs}' and predicted 'CD_{prd}' values. The projection pursuit regression finds the minimum distance.

$$y = \arg\min CD \quad (8)$$

Therefore, the feature with the minimum distance has lesser error that is projected for predictive analytics. Other remaining features are removed from the dataset. A projection on a relevant feature vector space is formulated as given below,

$$P: f_m \rightarrow f_k \quad (9)$$

Where, P denotes a projection vector, f_m original set of features, f_k denotes a relevant feature. In this way, relevant features are selected from the dataset. With the selected features are used to minimize the time consumption of accurate student's admission prediction. The algorithm of the kernel cook's proximity projection pursuit regression-based feature selection is given below,

Algorithm 2: Kernel Cook's Proximity Projection Pursuit Regression-based Feature Selection
Input: Preprocessed Datasets 'DS', data points '$Dp = \{Dp_1, Dp_2, \dots, Dp_n\}$', Features '$f = \{f_1, f_2, \dots, f_m\}$'
Output: select relevant features
Begin **Step 1:** Collect the preprocessed dataset as input **Step 2:** **For** each feature **Step 3:** Measure the Cook's distance based on Laplace kernel using (6) **Step 4:** Projection pursuit regression find minimum deviation $y = \arg\min \vartheta$ **Step 5:** Project the feature with minimum distance $P: f_m \rightarrow f_k$ **Step 6:** Select relevant features and remove other features **Step 7. End for** **End**

Algorithm 2 outlines the steps for feature selection with the goal of improving student admission predictions while minimizing time consumption. The process starts by taking a preprocessed dataset as input, which includes a number of features. Next, Kernel Cook's Proximity is computed between the features in the dataset. Following this, regression is employed to identify the minimum distance between the features. This method is applied to differentiate between relevant and irrelevant features within the dataset. As a result, the relevant feature selection process contributes to enhancing the accuracy of student admission prediction models.

4. EXPERIMENTAL SETTINGS

In this section, the experimental assessment of the proposed WRTPKPR technique and existing ENN and borderline SVM-based SMOTE [1] and decision trees [2] are implemented using Python coding. To conduct the experiment, Graduate Admission Analysis and Prediction dataset [21] with Admission_Predict_Ver1.1.csv is used. The main aim of predicting the chances of admission of a student based on his/her scores information. The dataset includes 9 features and 500 instances or records or student data for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

admission prediction. The features details are listed in Table 1.

Table 1: Feature Description

S.No	Features	Description
1	Serial No	Number of students 1-500
2	GRE Score	Graduate Record Examinations score (290 to 340)
3	TOEFL Score	Test of English as a Foreign Language score (92 to 120)
4	University Rating	1 to 5
5	SOP	Statement of Purpose (1 to 5)
6	LOR	Letter of Recommendation Strength (1 to 5)
7	CGPA	Undergraduate CGPA (6.8 to 9.92)
8	Research	Research Experience (0 or 1)
9	Chance of Admit	(0.34 to 0.92)

5. PERFORMANCE RESULTS AND DISCUSSION

In this section, performance of the proposed WRTPKPR technique and existing ENN and borderline SVM-based SMOTE [1] and decision trees [2] are evaluated with various performance metrics, including admission prediction accuracy, precision, recall, F1 score and admission prediction time across different number of student data.

5.1 Impact of Admission Prediction Accuracy

It is referred to the ratio of predicting the chances of admission with the number of student data from the total number of data. Therefore, accuracy is formulated as follows

$$\boldsymbol{APA} = \left(\frac{\boldsymbol{Tp+Tn}}{\boldsymbol{Tp+\ Tn+Fp+Fn}}\right) * \mathbf{100} \qquad (10)$$

Where, $\boldsymbol{APA}$ denotes an admission prediction accuracy, Tp denotes the true positive, Tn denotes the true negative, $\boldsymbol{Fp}$ represents the false positive, $\boldsymbol{Fn}$ represents the false negative. It is measured in the unit of percentage (%).

Table 2 : Performance comparison of admission prediction accuracy

Number of student data	Admission Prediction Accuracy (%)		
	WRTPKPR	ENN and borderline SVM-based SMOTE	Decision trees
50	90	86	82
100	91	85	83
150	91.33	88	86
200	91.5	86.5	84.5
250	92	87.2	85.6
300	90.66	85.66	83.33
350	90.57	86.85	84.28
400	91	88.5	86.5
450	92.22	88.88	86.66
500	91	87	85.8

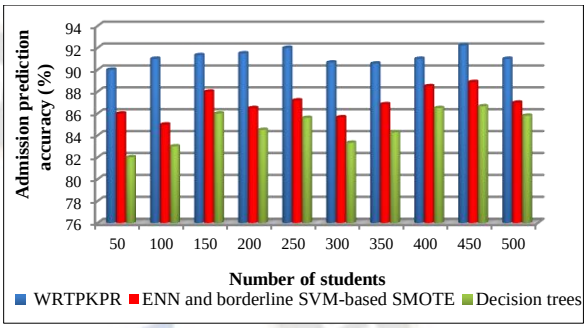

Figure 3 : Graphical Assessment of Admission Prediction Accuracy

Figure 3 given above depicts performance of student admission prediction accuracy using three methods namely WRTPKPR technique, ENN and borderline SVM-based SMOTE [1] and Decision trees [2]. As shown in figure 3, the horizontal axis indicates the number of student data ranging from 50 to 500, while the vertical axis indicates the evaluation results of prediction accuracy using the three methods. Among three methods, the performance of WRTPKPR technique is better when compared to the existing methods. For example, considering experiments with 50 student data for calculating student admission prediction accuracy, the WRTPKPR technique achieved an accuracy of 90%. In addition, the prediction accuracy of the existing models [1] and [2] was found to be 86% and 982%, respectively. Overall, the comparative analyses indicate that the student admission prediction accuracy of the WRTPKPR technique considerably increased by 5% and 8% compared to [1] and [2], respectively. This is because of applying the Kernel Cook's Proximity Projection Pursuit Regression. The regression function is employed to select target features from the dataset based on deviation measures. Consequently, accurate admission predictions are made using these selected relevant features to achieve high accuracy.

5.2 Impact of Precision

It is referred to the ratio of accurate predicting the admission with the number of student data. The precision is mathematically formulated as,

$$\boldsymbol{Pr} = \left(\frac{\boldsymbol{Tp}}{\boldsymbol{Tp+Fp}}\right) \qquad (11)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

Where, Pr denotes a precision, Tp denotes the true positive, Fp represents the false positive.

Table 3 Performance comparison of precision

Number of student data	Precision		
	WRTPKPR	ENN and borderline SVM-based SMOTE	Decision trees
50	0.930	0.902	0.871
100	0.932	0.882	0.860
150	0.940	0.921	0.900
200	0.944	0.910	0.885
250	0.947	0.914	0.901
300	0.940	0.900	0.877
350	0.935	0.911	0.894
400	0.940	0.918	0.903
450	0.95	0.922	0.906
500	0.943	0.910	0.9

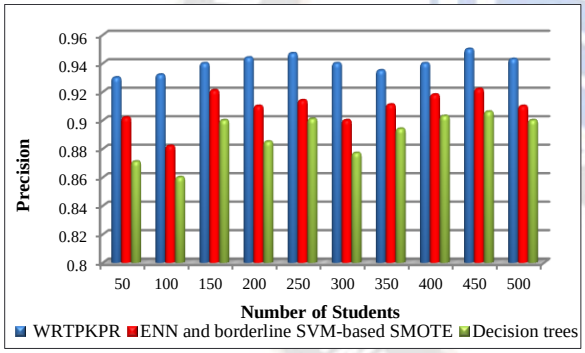

Figure 4 : Graphical assessment of precision

Figure illustrates the performance of precision versus number of student data varied from 50 to 500. There are three methods used for comparison of the precision. For each method, ten various results are observed along with the counts of input data. The observed results indicate that the WRTPKPR technique outperforms well in terms of achieving higher precision than the other two existing methods. The performance of WRTPKPR technique is compared to the results of existing methods. The comparison results prove that the performance results of precision are considerably improved using WRTPKPR technique by 3% when compared to [1] and 6% when compared to [2]. To achieve this performance, WRTPKPR technique employs a Kernel Cook's Proximity Projection Pursuit Regression for selecting the target features to enhance the admission prediction with a better true positive rate and lesser false positive results thereby increasing precision.

5.3 Impact of Recall

Recall, also known as sensitivity is the ratio of correctly predicted positive observations to the total actual positives. It is calculated using the following formula:

$$Rc = \left(\frac{Tp}{Tp+Fn}\right) \quad (12)$$

Where, Rc denotes a recall, Tp denotes the true positive, Fn represents the false negative.

Figure 5 illustrates the assessment of recall versus the number of student data, ranging from 50 to 500. The horizontal axis of the graph indicates the number of student data, while the vertical axis represents the performance outcomes of recall. The analysis reveals that the WRTPKPR technique consistently outperforms than the conventional methods in terms of achieving a better recall value. In the first run, where 50 student data are considered for calculating recall, the WRTPKPR technique achieved a recall of 0.952, while the recall of existing methods [1] and [2] was observed to be 0.925 and 0.902, respectively. Similarly, different performance results are observed for all three methods.

Table 4 Performance comparison of Recall

Number of student data	Recall		
	WRTPKPR	ENN and borderline SVM-based SMOTE	Decision trees
50	0.952	0.925	0.902
100	0.965	0.937	0.925
150	0.962	0.936	0.928
200	0.960	0.927	0.921
250	0.964	0.932	0.923
300	0.954	0.926	0.919
350	0.957	0.930	0.914
400	0.957	0.945	0.934
450	0.962	0.946	0.934
500	0.954	0.934	0.927

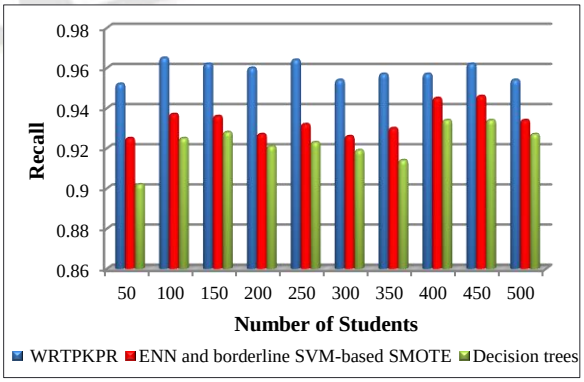

Figure 5 Graphical assessment of Recall

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

Finally, the results of the WRTPKPR technique are compared to existing methods. Overall recall comparison results show that the WRTPKPR technique improves the performance of recall by 3% and 4% when compared to [1] and [2], respectively. This is because of WRTPKPR technique's accurately predicting the chance of student admission for higher education based on the selected feature score value. Therefore, it enhances the performance of true positives and reduces false negatives, thereby improving recall.

5.4 Impact of F1 score

The F1 score is a metric to provide a balanced assessment of a model's performance. It is measured as the harmonic mean of precision as well as recall. It is calculated as follows,

$$\boldsymbol{F1\ score} = \boldsymbol{2} * \left(\frac{\boldsymbol{Pr} * \boldsymbol{Rc}}{\boldsymbol{Pr} + \boldsymbol{Rc}}\right) \qquad (13)$$

Where, $\boldsymbol{Rc}$ denotes a recall, $\boldsymbol{Pr}$ denotes the precision.

Table 5 performance comparison of F1 score

Number of student data	F1 Score		
	WRTPKPR	ENN and borderline SVM-based SMOTE	Decision trees
50	0.940	0.913	0.886
100	0.948	0.908	0.891
150	0.950	0.928	0.913
200	0.951	0.918	0.902
250	0.955	0.922	0.911
300	0.946	0.912	0.897
350	0.945	0.920	0.903
400	0.948	0.931	0.918
450	0.955	0.933	0.919
500	0.948	0.921	0.913

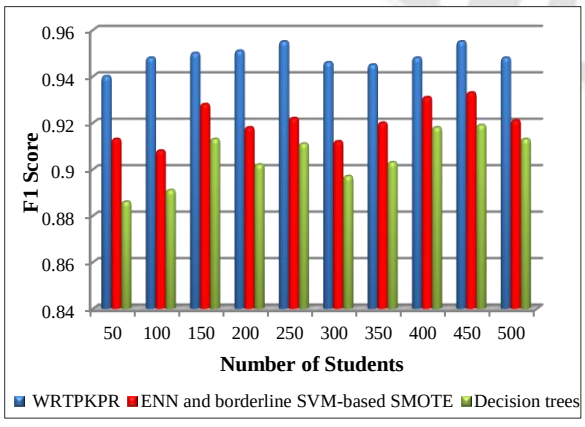

Figure 6 : Graphical assessment of F1 score

The performance of the F1-score for student admission prediction using three methods, including the WRTPKPR technique and [1] [2], is depicted in Figure 6. The dataset includes the number of student data ranging from 50 to 500. In the graph, the number of student data is represented on the horizontal axis, while the F1-score performance for all three methods is observed on the vertical axis. The graphical representation illustrates that the F1-score of the WRTPKPR technique is improved compared to the other two existing methods [1] [2]. This improvement is attributed to the WRTPKPR technique enhancing the performance of both precision and recall during admission prediction. The average of ten comparison results indicates that the F1-score of the WRTPKPR technique is significantly enhanced by 3% and 5% compared to existing methods [1] and [2], respectively. This improvement highlights the effectiveness of the WRTPKPR technique in achieving a balanced trade-off between precision and recall for student admission prediction.

5.5 Impact of Admission Prediction Time

It is measured as an amount of time consumed by algorithm for student admission prediction. The prediction time is formulated as follows,

$$\boldsymbol{APT} = \sum_{i=1}^{n} \boldsymbol{Dp_i} * \boldsymbol{Time}\ (\boldsymbol{AP}) \qquad (14)$$

Where, $\boldsymbol{APT}$ denotes a student admission prediction time based on the data '$\boldsymbol{Dp_i}$' and the actual time consumed in prediction for one data denoted by '$\boldsymbol{Time}\ (\boldsymbol{AP})$'. It is measured in terms of milliseconds (ms).

Table 6 : Performance comparison of Admission Prediction Time

Number of student data	Admission Prediction Time (ms)		
	WRTPKPR	ENN and borderline SVM-based SMOTE	Decision trees
50	18	20	22.5
100	22	25	28
150	27	30	33
200	30	33	34
250	32.5	35	37.5
300	36	39	42
350	38.5	40.25	42
400	40	42	44
450	42.75	44.1	47.25
500	46	47.5	50

Figure 7 illustrates the performance analysis of admission prediction time using three different methods namely WRTPKPR technique, [1], and [2]. The admission

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
Article Received: *10 July 2023* ***Revised:*** *12 August 2023* ***Accepted:*** *28 August 2023*

prediction time for all three methods gradually increases with an increase in the number of student data.

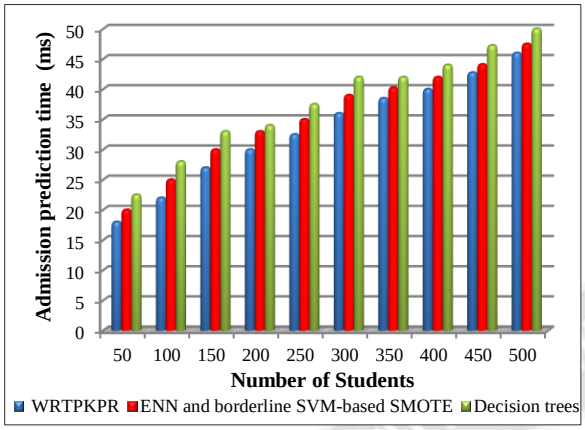

Figure 7 : Graphical assessment of admission prediction time

Especially, the admission prediction time for the WRTPKPR technique is significantly reduced compared to the existing methods [1] and [2]. Lt us consider the first iteration with 50 student data, where the admission prediction time for the WRTPKPR technique was18 ms. Similarly, the time consumption for [1] and [2] was 20 ms and 22.5 ms respectively. The results obtained from the WRTPKPR technique are then compared to the existing methods. The average comparison indicates that the admission prediction time is considerably minimized by 7% and 13% when compared to the existing methods [1] [2]. This is due to the WRTPKPR techniques performed the data preprocessing steps, which involve missing data imputation and outlier detection. The Horvitz–Thompson Weighted imputation method is employed for finding missing data points based on other known data points. In the second step, outlier detection is performed using the maximum normalized residual test. The preprocessing steps of the WRTPKPR technique reduce the time complexity of admission prediction. Additionally, the WRTPKPR technique utilizes a target feature selection process to select relevant features and eliminate others, further reducing time consumption.

6. CONCLUSION

Education data mining (EDM) is a field of study that involves applying data mining techniques to educational data to identify patterns, and make informed decisions about teaching and learning. The predictive modeling process help educational institutions make informed decisions during the admission process and identify factors that contribute to successful admissions. In this paper, a novel WRTPKPR technique has been developed for predicting the admissions of student to improve the quality of education. The WRTPKPR technique incorporates data preprocessing step to organize the dataset properly by filling the missing data and remove the outlier data before applying the machine learning technique. Followed by, the feature selection process is executed using Kernel Cook's proximity projection pursuit regression to predict probability of admission with higher accuracy and minimal time. The paper conducts a comprehensive experimental evaluation using a publicly available dataset. The results demonstrate that the WRTPKPR technique outperforms existing methods in terms of student admission prediction accuracy, precision, recall, and F1-score. Furthermore, the WRTPKPR technique also proves to be efficient in minimizing the time consumption for student admission prediction compared to existing approaches.

REFERENCES

[1] Md. Abul Ala Walida, S.M. Masum Ahmed, Mohammad Zeyad, S. M. Saklain Galib, Meherun Nesa, “Analysis of machine learning strategies for prediction of passing undergraduate admission test”, International Journal of Information Management Data Insights, Elsevier, Volume 2, Issue 2, November 2022,Pages1-12. https://doi.org/10.1016/j.jjimei.2022.100111

[2] Muhammad Nauman, Nadeem Akhtar, Adi Alhudhaif, And Abdulrahman Alothaim, “Guaranteeing Correctness of Machine Learning Based Decision Making at Higher Educational Institutions”, IEEE Access, Volume 9, 2021, Pages 92864-92880. DOI: 10.1109/ACCESS.2021.3088901

[3] Khrystyna Zub, Pavlo Zhezhnych and Christine Strauss, “Two-Stage PNN–SVM Ensemble for Higher Education Admission Prediction”, Big data and cognitive computing, Volume 7, Issue 2, 2023, Pages 1-13. https://doi.org/10.3390/bdcc7020083

[4] Hanan Abdullah Mengash, “Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems”, IEEE Access, Volume 8, 2020, Pages 55462-55470. DOI: 10.1109/ACCESS.2020.2981905

[5] Anietie Ekong, Abasiama Silas and Saviour Inyang, “A Machine Learning Approach for Prediction of Students’ Admissibility for Post-Secondary Education using Artificial Neural Network”, International Journal of Computer Applications, Volume 184 , Issue 27, 2022, Pages 44-49, DOI: 10.5120/ijca2022922340

[6] Aashish Singha and Saurabh Gautam, “Graduate University Admission Predictor using Machine Learning”, International Journal for Modern Trends in Science and Technology, Volume 6, Issue 12, 2020,Pages474-478. https://doi.org/10.46501/IJMTST061292

[7] Inssaf El Guabassi, Zakaria Bousalem, Rim Marah, Aimad Qazdar, “A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms”, International Journal of Online and Biomedical Engineering (iJOE), Volume 17, Issue 02, 2021, Pages135–147. https://doi.org/10.3991/ijoe.v17i02.20049

[8] Burhanudin Zuhri, Nisa Hanum Harani, Cahyo Prianto, "Probability Prediction for Graduate Admission Using CNN-LSTM Hybrid Algorithm", Indonesian Journal of Computer Science, Volume 12, Issue3,2023, Pages1105-119. DOI:10.33022/ijcs.v12i3.3248

[9] Wenhan Ju, Heran Wang and Zi Ye, "Analysis of Relative Importance of Factors Affecting the Chance of Admission of Applying to Graduate Students", Journal of Education, Humanities and Social Sciences, Volume-6, 2022, Pages1-11. DOI:10.54097/ehss.v6i.4048

[10] Yijun Zhao, Xiaoyu Chen, Haoran Xue, Gary M. Weiss, "A machine learning approach to graduate admissions and the role of letters of recommendation", PLoS One, Volume 18, Issue 10, 2023,Pages1-17. https://doi.org/10.1371/journal.pone.0291107

[11] Aga Maulana, Teuku Rizky Noviandy, Novi Reandy Sasmita, Maria Paristiowati, Rivansyah Suhendra, Erkata Yandri, Justinus Satrio and Rinaldi Idroes, "Optimizing University Admissions: A Machine Learning Perspective", Journal of Educational Management and Learning, Volume 1, Issue 1, 2023, Pages 1-7. DOI: 10.60084/jeml.v1i1.46

[12] Sujay S **,** "Supervised Machine Learning Modelling & Analysis For Graduate Admission Prediction ", International Journal of Trend in Research and Development (IJTRD), Volume 7, Issue-4 , 2020, Pages 1-3.
URL: http://www.ijtrd.com/papers/IJTRD22200.pdf

[13] Anmol Pawar, Rushikesh Patil, Kadayya Mathapati, Pratik Lonare, Prof.Laximikant Malphedwar, "College Admission Prediction Using Machine Learning", Journal of Emerging Technologies and Innovative Research, , Volume 10, Issue 11, 2023, Pages 229-233. :http://www.jetir.org/papers/JETIR2311129.pdf

[14] Meenakshi Bhrugubanda, Varsha Udutha, Sri Charan Yella, "Post Graduate Admission Prediction Using ANN", International Journal for Research in Applied Science & Engineering Technology, Volume 11, Issue VI, 2023, Pages 219-222. https://doi.org/10.22214/ijraset.2023.53463

[15] Xiaojun Wu, Jing Wu, "Criteria evaluation and selection in non-native language MBA students admission based on machine learning methods", Journal of Ambient Intelligence and Humanized Computing, Springer, Volume 11, 2020, Pages 3521–3533. https://doi.org/10.1007/s12652-019-01490-0

[16] Sharan Kumar Paratala Rajagopal, "Predicting Student University Admission Using Logistic Regression", European Journal of Computer Science and Information Technology, Volume 8, Issue 3, 2020, Pages 46-56. https://doi.org/10.37745/ejcsit.2013

[17] Adrita Iman and Xiaoguang Tian, "A Comparison of Classification Models in Predicting Graduate Admission Decision", Journal of Higher Education Theory and Practice, Volume 21, Issue 7, 2021, Pages 219- 230. https://doi.org/10.33423/jhetp.v21i7.4498

[18] Nasim Matar, Wasef Matar, Tirad Al Malahmeh, "A Predictive Model for Students Admission Uncertainty Using Naïve Bayes Classifier and Kernel Density Estimation (KDE)", International Journal of Emerging Technologies in Learning (iJET), Volume 17, Issue 8, 2022, Pages 75–96. https://doi.org/10.3991/ijet.v17i08.29827

[19] Chayaporn Kaensar, Worayoot Wongnin, "Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning", EURASIA Journal of Mathematics, Science and Technology Education, Volume 19, Issue 12, 2023, Pages 1-16. https://doi.org/10.29333/ejmste/13863

[20] Bruno P. Roquette, Hitoshi Nagano, Ernesto C. Marujo, Alexandre C. Maiorano, "Prediction of admission in pediatric emergency department with deep neural networks and triage textual data", Neural Networks, Elsevier, Volume 126, June 2020, Pages 170-177. https://doi.org/10.1016/j.neunet.2020.03.012

[21] Graduate Admission Analysis and Prediction dataset: https://github.com/divyansha1115/Graduate-Admission-Prediction