Int. j. inf. tecnol. (June 2024) 16(5):3193–3204
https://doi.org/10.1007/s41870-024-01818-w

ORIGINAL RESEARCH

Customized support vector machine for predicting the employability of students pursuing engineering

Suja Jayachandran[1,2] · **Bharti Joshi**[1]

Received: 17 November 2023 / Accepted: 2 March 2024 / Published online: 6 April 2024
© Bharati Vidyapeeth's Institute of Computer Applications and Management 2024

Abstract Higher Education Institutions are always concerned about their students' employability, so it is ideal to have a mechanism to forecast students' employability at an early stage so as to take appropriate action to improve the same. There could be many factors that can have a direct or indirect affect on employability. In this research, we have analyzed the students data who have graduated between the years of 2018 and 2022 from an engineering college and we considered both academic and socio-demographic factors. To identify the best attributes that impact employability, we have used our proposed feature selection approach, which is influenced by the Teaching Learning Based Optimization (TLBO) algorithm. The study employs several classifiers, including Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Gradient Boosting Adaptive Boosting, and Extreme Gradient Boosting and found that Support Vector Machine(SVM) provides the greatest accuracy of 74.37%. Then to improve the accuracy of SVM we have proposed Customized SVM where we have customized the hyperparameters like the kernel function's radial basis function and regularization parameter, C value. We found that with this novel approach, accuracy improved by 13.43%. Thus this proposed novel approach helps in finding an optimal set of features and the best classifier for predicting students' employability enrolled in a Technical Higher Education Institute.

Keywords SVM · Radial basis function · Employability · Feature selection · Prediction

1 Introduction

A nation's economy depends heavily on higher education since it produces a stable and skilled labour force. Higher education serves as the cornerstone for many advantages, including fostering talent, raising the standard of the nation's human capital, and enhancing a country's competitive status. As a result, educational institutions are under pressure to find better ways to increase their students' employability. In most institutions, each student's employability is their top concern, so predicting students' employability before they apply for jobs can enhance the percentage of students who are placed in institutions. Before an interview, students should assess their deficiencies so they may work on improving those areas.

In this data driven era, educational institute collect a lot of data like academic record, attendance, socio-demographic details, extra-curricular and co-curricular data etc. through learning management system and management information system. This can be used to predict performance and employability of a student.

This paper will analyse characteristics that are crucial in determining a student's employability. During data pre-processing, we can employ a variety of Feature Selection (FS) algorithms to enhance the functionality of the Machine Learning (ML) model. Filter, wrapper, and hybrid models are some

Suja Jayachandran and Bharti Joshi contributed equally to this work.

✉ Suja Jayachandran
suja.jayachandran@vit.edu.in

Bharti Joshi
bharti.joshi@rait.ac.in

[1] Department of Computer Engineering, Ramrao Adik Institute of Technology, D.Y. Patil Deemed to be University, Nerul, Navi Mumbai 400706, Maharashtra, India

[2] Department of Computer Engineering, Vidyalankar Institute of Technology, Wadala, Mumbai 400037, Maharashtra, India

examples of different FS algorithms. The filter approach is used in the pre-processing stage and depends on all features. The statistical association between the target variable and their score is taken into account. The wrapper approach assesses the features through the application of learning algorithms. The characteristics of both filter and wrapper approaches are combined in hybrid FS.

In this paper, we are optimizing the FS method using a nature-inspired algorithm. In paper [1] has covered several algorithms that are inspired by nature, including the Cuckoo Search, Ant Colony Optimization, Firefly, Genetic Algorithm, and Swarm Intelligence based algorithm. An algorithm inspired by teaching and learning called Teaching Learning Based Optimization (TLBO), is discussed in [2]. It has Teacher's and Learner's phases. Here population is the set of learners(students) and variables are features that affect the employability and fitness value is their placement in a company. The best solution to any problem discussed will be given by the tutor.

In Teacher's Phase, the teacher imparts information to understand the problem to the learners such that they provide the optimized solution to the said problem. The one who finds the optimal solution is the best learner. Now find a new solution Ynew (Eq. 1) for each Y solution with the help of an optimized solution and the average outcome. The fitness value will then be updated, and if there is an improvement, the new solution will be accepted using greedy selection to improve learning.

Let us consider the following: Y is the present answer, Ybest is the person who has worked to enhance the overall result, Tf is the teaching parameter (which is the same for all variables, either 1 or 2), Ymean is the overall result, and r is a random number (between 0 and 1). The learning phase will use this updated solution as its input.

$$\begin{aligned} Ynew &= X + r(Ybest - TfYmean) where\ Tf \\ &= round(1 + random) \end{aligned} \tag{1}$$

In the Learners Phase, a learner will increase their knowledge by interacting with their peers. The probability of acquiring more knowledge is high if the peers are better in their studies. With the partner's assistance (Yp), a new answer (Ynew) is obtained. Then, we must apply a greedy algorithm to identify an optimal solution.

To maximize optimization

$$\begin{aligned} Ynew &= Y + r(Y - Yp) iff > fp\ and\ Ynew \\ &= Y - r(Y - Yp) iff < fp \end{aligned} \tag{2}$$

To minimize the optimization

$$\begin{aligned} Ynew &= Y + r(Y - Yp) iff < fp\ and\ Ynew \\ &= Y - r(Y - Yp) iff > fp \end{aligned} \tag{3}$$

Let f be the learner's knowledge and fp be the partner's knowledge. For maximizing the optimization (Eq. 2), i.e. f is greater than fp and if f is less than fp and then minimization (Eq. 3) is performed and updates the solution according to the requirement.

Our contribution to this research is to create a dataset for which we have collected student details from the Institute's Management Information System(MIS).

- We have collected details about 1647 students who had been admitted to a direct second year(Diploma holders) of a 4-year engineering course and passed out with the engineering degree in the year 2018–2022. Their academic, socio-demographic, extra-curricular, co-curricular, and pre-placement details were considered as features. We studied it and extracted 72 features from these details.
- For performing predictive analysis we used a novel feature selection approach inspired by a TLBO algorithm to get the best set of features and then designed the machine learning model for prediction of students' employability or placement in a higher educational institute.
- Then the support vector machine ML model is customized to improve its performance.

This paper has six sections, in section 2 we have discussed work published in the area of employability prediction, feature engineering, and the TLBO. Section 3 has details about the proposed methodology. Section 4 has information about the experimental setup then section 5 discusses the result. Lastly, section 6 has a conclusion and future work.

2 Related work

This paper focuses on exploring different features affecting a student's employability in a higher education institute. In the paper [3] authors have surveyed relevant 20 papers from different academic libraries and concluded how using machine learning and artificial intelligence one can predict employability. They found few challenges like in most of the papers, features like the country's economy, psychometric attributes of the student, and gender gap were not considered. In the paper [4] authors have considered nine features based on personality traits and final year results and used models like decision tree, SVM, and random forest and concluded that the SVM model predicts better. The authors of paper [5] have surveyed how SVM has performed comparatively better with medical image MRI data, Hyperspectral data, chemical pattern data, fault diagnosis, image classification etc. The authors of [6] have discussed how factors like an institute's accreditation, facilities around the institute, previous four-year placement record, department accreditation,

and the student-faculty ratio also affect the placement of a student. They evaluated the performance of various ML models like k-nearest Neighbour model, extreme grading boost(XGB), classification and regression tree (CART), and artificial neural network model based on Mean Squared Error(MSE) and Root Mean Square Error(RMSE), Mean Absolute Error(MAE) and R-Squared(R^2) and concluded that the XGB model is better as it had lowest RMSE and highest R^2 value. Some authors have predicted the employability of students pursuing Post graduation degree like in the paper [7]. They identified essential parameter like academic factors like their 10th and 12th/diploma marks, graduation branch and marks, certification, and subject knowledge. Along with academic and socio-demographic factors some authors have considered emotional skill parameters also. Similarly, the authors in [8] have considered student data who are pursuing MBA (Masters in business administration)so features like academic, socio-demographic, work experience, specialization, and salary were considered. The authors discovered random forests could be a useful model for the placement prediction of college students. In the paper, [9] authors have considered emotional skill parameters by analyzing the response to psychometric-based questions like assertion(honesty and integrity), leadership skills, stress management skills, empathy, decision-making skills, and time management skills. Then these features were considered as input variables for J48, Random forest, sequential minimal optimization, and the Multilayered perceptron model to predict the employability and student performance of post-graduate students. They found the J48 model as the best model with 70.19% accuracy. In the paper [10] authors have tried to find whether internship has a direct impact on fostering employability using gradient boosting models. They considered hard skill, soft skill, academic, socio-demographic, internship etc. They found that internship-based context is more important in predicting placement than any other feature related to student context. Both the duration and quality of assignments completed during an internship have the highest impact on student placement. To establish a ground for the proposed work we have tabulated in Table 1 the literature review by performing a meta-analysis of available work.

The TLBO is an advanced metaheuristic optimization approach that draws inspiration from a traditional classroom setting. The authors of the paper [11] and [12] TLBO algorithm is a self-regulating. Unlike other algorithms inspired by nature, it only requires the size of the population and the number of generations, without requiring any specific trait or attribute. In the paper [13] five swarm-based nature-inspired algorithms that have been the subject of discussion by authors are the binary versions of the firefly, bat, whale(WOA), grey wolf(GWO), and particle swarm optimization algorithms. In the absence of empirical data, WOA and GWO are utilized to identify the best feature subsets. The fusion of GWO with opposition-based initialization is discussed in the paper [14]. In this paper, authors have tested this fusion and discovered a way to minimize the cost of searching while striking a balance between feature search space exploration and utilization. One more variant of TLBO is discussed in paper [15] enhanced TLBO algorithm for neural network training that includes a self-learning phase to boost efficiency.

A crucial part of data pre-processing is feature selection (FS). It aids in enhancing the ML model's performance. In paper, [16] authors considered feature selection is considered as dimensionality reduction. Better learning performance, such as increased learning accuracy, reduced computing cost, and improved model interpretability, is typically a result of feature selection. They concluded that unsupervised FS algorithms improve the performance in clustering algorithms.

In the paper, [17] authors have experimented with a Filter feature selection algorithm to predict student performance on two student dataset with different number of features and concluded that Chi-square feature selection algorithm helps in finding the optimal set of features. They have not considered wrapper or hybrid method of feature selection. In paper [18] authors have experimented with genetic algorithms, SVM, information gain, and minimum redundancy and maximum relevance techniques for feature selection with four supervised classifiers: naïve bayes, decision tree, k-nearest neighbor(kNN), and neural network and concluded that minimum redundancy and maximum relevance technique with kNN gives the highest accuracy with 10 features within the pool of 15 features.

In the paper [19] authors have done a survey on different feature selection algorithms and found some open challenges like scalability as per authors feature selection algorithms don't scale up for ultrahigh dimensionality datasets used in information retrieval, text mining type of application. The stability concerning an unsupervised feature selection algorithm is also a challenge along with ML model selection. In the paper [20] authors have experimented with binary TLBO and found an improvement in the performance of LR and linear discriminate analysis model to predict student performance. Authors of paper [21] have proposed a novel wrapper-based feature selection algorithm that needs a common controlling parameter to obtain an optimal set of features. In the paper [22] authors have stated the feature selection is useful for discarding redundant, noisy, and irrelevant features. They have used three methods correlation-based(pearson correlation coefficient), sequential feature selection (greedy search approach)and information gain (Decision Tree) based FS model. The sequential feature selection is slower than survey on different feature selection algorithms and found some open challenges like scalability

Table 1 Literature review over student data set to predict their employability by various authors

Reference	Sample size	Feature selection method	Type of features	Algorithm used	Accuracy	Limitation
Causat et al. [4]	3000	Not discussed	9 features (general appearance, manner of speaking, physical condition, mental awareness, self-confidence, ability to present ideas, communication skill, student performance rating, general point grade)	Decision Trees, Random Forest, and Support vector machine	91.2% (SVM)	Engineering domain, socio-demographic factor, year of passing, skill set, Backlog record were not considered
Çakıt and Dağdeviren [6]	314	Not discussed	29 features (No. of students and faculties, medium of education (language), program accreditation, sports participation, entertainment options, previous 4-year placement records	Extreme Gradient Boosting (XGBoost), kNN, Artificial Neural Network(ANN), Classification and Regression Tree, gradient boosting machines (GBM)	RMSE−6.45 and R^2 − 0.962 (XGB)	Engineering domain, academic record, skill set, Backlog record were not considered
Kumar et al. [8]	215	Not discussed	14 features (gender, academic, work, work-experience, specialization, placement test result)	Logistic Regression (LR), Naive Bayes, SVM, Random Forest, kNN	96% (Random Forest)	Socio-demographic factor, Backlog record, personal characteristics were not considered
Mishra et al. [9]	1400	Not discussed	31 features (socio-demographic, academic, emotional skill parameter	J48, Random Forest, Naive Bayes, Multilayered Perceptron, Sequential Minimal Optimization	70.19% (J48)	Engineering domain, logical reasoning, participation in extracurricular activities, Backlog record were not considered
Saidani et al. [10]	283	Context-aware feature selection method	18 features(hard skill, soft skill, demographic features, extracurricular, work experience, internship)	eXtreme Gradient Boosting (XGBoost), Category Boosting (CatBoost) and Light Gradient Boosted Machine (LGBM)	77.53% (Light GBM)	Engineering domain, Backlog record were not considered

as per authors feature selection algorithms don't scale up for ultrahigh dimensionality datasets used in information retrieval, text mining type of application. The stability concerning an unsupervised feature selection algorithm is also a challenge along with ML model selection. In the paper [20] authors have experimented with binary TLBO and found an improvement in the performance of LR and linear discriminate analysis model to predict student performance. Authors of paper [21] have proposed a novel wrapper-based feature selection algorithm that needs a common controlling parameter to obtain an optimal set of features. In the paper [22] authors have stated the feature selection is useful for discarding redundant, noisy, and irrelevant features. They have used three methods correlation-based(Pearson correlation coefficient), sequential feature selection (greedy search approach), and information gain (Decision Tree) based FS model. The sequential feature selection is slower than the correlation-based approach.

3 Proposed methodology

3.1 Design

Here we have used various FS models and TLBO to find the optimal set of features. To boost the ML model's efficiency we have customized the hyperparameter. Research questions(RQ):

- **RQ1**—what characteristics influence the employability of a student?
- **RQ2**—which ML classifier gives optimal prediction?

To answer these questions we have proposed a framework as shown in the following Fig. 1.

As shown in the diagram the proposed model has the following steps:

Step 1: Import the student information dataset from the institute's MIS. It will contain academic, socio-demographic, co-curricular, and extra-curricular, and placement details.

Step 2: Data preprocessing will include cleaning, transformation, factor analysis, and normalization.

Step 3: Apply different FS and rank all the features as per their co-relation with the target variable. Select all features with correlation>0.4. If rank>3 will be selected for further processing.

Step 4: Apply the TLBO to select the best feature based on Eq. (1). We will design the ML model and consider the feature set as population and the best feature will be selected randomly based on fifty iterations.

Step 5: Evaluate the performance based on Eq. (2) and find the best classifier with the highest accuracy.

Step 6: To improve the accuracy we fine-tuned the hyperparameter and evaluated the ML model again.

Step 7: We then customized the best ML model i.e. SVM by experimenting with the kernel function.

Algorithm 1 Calculate feature-target correlation

Algorithm 1 Calculate Feature-Target Correlation

Data:DataseC) with features and a target variable
Result:Select Optimal feature
Load the dataset
for each feature X_i **do** Calculate the correlation between X_i and the target variable Store the correlation coefficient Y_i
return Correlation coefficients for all features
if $Y_i \geq 0.4$ **then** $Feature\_count \leftarrow Feature\_count + 1$ Else reject
for each Feature_count ≥ 3 **do** analyze the ML model's performance and after 50 iterations we get best features

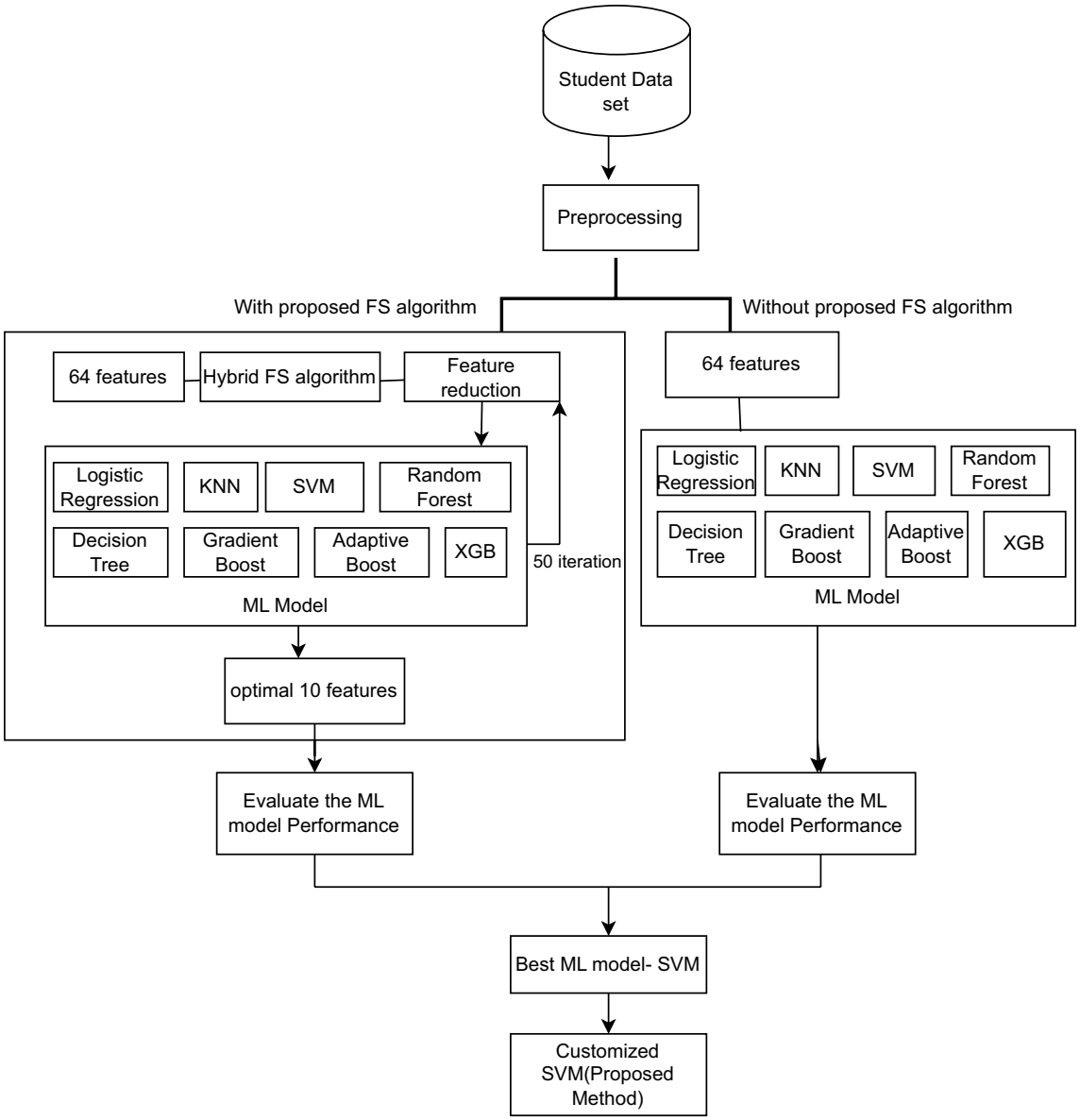

Fig. 1 Proposed method

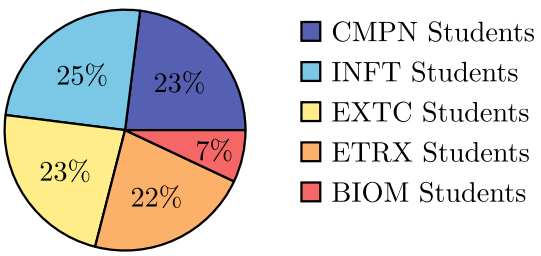

Fig. 2 Distribution of students

4 Experimental setup

We have written all programs in Python programming language using Numpy, Pandas and Scikit-Learn.

4.1 Data description

In this research article, we have collected the student details from our Institute's MIS for the 2018 to 2022 pass-out year. For the employability prediction, we have considered students, enrolled in different branches like Computer Engineering(CMPN), Information Technology(INFT), Electronics and Telecommunication(EXTC), Bio-medical Engineering(BIOM), and Electronics Engineering(ETRX). The Fig. 2 shows student distribution across all the departments. There are 57% male and 43% female students. The features which we have considered as academic factors like their 10th and 12th standard/diploma mark and year of passing, engineering semester marks and year of passing,

backlog details(KT count), further studies(information retrieved from students who have applied for Transcript for pursuing MS/MTech/MBA). Socio-demographic factors like are they OHU (other than home university) means they must be far from their family, their parent's qualification and profession. We also considered factors like their participation in Pre-placement activities, coding competitions like Hackathon, and extracurricular activities like sports, cultural etc.

4.2 Data pre-processing

The pre-processing stage of the machine learning process involves cleaning, normalizing, and transforming data.

1. One-hot encoding: Using this categorical values were transformed into numerical values. As a result, some new features were added.
2. Factor analysis: Using exploratory factor analysis we brought down the number of variables. As per the authors of the paper [23], the relation between variable/features size and sample size should be at least 1:5 whereas ideally, it should be 1:20. We have details about 1647 diploma holders and 64 features so we have an ideal dataset.
3. Normalization: The Min-max scaler and all numeric values were in the range of 0 and 1.

4.3 Feature selection

In this experiment, we have used a novel approach to select features. We proposed this approach in our earlier work [24] to select the optimal set of features for predicting the performance of students. In this work, we used the filter method (Pearson correlation and Chi-Square) the Wrapper method(Random forest (RF), Linear regression (LinR), Recursive feature elimination (RFE), Gradient boosting Method (GBM))and the hybrid method (Lasso and Ridge method). Table 2 depending on several approaches, displays the association between each feature and the goal feature. We have counted the features if it has a co-relation above the threshold > 0.4. Now based on the feature selection count (FS count > 3) we have selected the best feature that shows the highest co-relation with the target variable.

4.4 Supervised machine learning model

Once the selection of optimal features was done we evaluated the performance of eight machine learning models like LR, K-Nearest Neighbours Classifier (kNN), Support Vector Classifier (SVM), Decision Tree Classifier (DT), Random Forest Classifier (RF), Gradient Boosting Classifier (GB), Adaptive Boosting Classifier (AB), and Extreme Gradient Boosting Classifier (XGB). We evaluated the performance twice on the dataset to compare the performance of ML model with and without the proposed feature selection methodology.

4.5 Evaluating the performance of ML model and result

We have evaluated the different ML models based on accuracy, precision, F1 score, and Area under the curve(AUC) score with and without the application of the proposed feature selection method. The Tables 3 and 4 shows how each model performed on the dataset without and with the proposed feature selection method.

As per Tables 3 and 4 in some ML models the performance may be better if we select all 64 features but ideally it is not feasible to capture these many features so we selected around 10 features based on 50 iterations of the proposed feature selection algorithm and then found out how these ML model performed. SVM has performed better than other ML model. To improve the accuracy further we experimented with the hyperparameters of these ML models. We tuned the hyperparameters of the SVM, LR, and RF algorithm as shown in Table 5.

4.6 Proposed customized SVM ML model for better accuracy and result

With all these experiments we found that the SVM ML model gives us the best accuracy and it is suitable for our dataset. Based on these findings we tried to customize the hyperparameter of SVM. We experimented with different kernel functions like radial basis function (RBF), polynomial (Poly) and linear by importing respective libraries from various python libraries. We also experimented with hybrid kernel and found these results Table 6.

We found that the highest accuracy is achieved if the RBF is used as the kernel function. So we tried customizing the same. As discussed in [25] A supervised machine learning algorithm called SVM is primarily used for classification, while it can also be used for regression. SVM uses an ideal hyperplane that optimizes the margin between the classes to categorize the observation. SVM has two variants linear and non-linear. In our study, we have used linear SVM. The complexity of linear SVM depends on the number of features. In Fig. 3 there are two classes + and − and observations should be classified in these two classes. In a linear SVM, two types of margin need to be maximized: soft margin, which allows for some misclassification of fresh data, and hard margin, which allows for no training errors.

There are a few essential parameters of SVM that determine the performance of the model. As discussed in paper

Table 2 Feature's Rank

Feature	Chi-Square	Lasso	LinR	Pearson	RF	RFE	Ridge	GBM	FS count
Sem4 marks	0	0.78	0	1	0.93	0.44	0.28	0.85	5
F Non-Working	0.46	0	1	0.04	0.02	0.77	0.47	0.04	4
F Working	0.85	0	0.96	0.04	0.02	0.78	0.47	0	4
Sem5 marks	0	1	0	0.96	1	0.3	0.24	0.79	4
F-Govt	0.71	0	0.04	0.02	0.05	0.64	0.44	0.02	3
F-mechanic	0.52	0	0.04	0.04	0.03	0.72	0.47	0.01	3
M-Education	0.49	0	0.11	0.04	0.01	0.86	0.61	0	3
M-Engineer	0.68	0	0.11	0.02	0	0.84	0.51	0	3
IS-INFT	0	0	0.12	0.45	0.08	0.48	0.92	0.08	3
IS-CMPN	0	0	0.12	0.55	0.07	0.47	1	0.06	3
IS-ETRX	0	0	0.12	0.63	0.12	0.53	0.75	0.07	3
Sem3 marks	0	0	0	0.92	0.51	0.36	0.28	0.81	3
Sem6 marks	0	0.21	0	0.87	0.41	0.11	0.02	0.95	3
Sem8 marks	0	0.16	0	0.51	0.48	0.31	0.31	0.76	3
SSC marks	0	0.07	0	0.52	0.53	0.03	0.02	1	3
Diploma marks	0	0.35	0	0.87	0.7	0.08	0.03	0.99	3
Pass out Year	0.96	0	0	0.09	0.02	0.39	0.73	0.04	2
OHU	0.11	0	0	0.09	0.01	0.42	0.84	0.02	2
F-Accounts	0.12	0	0.04	0.09	0.01	0.75	0.6	0	2
F-Agri	0.18	0	0.04	0.08	0.03	0.59	0.5	0.04	2
F-Arts	0.95	0	0.04	0	0.01	0.73	0.09	0	2
F-Business	0.94	0	0.04	0	0.04	0.67	0.17	0.03	2
F-Finance	0.12	0	0.04	0.09	0.02	0.56	0.87	0	2
F-Engineer	0.86	0	0.04	0.01	0	0.62	0.02	0	2
F-Healthcare	0.88	0	0.04	0.01	0.02	0.66	0.36	0	2
F-Law-Order	0.44	0	0.04	0.04	0.02	0.61	0.07	0	2
F-manufacturing	0.14	0	0.04	0.08	0	0.55	0.52	0	2
F-Office	0.69	0	0.04	0.02	0.03	0.7	0.16	0.02	2
F-Others	0.64	0	0.04	0.03	0.04	0.69	0.16	0.1	2
M-Agri	0.68	0	0.11	0.02	0	0.83	0.23	0	2
M-Arts	0.14	0	0.11	0.08	0	0.98	0.83	0	2
M-Business	0.88	0	0.11	0.01	0	0.91	0.08	0	2
M-Finance	0.18	0	0.11	0.07	0	0.81	0.51	0	2
M-Govt	0.2	0	0.11	0.07	0	0.97	0.47	0	2
M-Healthcare	0.92	0	0.11	0	0.01	0.88	0.3	0	2
M-Law Order	0.2	0	0.11	0.07	0	1	0.84	0	2
M-mechanic	0.46	0	0.11	0.04	0	0.95	0.26	0	2
M-Non-Working	0.81	0	0.03	0.04	0.01	0.89	0.26	0.02	2
M-Office	0.84	0	0.11	0.01	0	0.94	0.01	0	2
M-Others	0.65	0	0.11	0.02	0.02	0.92	0.13	0	2
M-Working	0.47	0	0.14	0.04	0.01	0.45	0.26	0	2
IS-BIOMED	0	0	0.12	0.28	0.04	0.52	0.55	0.03	2
IS-EXTC	0	0	0.12	0.21	0.13	0.5	0.62	0.06	2
Sem7 marks	0	0	0	0.57	0.39	0.16	0.1	0.71	2
Sem6 Year	0.93	0	0	0.17	0.02	0.41	0.39	0	2
Gender	0.67	0	0	0.03	0.07	0.34	0.56	0.12	2
SSC pass out year	0.8	0	0	0.43	0.09	0.09	0.22	0.16	2
HSC pass out year	0.8	0	0	0.42	0.07	0.25	0.01	0.05	2
F Education	0.21	0	0.04	0.07	0.02	0.58	0.38	0.02	1
M Accounts	0.18	0	0.11	0.07	0	0.8	0.36	0	1
M-manufacturing	nan	0	0.79	nan	0	0.28	0	0	1

Table 2 (continued)

Feature	Chi-Square	Lasso	LinR	Pearson	RF	RFE	Ridge	GBM	FS count
Sem1 marks	nan	0	0.51	nan	0	0.17	0	0	1
Sem3 Year	0.9	0	0	0.23	0.07	0.2	0.26	0.11	1
Sem4 Year	0.96	0	0	0.09	0.03	0	0.03	0.05	1
Sem5 Year	1	0	0	0	0.04	0.12	0.08	0.06	1
Sem7 Year	0.96	0	0	0.11	0.02	0.23	0.21	0.01	1
Sem8 Year	0.94	0	0	0.15	0.02	0.22	0.28	0.01	1
Extracurricular	0.59	0	0	0.03	0.01	0.19	0.02	0	1
Diploma pass out year	0.89	0	0	0.27	0.03	0.27	0.35	0.03	1
Is Preplacement	0	0	0	0.35	0.04	0.33	0.45	0.05	1
Further studies	0.17	0	0	0.08	0.04	0.05	0.06	0.02	0
Sem2 marks	0.07	0	0	0.04	0	0.38	0.08	0	0
IsHackathon	0.24	0	0	0.07	0	0.14	0.08	0	0
KT count	0	0	0	0.32	0.05	0.06	0.06	0.05	0
HSC marks	0	0	0	0.04	0.1	0.02	0.01	0.1	0

Table 3 ML model performance evaluation without feature selection

ML model	Accuracy	Precision	Recall	F1score	AUC score
LR	72.16	65.28	60.64	62.88	71.06
kNN	71.34	63.56	54.89	58.92	69.34
SVM	**73.36**	**68.82**	**53.74**	**60.37**	**69.64**
DT	64.46	50.77	41.67	45.67	56.24
RF	67.32	57.93	46.72	51.78	62.60
GB	72.33	63.81	55.74	59.50	69.30
AB	68.73	58.41	54.77	56.53	65.53
XGB	68.93	59.33	53.04	56.02	65.30

SVM ML model is giving highest accuracy so it is written in bold

Table 4 ML model performance evaluation with feature selection

ML model	Accuracy	Precision	Recall	F1score	AUC score
LR	73.96	67.27	56.77	61.59	70.16
KNN	71.14	69.54	39.53	50.62	65.90
SVM	**74.37**	**66.13**	**61.36**	**63.66**	**71.39**
DT	69.32	57.10	60.21	69.62	67.23
RF	73.96	67	57.34	61.81	70.16
GB	73.36	67.58	53.32	59.61	68.77
AB	73.76	66.03	59.49	62.04	70.27
XGB	72.35	64.33	55.04	59.34	68.39

SVM ML model is giving highest accuracy so it is written in bold

Table 5 Accuracy

Hyper parameter tuned ML Model	Accuracy
Support Vector Machine	77.37
Logistic Regression	76.16
Random Forest	74.34

Table 6 Modified SVM model

SVM model with different kernel function	Accuracy
SVM with kernel = RBF	77.37
SVM with kernel = Poly	73.73
SVM with kernel = Linear	74.74
SVM with kernel = hybrid	74.55

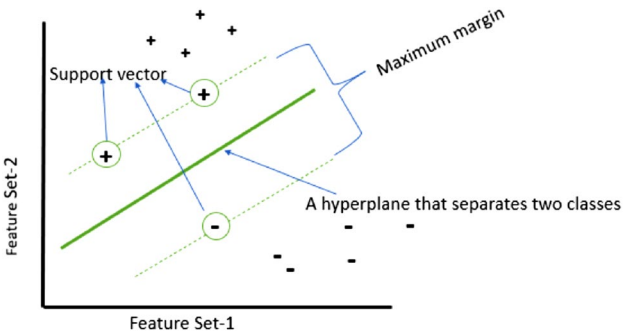

Fig. 3 Graphical representation of SVM

[26] authors have proposed a model to improve RBF kernel by hypertuning the soft margin value of C and gamma value for the classification of tweets based on polarity. With the help of WordNet, more than one polarity was calculated for opinion mining. The value of C (Regularization parameter) aids in determining the balance between margin and training error. It helps in controlling the error. Hard margin has a larger value of C and soft margin has a lower value of C. For our dataset parameters of SVM were experimented with different value of C like 100, 150, 200, 1000 and found C = 100 as best. We experimented with different kernel functions like linear, radial basis function(RBF), and polynomial and

found RBF as best. The Kernel function helps to take data as input and use mathematical functions to transform it into the required form. It provides you with the standard feature dimension's inner product of two points. Lastly, we checked with the gamma hyperparameter. This value we should set before training the model as it determines how much curvature we want in the decision boundary.

We tried modifying the RBF function with different distances. The radial distance between the elements of each pair of data points is customized using Euclidian as discussed in Eq. (4) and pairwise distance in Eq. (4), Customized Radial distance discussed in Eq. (6) and pairwise distance between features in Eq. (6), Manhatten discussed in Eq. (8), and Hamming distance discussed in Eq. (9) and it is found that RBF as a kernel function with Euclidian distance gives the highest accuracy.

Euclidian Distance: the radial_distance contains the pairwise Euclidian distance between each pair of data items in X and Y, where the Euclidian distance is calculated as the square root of the sum of squared feature differences.

$$radial\_distance = np.sqrt(np.sum(pairwise\_diff \ ** \ 2, axis = 2)) \tag{4}$$

$$pairwise\_diff = np.abs(X[:, np.newaxis] - Y) \tag{5}$$

Customized Radial Distance: The radial_distance matrix represents a measure of similarity or distance between the data points in X and Y, with the exact nature of this measurement being controlled by the value of gamma. The value of gamma is a hyperparameter that influences the weighting of the absolute differences and, in turn, the similarity or distance measurement. Smaller values of gamma will give less weight to feature differences, while larger values will emphasize feature differences more. we have set the value of gamma as 1.

$$radial\_distance = np.sum(pairwise\_diff \ ** \ gamma, axis = 2) \tag{6}$$

$$pairwise\_diff = np.abs(X[:, np.newaxis] - Y) \tag{7}$$

Manhatten Distance: The pairwise Manhatten distance between each pair of data points, X and Y, is contained in the distance matrix.

$$distance\_matrix = np.sum(np.abs(X[:, np.newaxis] - Y), axis = 2) \tag{8}$$

Hamming Distance: It counts the number of places where comparable elements in two binary strings differ from one another.

$$hamming\_distance = np.sum(X[:, np.newaxis]! = Y, axis = 2) \tag{9}$$

The Table 7 shows the accuracy obtained after this customization.

The Table 8 shows the suggested model's output in contrast to existing work. As discussed earlier in the proposed work the sample size and feature set are three times more than the existing work, still the accuracy is higher because of the proposed customization in the SVM model.

Table 7 Customised RBF

Customised RBF	Accuracy(%)
Euclidian distance as radial distance	87.80
Customized radial distance	86.08
Manhatten distance as radial distance	86.00
Hamming distance as radial distance	79.50

5 Discussion

In this paper, we have analyzed 1647 Direct second-year students(Diploma completed) who had enrolled in an engineering college located in a metropolitan and cosmopolitan city. We have analyzed the placement record of 5 years 2018, 2019, 2020, 2021, 2022 pass-out years. A four-year engineering course is divided into eight semesters the students who are considered here in this research work are in the direct second year which means they have enrolled in semester 3 directly instead of semester 1. The overall placement has been affected by the pandemic during 2020, 2021, and 2022, as the off-campus recruitment drive was reduced, and

Table 8 Comparison of proposed model with existing model

Reference	Methodologies	Best Model	Accuracy
Mishra et al. [9]	J48, Random Forest, Naive Bayes, Multilayered Perceptron, Sequential Minimal Optimization	J48	70.19%
Saidani et al.[10]	XGBoost, CatBoost, LGBM	LGBM	77.53%
Proposed Work	LR, kNN, Random Forest, SVM, XGB, AdaBoost, DT, Gradient Boost	Customized SVM	87.8%

student's academic activities and pre-placement activities were conducted in either hybrid or online mode. Although it is found that few students have received multiple offers for example 21 students in the Computer department, 25 students in the INFT department, and 17 students in EXTC have also received more than two offers. Students were placed in different sectors like—Information Technology, the Healthcare sector, Ed-Tech companies, and the financial sector, etc. Post pandemic there is growth in placement offers in healthcare sectors. Through this experiment setup, let's answer:

- RQ 1: What characteristics influence the employability of a student? Here in this research work, We determined the relationship between every feature and the desired variable. The factors correlating more than 0.4 were selected and created a feature rank pool. The features with rank 3 were selected. We found that features like semester 4 and 5 marks of eight semesters, diploma marks, specialization, or branch of engineering, and socio-demographic factors like their parent's education and occupation also affected student's employability. We also experimented with other factors like their 10th and 12th standard marks, being an outside-home university(OHU) student, extra-curricular activities, etc. In our dataset, it was found that these factors did not contribute much to employability.
- RQ2: Which ML classifier gives optimal prediction? To answer this question we have experimented with different ML models both with and without the proposed feature selection algorithm. It was found that in both cases the Support vector machine classifier is performing better in comparison to other ML models. So we tuned the SVM by customizing its kernel function. After tuning we found a rise of 14% in the accuracy of the ML model.

6 Conclusion and future work

This paper aims to find the optimal factors which affect employability. In this research work, we have worked with a real live dataset of students who have passed out in the years 2018,2019,2020,2021, and 2022. During these years there were different pedagogical approaches like before the pandemic traditional/classroom, during the pandemic online, and in post-pandemic blended environments. Even the placement opportunities from the employers' side also had a significant effect on employability. We found that our proposed approach to finding the optimal set of features when fed to SVM gives the highest accuracy. Earlier without the feature selection algorithm, all 64 features were considered and we found the best accuracy by SVM as 73.36% and with the use of the proposed feature selection algorithm we found an optimal set of features and a better accuracy with 74.37%. We have further improved the performance of our SVM model by hypertuning the kernel function. We experimented with different kernel functions like RBF, Polynomial, linear, and hybrid but found RBF gives the highest accuracy and when its radial distance between each pair of data points was customized the model showed an improved accuracy of 87.8%.

We can extend this work in the future by adding a recommendation engine to prescribe measures to be taken by the institute and students at correct intervals to improve the placement records. A proper feedback system should be designed between institutes and recruiters (post-hiring) for the improvement of records. Although academic and socio-demographic factors significantly impact employability, we should consider quantitative aptitude, logical reasoning, technical skills, communication skills, management skills, decision-making skills, critical thinking, and indirect factors like the global economy, recession, pandemic, etc. as well. We should experiment by considering these factors to see whether academically average students can also obtain good placement through other abilities and virtues. We can conclude that for a successful career in the engineering domain, students should have today's skills and capability to adapt to the changing requirements.

Acknowledgements Authors would like to thank the Vidyalankar Institute of Technology for sharing the student data and for supporting the research work under the Higher Studies scheme of the Institute.

Author contributions Both authors have equally contributed to conceptualization, methodology, software, data curation, writing—reviewing, editing, visualization, investigation, and project administration.

Funding The authors declare that no funding was received to assist with the preparation of this manuscript or conducting this study.

Data availability The dataset analyzed during the current study is not publically available as per the policy set by the Institute.

Code availability The source code can be made available based on request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethics approval This research work is approved by the Institute.

Consent for publication The Institute gave informed consent to publish this research work.

Consent to participate The Institute gave informed consent to participate in this research.

References

1. Yang X-S (2020) Nature-inspired optimization algorithms: challenges and open problems. J Comput Sci 46:101104
2. Rao R, Patel V (2012) An elitist teaching-learning-based optimization algorithm for solving complex constrained optimization problems. Int J Ind Eng Comput 3(4):535–560
3. Mezhoudi N, Alghamdi R, Aljunaid R, Krichna G, Düştegör D (2023) Employability prediction: a survey of current approaches, research challenges and applications. J Ambient Intell Humaniz Comput 14(3):1489–1505
4. Casuat CD, Festijo ED (2019) Predicting students' employability using machine learning approach. In: 2019 IEEE 6th international conference on engineering technologies and applied sciences (ICETAS), IEEE, pp 1–5
5. Chandra MA, Bedi S (2021) Survey on svm and their application in image classification. Int J Inf Technol 13:1–11
6. Çakıt E, Dağdeviren M (2022) Predicting the percentage of student placement: a comparative study of machine learning algorithms. Educ Inf Technol 27(1):997–1022
7. Patel C (2018) Identification of essential parameters for post graduate students' job placement in computer applications in india. Int J Inf Technol 10(4):511–518
8. Kumar M, Walia N, Bansal S, Kumar G, Cengiz K (2023) Predicting college students' placements based on academic performance using machine learning approaches. Int J Modern Educ Comput Sci 15:1–13
9. Mishra T, Kumar D, Gupta S (2016) Students' employability prediction model through data mining. Int J Appl Eng Res 11(4):2275–2282
10. Saidani O, Menzli LJ, Ksibi A, Alturki N, Alluhaidan AS (2022) Predicting student employability through the internship context using gradient boosting models. IEEE Access 10:46472–46489
11. Sarzaeim P, Bozorg-Haddad O, Chu X (2018) Teaching-learning-based optimization (tlbo) algorithm. Adv. Optim. Nat.-Inspir. Alg. 2018:51–58
12. Zhou G, Zhou Y, Deng W, Yin S, Zhang Y (2023) Advances in teaching-learning-based optimization algorithm: a comprehensive survey. Neurocomputing 2013:126898
13. Karlupia N, Abrol P (2023) Wrapper-based optimized feature selection using nature-inspired algorithms. Neural Comput Appl 35(17):12675–12689
14. Tripathi A, Bharti KK, Ghosh M (2023) A fusion of binary grey wolf optimization algorithm with opposition and weighted positioning for feature selection. Int J Inf Technol 15(8):4469–4479
15. Khorashadizade M, Hosseini S (2023) An intelligent feature selection method using binary teaching-learning based optimization algorithm and ann. Chemom Intell Lab Syst 240:104880
16. Miao J, Niu L (2016) A survey on feature selection. Procedia Comput Sci 91:919–926
17. Zaffar M, Hashmani MA, Savita K, Rizvi SSH (2018) A study of feature selection algorithms for predicting students academic performance. Int J Adv Comput Sci Appl 9:5
18. Punlumjeak W, Rachburee N (2015) A comparative study of feature selection techniques for classify student performance. In: 2015 7th international conference on information technology and electrical engineering (ICITEE), IEEE, pp 425–429
19. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Comput Surv (CSUR) 50(6):1–45
20. Alraddadi S, Alseady S, Almotiri S (2021) Prediction of students academic performance utilizing hybrid teaching-learning based feature selection and machine learning models. In: 2021 international conference of women in data science at Taif University (WiDSTaif), IEEE, pp 1–6
21. Allam M, Nandhini M (2022) Optimal feature selection using binary teaching learning based optimization algorithm. J King Saud Univ-Comput Inf Sci 34(2):329–341
22. Sharma A, Mishra PK (2022) Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. Int J Inf Technol 2022:1–12
23. Comrey AL, Lee HB (2013) A first course in factor analysis. Psychology press, London
24. Jayachandran S, Joshi B (2024) Optimizing the feature selection methods using a novel approach inspired by the tlbo algorithm for student performance prediction. Adv Comput 2021:14
25. Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine Learning, pp 101–121. Elsevier, London
26. Gopi AP, Jyothi RNS, Narayana VL, Sandeep KS (2023) Classification of tweets data based on polarity using improved rbf kernel of svm. Int J Inf Technol 15(2):965–980

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.