2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Oct 13-14, 2022

2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-7433-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRITO56286.2022.9965116

Student Satisfaction and Churn Predicting using Machine Learning Algorithms for EdTech course

Rinika Paul
Department of Electrical and Electronics Engineering,
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Bengaluru, India
rinikapaul0506@gmail.com

MR. Rashmi
Department of Electrical and Electronics Engineering,
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Bengaluru, India
mr_rashmi@blr.amrita.edu

***Abstract*—EdTech courses are believed to provide industry-oriented practical skills to the students to bridge the gap between academics and the industry. They also provide supporting education to the students to crack competitive exams in various sectors. With education institutions being shut down due to lockdown and moved to online mode, many students enrolled themselves in EdTech courses online creating competition amongst EdTech organizations. Reduced prices at times lead to reduced quality of the course content and the delivery. Students are often the prey in the competition who are left unsatisfied with the course and would never return to the organization for further courses leading to a churn. At least 50% of college students enroll in online courses in India but the completion rate is only about 13% which is very low. Customer churn is a major challenge to an organization reducing their growth as well as revenue. Two different machine learning algorithms viz. K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) are proposed in this paper to predict the churn rate of students for EdTech courses based on the feedback of the students collected through the course end survey. The model also tells if the student is interested to enroll further in any other courses offered by the organization or refer the courses to their friends or relatives. The data is collected in real-time based on the online courses offered by a startup company and the churn rate for the organization is calculated.**

Keywords— Machine Learning (ML), Student Satisfaction, Churn Prediction, Student Attrition, E-Learning, EdTech Course, Support Vector Machine (SVM), K-Nearest Neighbor (KNN)

I. INTRODUCTION

The year 2020 has seen the onset of the Covid-19 pandemic that had shattered the lives of humans forcing them to stay in their homes for months. The world came to a stagnant with industries, educational institutions, and libraries closed. The situation created a negative impact on students who are the future of our country. The second decade of the 21st century has witnessed some great fast-paced innovations in terms of digitization. Digitization has reframed the world with more connectivity, networking, and transactions and the onset of massive online courses has changed the lives of students. India has seen significant growth in online learning as the internet and education have merged to provide people with the possibility to master new skills. But during the Covid-19 pandemic, the growth of e-learning has reached its peak. The pandemic has caused schools, institutions, and businesses to operate remotely, which has increased the use of online learning and virtual classrooms. A web-based distance learning program aims at huge groups of students from all over the world typically utilized for higher education and professional development. However, many public school districts and undergraduate degree programs have adopted MOOCs as the new standard to keep up with the advances in technology. The Ministry of Education, India launched the New Education Policy in 2020 which has brought about changes in the way Indian Education System existed and is now experimental learning and critical thinking. With the shift in the conventional education system and emerging online courses for upskilling, each organization aims to attract students through tailor-made courses. About 4335 EdTech startups were alone founded in 2020 – 2021. This increases the competition between organizations to attract more students for enrollment thus providing huge discounts. The term EdTech is a combination of "Education with "Technology" that defines new-age technologies included in learning methodologies apart from traditional classroom learnings. This may include better hardware and software such as tablets or animations and video-based learning and even online courses.

The goal of EdTech courses is to provide quality education to the students to shape their future and provide them with the required skillset [1]. Many students have high expectations for the classes they are enrolled in. They begin by believing that they can complete the course without putting up much effort and yet learn a lot of new things [2]. When the course is not up to the mark, many will withdraw from the course in the middle of the program. They eventually drop out since they can't keep up with the course's demands. At least 50% of college students enroll in online [3] courses in India but the completion rate is only about 13% which is very low. If an E-learning platform, if course delivery is inadequate, it might result in a high student turnover rate. Problems with technology can lead to poor course delivery that will need effective methods to improve [8]. It could also be related to a breakdown in communication between students and professors leading to dropouts or students failing to organize themselves and manage time and hence failing to cope with the certificate courses. When students enroll in a course, they have certain expectations about what they will learn and how the course will proceed. If they discover that what they expected is different from how the class is run or what they are learning is different from what they expected, a student may churn [7]. In each of these cases, it is a loss for the EdTech company as it costs 5 times more to acquire new customers than to retain an existing one [13].

Customer churn is the percentage of customers that discontinued using the company's products or services in a certain period. Customer Churn prediction is identifying the non-potential customers of an organization who are at risk of discontinuation of any services or products produced by the organization [3-4]. An ideal Churn rate should be as close to 0% for any organization. Machine Learning provides computers with the intelligence to learn from data and past experiences to make predictions with minimal human interventions. The application of machine learning is vast and has been used currently in medical fields for early-stage disease detections, weather forecasts, social causes, market analysis, customer predictions, and much more. Several ML techniques are used for classification and regression that has been applied in areas such as water monitoring, environmental pollution predictions, weather forecast, business analysis, and any predictions of huge data [9 - 12]. Classification methods are required to determine student satisfaction with a certain course which includes Support Vector Machines [4] Neural Networks, Decision Trees, and many more. However, KNN has proven to be the best machine learning algorithm that can be used for multiclass classification in any application [5] [6]. It is also believed that no single algorithm can satisfy all business goals. Therefore, a specific algorithm is apt for a specific application, or in certain cases, an ensemble of many algorithms proves to provide the best accuracy.

In this work, student satisfaction on an EdTech course is predicted based on the feedback data obtained through the course feedback filled by the students. The designed model can predict that if the student is satisfied with the course, they would continue with other courses offered by the organization or would refer their friends and relatives, or else the student wants to discontinue. The real-time data is obtained from students' feedback on certain courses provided by Zikshaa, Odisha's leading EdTech startup. Zikshaa focuses on research and development in Electric vehicles, robotics, unmanned aerial vehicles, Artificial Intelligence, Machine Learning, Cybersecurity, and other tailor-made courses. Although the startup has upskilled over 3000 students in 2 years, a certain fluctuation in the course enrollment is observed. Hence to understand the student's perspective, satisfaction, and feelings, feedback is collected and analysis is made to determine the overall churn for the Zikshaa courses. The model developed helps Zikshaa academy to identify dissatisfied students and contact them personally for improvement and tailoring of courses. The algorithm used to predict combines K-Nearest Neighbor with Support Vector Machine, an ensemble framework to increase the model accuracy and faster prediction.

Further discussion in the following sections talks about various steps in measuring student satisfaction and churn prediction. Section II describes the complete flow of the designed system and the data processing techniques right from data collection and preprocessing to the data split. In section III, the machine learning models have been discussed and the results are obtained.

II. Methodology

The workflow for the model built to identify the satisfaction of a student on a course provided by the EdTech company and the churn that occurs for the application is shown in Figure 1. The steps involved in the model development include:

1. Identifying data sources: Feedback forms for a few courses are collected from Zikshaa which gives the overall idea about the satisfaction of the student.
2. Data preparation and pre-processing: As the datasets are not completely balanced and have null rows, it needs to be preprocessed and null data should be filled and unnecessary rows and columns should be dropped to remove inconsistencies. Feature engineering, extraction, and selection processes are a part of preprocessing and data visualization.
3. Training ML model: SVM classification is used for binary classification for churn prediction. K Nearest Neighbor is used for multiclass classification for predicting if a student may refer Zikshaa to her friends or if a student may enroll in the next course.
4. Data Analysis

- Identify if the student is satisfied, dissatisfied, or neutral
- Predict the churn rate
- Predict if the student will recommend the course

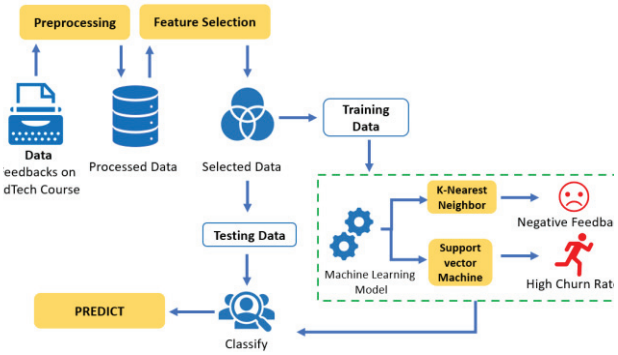


Fig. 1. Data Flow of the proposed model

A. Data Collection

Before moving to the model development, an understanding of the data to be fed into the model is required. Feedback from around 495 students who enrolled for different courses provided by Zikshaa is collected. The data consisted of the date of the feedback, course, and trainer ratings from 0 to 5 with 0 being the least which were mandatory. Some optional questions such as future recommendations of Zikshaa ad interest in further enrollment were asked.

Date Submitted	How satisfied were you with the workshop?	How relevant and helpful it was for your skill?	How satisfied were you with the Knowledge of speaker	How satisfied were you with the session conten	Would you recommend your friends to Zikshaa?	Would you like to enroll for our next Workshop?
09-17-2020	4	3	4	3	no	maybe
09-17-2020	4	5	5	5	yes	maybe
09-17-2020	4	4	4	4	maybe	maybe
09-17-2020	2	3	3	2	no	no
09-17-2020	5	5	5	5	yes	yes
09-17-2020	2	3	4	3	maybe	yes
09-17-2020	3	3	3	3	maybe	no
09-17-2020	2	3	3	3	no	no

Fig. 2. Raw Data – Feedback

A student status column was developed that monitored if the student continued for further courses or never came back. The optional questions were not filled by all and thus data set had null cells. Although the dataset is uniform it is a mix of numbers and characters. Data preprocessing is necessary before feeding into the model to remove inconsistencies. Figure 2 shows the few fields of feedback data collected from 495 students which represents the raw dataset.

B. Data Preprocessing

Understanding the available data for analysis is the first step in data preparation. There are 11 fields in student feedback data. Determining which fields may be utilized for classification is one of the most significant responsibilities. There are category data variables such as Course satisfaction, Doubt clearing session satisfaction, and others. Some of these elements, such as Additional Comments and Date, may not be necessary for analysis. Null values in the data must be handled properly if they are detected.

Data Preparation is a tedious job and hence Python has inbuilt libraries such as NumPy, Seaborn, Pandas, and Matplotlib that are used in the present application for preparing the data before using it in the model. The columns that are not used in the classification model are dropped and the empty data is filled with the mode of the feature column values. Satisfaction of the workshop, knowledge of the speaker, and other categorical data variables are transformed into matching columns. Mathematical operations cannot be performed on categorical data such as No/Maybe/Yes. As a result, these categorical data columns become columns with 0 or 1 or 2 entries replaced using the One-hot Encoding technique. After data preparation, the count of dissatisfied students is replaced by 0, neutral feedback is replaced by 1 and positive feedback is replaced by 2.

C. Feature Extraction

While features are important to developing an ML model, too many features can harm the accuracy of the model. Hence the optimum number of necessary features should be selected to feed into the model. The Feature Extraction method reduces the dimension of the data and provides us with exactly the columns in the dataset that affect the classification of the students into exactly 3 classes – positive, neutral, and negative.

TABLE I. STATISTICS ON FEATURE COLUMN FEEDBACK

Features	Mean	Min	Max	IQR – 25%	IQR – 75%
Workshop	3.775758	2	5	3	5
Content	3.876768	2	5	3	5
Skill Acquisition	3.885947	2	5	3	5
Speaker	3.927273	2	5	3	5

Four feature columns are identified that give the student's view on the overall workshop, relevance as a new skill acquired, content, and the speaker of the workshop as shown in Figure 3. This will help to identify whether the student likes to continue further with other courses. Table 1 shows the statistical analysis of the feedback obtained in the feature columns. From the dataset, test and train data are split in the 20:80 ratios respectively indicated in the pie chart in Figure 4. The test data represent the part of the new feedback in which the model will be used to predict student satisfaction in the future.

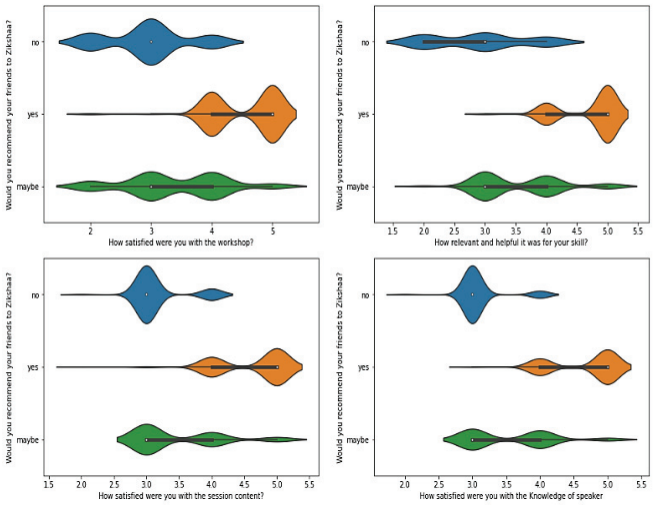


Fig. 3. Feature Column data visualization

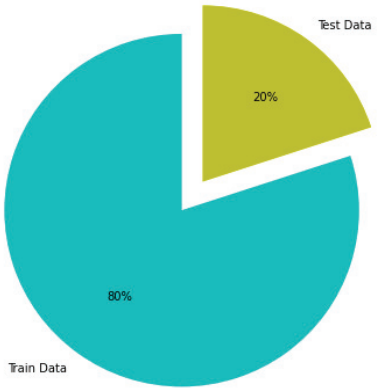


Fig. 4. Test Data and Train Data Splitting

III. MODEL FORMATION

A. K-Nearest Neighbor Algorithm

KNN is a supervised machine learning algorithm best suited for multiclass classifications but used for both regression and classification problems. It uses 'feature similarity' to predict the values of new data points that assign a value based on the closeness of matching points in the dataset. The best value of K is chosen i.e., the nearest data points by a method called the "Elbow Method".

$$d = \sum_{i=1}^{n} | Xi - Yi | \quad (1)$$

To find the optimum K value, Manhattan distance 'd' indicated in (1) is used as it calculates the distance between two real-valued vectors X_i and Y_i. For each point in the test data,

- The distance between the test data and each row of training data is calculated with the help of the Manhattan distance.
- Based on the distance value, ascending order sorting of the data is performed.
- The top K rows are chosen.

- A class is assigned to the test point based on the most frequent class of these rows.

The Manhattan distance of the features is plotted with k values ranging from 0 to 10 as shown in Figure 5. The blue line represents the best K value for determining if the student will enroll in the upcoming courses offered by Zikshaa. The red line represents the K value to be chosen to predict if the students will recommend the courses to others. The highest K value obtained is 5 and 6 respectively that is used in the KNN model. KNN algorithm is used to predict the satisfaction of the student regarding the course. Along with KNN, the SVM algorithm is also used.

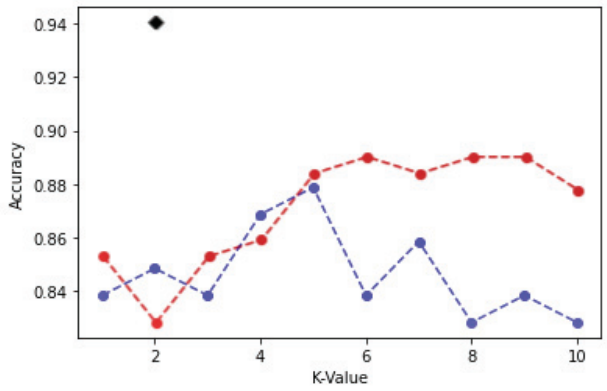


Fig. 5. K value for Manhattan Distance

B. *Support Vector Machine*

SVM is another supervised learning algorithm used mainly for binary classification. The algorithm finds the optimum hyperplane in an N-dimensional space that separates the data points into distinct classes. The Radial Basis Function (RBF) kernel function given in (2) is used for two points X_1 and X_2 that computes the similarity or how close the data points are to each other.

$$K(X1, X2) = \exp\left(\frac{-|X1-X2|^2}{2\sigma^2}\right) \quad (2)$$

where 'σ' is the variance and $||X_1 - X_2||$ is the Euclidean distance between the two points X_1 and X_2. SVM with RBF Kernel is used to classify if the student will churn or retain.

Scikit-learn library for python is an open-source tool specially designed for machine learning applications that contact various algorithms for classification and regression. This library is used to develop KNN and SVM models for the current application.

C. *Performance Metrics*

For every classification problem using a machine learning algorithm, there should be a method to evaluate the performance of the classifier. A confusion matrix is one such performance metric that is used to evaluate model performance and accuracy. In the student satisfaction and churn analysis, the confusion matrix has multiple classes – Yes, No, Maybe for course enrollment and recommendation.

TABLE II. THREE-CLASS CONFUSION MATRIX

		Predicted Class		
		A	B	C
True Class	A	TP_A	E_{AB}	E_{AC}
	B	E_{BA}	TP_B	E_{BC}
	C	E_{CA}	E_{CB}	TP_C

A confusion matrix is a table with dimensions - "Actual" and "Predicted" classes that have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)" and "False Negatives (FN)". The number of classes determines the size of N in a N x N confusion matrix.

- True Positive and True Negative mean that the actual class is the same as the predicted class.
- False Positive and False Negative mean that the actual class and the predicted class are not the same.

Classification Accuracy refers to the ratio of several correct classifications made concerning the total number of predictions made. The accuracy of the classification should be near 100% of the best model. To measure the churn rate, accuracy shouldn't be considered as the majority of the students will tend to retain back which will already show high prediction accuracy. In such cases, precision and recall are better measures.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

A classification reports consist of Precision rate, Recall rate, and F1 score. The precision determines how many of the Positive predictions are positive and recall determines how many of the Positive predictions are identified wrongly. To measure the accuracy of the model, the recall rate should be high to classify the students who are prone to churn.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

F1 score is the harmonic mean of both precision and recall so that equal weights are assigned to each of them and ignore the impact of outliers. A high F1 score is an indication of student satisfaction where precision rate and recall rate are equally important.

IV. RESULTS & ANALYSIS

The feature variables in student satisfaction data after the initial data cleaning are obtained. All other variables are used as predictor variables and the feature columns are:

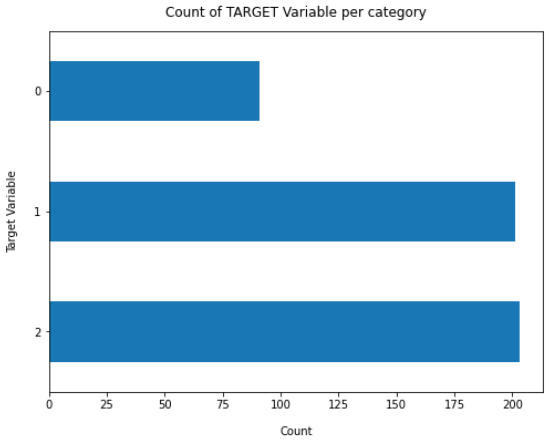
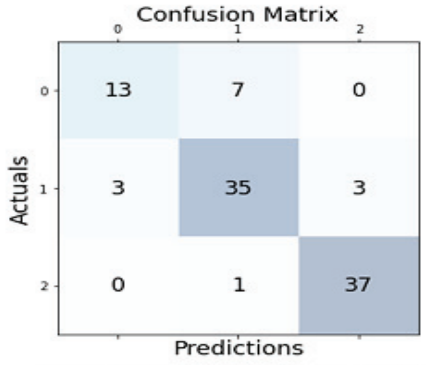


Fig. 6. Count of student satisfaction

- How satisfied were you with the workshop?
- How satisfied were you with the session content?
- How relevant and helpful it was for your skill?
- How satisfied were you with the Knowledge of the speaker?

The features are introduced to the model for training and K- Nearest Neighbor is used to train the dataset to identify the satisfaction of the students after they have completed the course and SVM is used to predict if the student is prone to churn. To feed into the model, the dataset in the form of a data frame is converted to a Numpy array. Figure 6 displays that out of 495 students who enrolled for the course, 89 students are dissatisfied (0) with the overall course, and about 195 students have neutral feedback (1) wherein they dint completely enjoy the course but don't have negative feedback either. It is also observed that 211 students would like to continue with further courses and are happy with the knowledge Zikshaa provided.



	precision	recall	f1-score
0	0.81	0.65	0.72
1	0.81	0.85	0.83
2	0.93	0.97	0.95
accuracy			0.86
macro avg	0.85	0.83	0.83
weighted avg	0.86	0.86	0.86

Fig. 7. KNN model accuracy

It is observed that the model has an F1 score of 0.95. In this case, it is only related to whether a student will recommend Zikshaa to other friends or a student will opt for the next course. As prediction is all about finding negative feedback, accuracy plays a lesser role. F1 score is a harmonic mean of both precision and recall which are 0.93 and 0.97 respectively. Therefore, a high precision and recall rate is required to measure the classification performance. Root Mean Square error obtained is 0.338738 which indicates that the data points are close to each other and have lesser deviation. There the model can be used to measure student satisfaction and churn rate. Another CSV file of feedback from 133 students who enrolled for a workshop is introduced to the developed model for prediction which is shown in Figure 8 and the results are as follows:

- 32 students are predicted to Churn out
- 16 students don't want to recommend Zikshaa to any of their friends or relatives.
- 22 students are not satisfied with the course and will not enroll in future courses.

The organization should personally contact the students who are at risk of churn to try to retain them in the future to avoid any kind of loss.

How satisfied were you with the workshop?	How relevant and helpful it was for your skill?	How satisfied were you with the Knowledge of speaker	How satisfied were you with the session content?	Name	Would you recommend your friends to Zikshaa	Would you like to enroll for our next Workshop?	Churn
4	3	4	3	Dhruv	no	maybe	1
4	4	4	4	Murali	maybe	maybe	0
2	3	3	2	Isha	no	no	1
2	3	4	3	Aryan	maybe	yes	0
3	3	3	3	Butala	maybe	no	1
2	3	3	3	Ankitha	no	no	1
3	3	3	3	Athul	maybe	no	1
5	4	5	4	Madhushankar	yes	maybe	0
3	3	4	4	Chahal	maybe	yes	0
3	4	3	3	Anagha	no	maybe	1
4	4	4	3	Chand	maybe	maybe	0
5	4	5	4	Aryan	yes	maybe	0
2	3	3	3	Amit	no	no	1
4	4	3	3	Sneha	maybe	maybe	0
5	4	5	4	Shreta	yes	maybe	0

Fig. 8. Predicted Data

V. CONCLUSION & FUTURE SCOPE

The work aims to predict the student satisfaction and the churn rate of those who enrolled for the online courses provided by an EdTech company so that the company can take precautionary measures and tailor their courses to retain the students. The churn rate is observed to be 24 % which is moderately high. This means that for every 100 students who enroll for a course, 24 students will never return and enroll for another course. A machine learning model was designed using KNN and SVM together to measure when the student is willing to enroll in new courses or refer their friends which helps understand the satisfaction of the student and determine the churn rate. The model helps the organization understand its customers better and methods to retain them. Based on the prediction, the organization can implement newer techniques to satisfy the students and retain them and grow their business and also help in the growth of the students.

The model can be modified in the future with deep learning algorithms for sentiment analysis and customization options to be provided to the students for a better customer experience. The written comments can be analyzed through LSTM networks in deep learning and improvements can be made by the company based on the analysis to attract more audience. It can also help the company to design new courses based on customer interest.

REFERENCES

[1] Indy Man Kit HoID, Kai Yuen Cheong, Anthony Weldon, "Predicting student satisfaction of emergency remote learning in higher education duringCOVID-19 using machine learning techniques," *PLOS ONE,* 2021.

[2] Ragad M Tawafak, Awanis BT Romli, Ruzaini Bin Abdullah Arshah, "E-learning Model for Students' Satisfaction in Higher Education Universities: Review Paper," in *International Conference on Fourth Industrial Revolution (ICFIR)*, 2019.

[3] Angitha A.U., Supriya M, "Ranking of Educational Institutions Based on User Priorities Using AHP-PROMETHEE Approach," *Advances in Computing and Network Communications, Lecture Notes in Electrical Engineering,* vol. 736, p. 127–142, 2021

[4] Sejal Badgujar, Anju S. Pillai, "Fall Detection for Elderly People using Machine Learning," in *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020.

[5] M. M. Ulkhaq, A. T. Wibowo, M. R. Tribosnia, R. Putawara, A. B. Firdauz, "Predicting Customer Churn: A Comparison of Eight Machine Learning Techniques: A Case Study in an Indonesian Telecommunication Company," in *International Conference on Data Analytics for Business and Industry (ICDABI)*, 2021.

[6] Akash Kumar, Aniket Verma, Gandhali Shinde, Yash Sukhdeve, Nidhi Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," in *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020.

[7] Soumi De, Prabu P, Joy Paulose, "Effective ML Techniques to Predict Customer Churn," in *Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021.

[8] Suresh K, Pooja M E, Meghana J, "Improvement of e-learning in Ontology using Machine Learning Techniques," in *Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021.

[9] T. M. Swetha, T. Yogitha, M. K. Sai Hitha, P. Syamanthika, S. S. Poorna, K. Anuraj, "IOT Based Water Management System For Crops Using Conventional Machine Learning Techniques," in *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021.

[10] B.N.Krishna Sai, T. Sasikala, "Predictive Analysis and Modeling of Customer Churn in Telecom using Machine Learning Technique," in *3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019.

[11] M. Monisha, A. Suresh, M. R. Rashmi, "Artificial Intelligence Based Skin Classification Using GMM," *Journal of Medical Systems,* 2019.

[12] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba and N. Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 349-353.

[13] Y. Liu, Z. Li, and H. Nan, "Motivation and Retention Strategies of Undergraduates' Online Course Learning Motivation in the Information Age," 2021 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE), 2021, pp. 74-77.