International Journal of Information and Communication Technology Education
Volume 20 • Issue 1 • January-December 2024

Integration of Collaborative Filtering Into Naive Bayes Method to Enhance Student Performance Prediction

Venera Nakhipova
South Kazakhstan Pedagogical University, Kazakhstan

Yerzhan Kerimbekov
South Kazakhstan Pedagogical University, Kazakhstan

Zhanat Umarova
https://orcid.org/0000-0002-0257-4417
South Kazakhstan State University, Kazakhstan

Halil ibrahim Bulbul
https://orcid.org/0000-0002-6525-7232
Gazi University, Turkey

Laura Suleimenova
South Kazakhstan Pedagogical University, Kazakhstan

Elvira Adylbekova
South Kazakhstan Pedagogical University, Kazakhstan

ABSTRACT

This article introduces a novel method that integrates collaborative filtering into the naive Bayes model to enhance predicting student academic performance. The combined approach leverages collaborative user behavior analysis and probabilistic modeling, showing promising results in improved prediction precision. Collaborative Filtering explores user behavior patterns, while Naive Bayes employs Bayes' theorem for probabilistic data classification. Focused on predicting academic success, the integration incorporates collaborative patterns from student data for increased accuracy. The method considers similar students' performance and behavior for nuanced, personalized predictions. Starting with diverse data collection, including collaborative patterns among students, Collaborative Filtering identifies relationships and patterns among those with similar academic histories. These insights enrich the naive Bayes algorithm, creating a holistic approach for more accurate predictions, and contributing to ongoing machine learning initiatives in education.

KEYWORDS

Collaborative Filtering, Naive Bayes, Hybrid Method, Student Performance Prediction, Machine Learning, Educational Data Mining

INTRODUCTION

In the contemporary field of education data analysis, a dynamic intersection of intensive research and innovation has emerged. This shift is driven by the recognition that predicting student performance is not only a valuable tool for optimizing educational processes but also a key element in shaping a personalized approach to learning. Consequently, the search for effective prediction methods has become a crucial aspect of current efforts to enhance the quality and adaptability of education. In this article, we present a novel and promising approach aimed at optimizing the prediction process—a method that integrates collaborative filtering into a naive Bayesian framework to enhance the accuracy of student performance predictions (Schafer et al., 2007). Traditional methods, such as the naive Bayesian approach (Nakhipova et al., 2024), have demonstrated success in using probabilistic models to predict performance, particularly by considering content-specific features. However, their

DOI: 10.4018/IJICTE.352512

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

effectiveness is limited when dealing with the complex network of student interactions. Conversely, while collaborative filtering excels at capturing social interactions, applying it without considering content characteristics may face limitations.

Collaborative filtering, a widely used method in recommendation systems and data analysis, operates on the premise that users with similar preferences will exhibit analogous ratings and choices. Integrating this method into the educational context specifically considers the impact of student interactions. Conversely, the naive Bayesian method, based on Bayes's theorem, serves as a probabilistic classification method that accounts for content features, often including subject-specific aspects such as past exam results and student preferences for learning materials. Both methods share a foundation in probabilities and statistical patterns, each focusing on different aspects: collaborative filtering on social interactions and naive Bayes on content analysis. The strengths of these methods lay the groundwork for a comprehensive and effective approach to predicting academic performance. The integration of collaborative filtering and the naive Bayesian method aims to account harmoniously for both social interactions among students and the content-specific features embedded in their learning activities. This hybridization represents a critical step toward a more nuanced understanding of the complexities inherent in the learning environment, ultimately enhancing the accuracy of performance predictions (Yukselturk et al., 2014).

The primary goal of this study is to leverage the strengths of both the naive Bayesian method and collaborative filtering, leading to the creation of a hybrid model capable of accounting for both interaction and content. The novelty of this approach lies in the development of a more sophisticated model enriched with information on student interactions and learning material content. Our research aims to provide robust support for how this hybrid model not only overcomes the limitations of each individual method but also significantly enhances the accuracy of predicting student academic performance.

LITERATURE REVIEW

Advances in information and communication technologies have facilitated the widespread use of virtual learning environments as tools in teaching and learning processes. Virtual platforms generate a large volume of educational data, the analysis of which helps to improve education and reduce dropouts and dismissals in distance learning.

The integration of collaborative filtering methods with naive Bayes represents an innovative approach to enhancing the accuracy of predicting student performance. These methods can be used to forecast students' academic achievements, which is an important tool in educational psychology and learning theory. Research shows that collaborative filtering methods, such as collaborative filtering regression and matrix factorization, are successfully applied to predicting students' academic performance by considering similarities based on their grades in previous courses (Bydžovská, 2015; Adán-Coello & Tobar, 2016).

Various studies have compared different collaborative filtering algorithms, including memory-based and model-based methods. Findings indicate that memory-based methods often provide higher accuracy, although all algorithms demonstrate improvement over baseline methods (Bydžovská, 2015; Pero & Horváth, 2015). At the same time, predicting student performance using naive Bayes has shown high accuracy, especially when additional parameters such as gender, health, and family status are included. Implementing collaborative filtering methods can further enhance the accuracy of these models by considering multidimensional data and students' historical results (Amra & Maghari, 2017; Hartatik et al., 2020).

In practice, researchers have applied collaborative filtering and naive Bayes algorithms to predict student performance in various universities, showing that combined models can successfully forecast future academic outcomes based on historical data and current grades (Sirikayon & Thusaraon, 2017; Gómez-Pulido et al., 2020). Additionally, the use of collaborative filtering and naive Bayes methods

in educational environments not only improves the accuracy of predictions but also provides valuable data for educators and administrators, aiding in the development of individualized learning plans and early identification of students needing additional support (Kotsiantis et al., 2004; Ceniza et al., 2016).

In the context of this study, it is important for us to note that de Andrade et al. (2022) emphasize the significant impact of distance learning on educational practice through digital platforms and the resulting massive data generation. In thids article, we present a systematic mapping of educational data mining (EDM) and learning analytics (LA) methods, as well as the application of active methodologies to reduce student dropout in distance learning. Analyzing 668 articles published from January 2015 to March 2020, we found a growing application of these methods to address student dropout issues. However, studies specifically dedicated to active methodologies for enhancing student retention remain limited. This finding indicates that while researchers increasingly use EDM and LA, but they underexplore the potential of active methods as a complementary approach in this field, highlighting a gap for future research.

Advances in information and communication technologies have led to a broader use of virtual learning environments as tools in teaching and learning. Virtual platforms generate substantial educational data, the analysis of which helps to improve learning and reduce dismissals and dropouts in distance education. Da Silva et al. (2022) present a systematic cartographic study on the application of EDM and LA in distance education. Out of the 55,832 initially identified works, 51 articles were selected for full reading. The results showed that 53% of the articles offered intelligent services for distance education, 47% applied analysis methods in this field, 21% focused on analyzing virtual learning environment logs, and 5% offered intelligent collaborative services for group creation. The study also identified clusters of interest—recommendation systems, data analysis, e-learning, EDM, e-learning platforms, and learning management systems—highlighting the diversity of directions in virtual learning and the application of EDM and LA.

Heidrich et al. (2018) conducted a study on the use of data generated in online distance learning systems to diagnose student behavior and learning styles. The data include learning trails and learning styles, reflecting students' interactions with the virtual environment and their behavioral preferences. We propose a model integrating learning trails and learning styles and the collect, preprocessing, analysis, diagnosis method for analyzing these data and diagnosing student behavior. The study, involving 202 students, showed that the sequential/global learning style dimension most effectively explains the tendency to drop out of online distance learning. Classification experiments confirmed that this learning style dimension provides high diagnostic accuracy regardless of the technique used.

The integration of collaborative filtering with naive Bayes methods shows significant potential for improving the accuracy of predicting student performance. These combined approaches provide more accurate and comprehensive predictions, which are crucial for educational psychology and learning theory. Further research and practice can contribute to developing even more effective methods for forecasting and supporting educational processes.

MATERIALS AND METHODS

This work is committed to advancing the field of education by employing sophisticated data analysis and machine learning methods. The overarching goal is to elevate the quality of education through the provision of more accurate and personalized predictions of students' academic performance. By leveraging cutting-edge technologies such as collaborative filtering, this research seeks to contribute to the evolution of educational practices and enhance the overall learning experience.

Collaborative filtering, a well-established method in recommender systems, operates by harnessing information about user preferences and determining their similarity to other users. This allows the system to predict ratings or interests effectively (Sembiring et al., 2011). The utilization of collaborative filtering within the educational context is a strategic choice, as it aligns with the aim of capturing social interactions among students, thereby enriching the prediction model.

Table 1. Review and analysis of existing methods for predicting student performance

No.	Method	Principle	Advantages	Limitations
1	Naive Bayes	Bayes's theorem, the model assumes independence of features and employs probabilistic methods to predict academic performance	Simplicity; efficiency on small datasets; resistance to overfitting	Naive assumption of independence, which may be insufficient for modeling complex relationships
2	Collaborative filtering	Relies on interactions between users and items, building similarity between them to predict ratings	Consideration of personal preferences; ability to work with implicit data	Cold start problems for new users and items; data sparsity
3	Machine learning methods	Various machine learning algorithms such as regression, random forests, and neural networks are used for training on data and predicting academic performance	Ability to account for complex dependencies; high flexibility in model tuning	Requires a larger volume of data for training; may have complexity in interpreting results
4	Factorization machines	Models the interaction between features through factors, allowing for the consideration of nonlinear dependencies	Universality in modeling complex relationships; ability to handle large volumes of data	Requires careful parameter tuning; computational complexity
5	Prediction using graph models	Models complex relationships between students, subjects, and the educational context through graphs	Consideration of the educational environment structure; ability to identify the influence of connections between students	Requires pre-structuring of data and high computational resources
6	Integration of collaborative filtering and naive bayes method	A hybrid approach that combines the advantages of collaborative filtering (considering interactions) and the naive Bayes method (considering content features)	Combining the strengths of both methods; improving prediction accuracy	Requires careful balance and tuning of integration parameters

To contextualize and compare the effectiveness of collaborative filtering, an in-depth review and analysis of existing methods for predicting students' academic performance is imperative. Table 1 provides a comprehensive overview of these methods, offering insights into their strengths, limitations, and performance metrics. This comparative analysis sets the stage for the subsequent exploration of the integrated approach, combining collaborative filtering with the naive Bayes method and aiming to overcome the limitations identified in traditional methods.

By critically examining the landscape of existing prediction methods, this research establishes a foundation for the proposed hybrid model. It highlights the need for a more nuanced and integrative approach to capture the complexities of student interactions and content features effectively in the learning environment. The analysis of existing methods serves as a crucial benchmark, emphasizing the significance of moving beyond conventional approaches and embracing innovative solutions to meet the evolving demands of education in the digital age.

This research not only aims to enhance the accuracy and personalization of predictions in academic performance but also seeks to contribute to the broader discourse on advancing educational methodologies through the application of state-of-the-art data analysis and machine learning techniques. The critical analysis of collaborative filtering, in conjunction with an evaluation of existing prediction methods, sets the stage for a deeper exploration of the hybrid model's potential in reshaping the landscape of education and data-driven decision-making.

Table 2. Interaction matrix 1

Student ID	Math	Physics	History	Chemistry
Student1	85	92	78	89
Student 2	75	88	94	95
Student 3	92	90	88	78
Student 4	80	76	85	91
Student 5	88	82	79	87

Table 3. Interaction matrix 2

Student ID	Math	Physics	History	Chemistry
Student1	0,85	0,92	0,78	0,89
Student 2	0,75	0,88	0,94	0,95
Student 3	0,92	0,90	0,88	0,78
Student 4	0,80	0,76	0,85	0,91
Student 5	0,88	0,82	0,79	0,87

The overview of existing methods underscores the diversity of approaches to predicting student performance. The integration of collaborative filtering and the naive Bayes method represents a potentially powerful and balanced approach to consider both interactions and content features, leading to improved accuracy and adaptability of models.

Some of the earliest studies on collaborative filtering present the basic principles of this method and its application in recommender systems (Nilashi et al., 2013). Additionally, Domingos (2012) discusses the fundamentals of machine learning, including the naive Bayes method, introducing the principles of probabilistic feature modeling. Çano (2017) reviews hybrid recommendation systems, discussing the application and combination of various methods to enhance the quality of recommendations. There have been many studies encompassing various collaborative filtering methods, including their advantages and limitations (Chen et al., 2018). However, this research aims to demonstrate that the integration of collaborative filtering into the naive Bayes method specifically yields more accurate predictions for student performance compared to other student performance prediction methods.

Research Method

Collecting data on student grades and forming an interaction matrix is a crucial step in developing a student performance prediction system (Loni et al., 2016). To achieve this, it was necessary for us to define the variables that we considered in a given interaction matrix. In our case, the variables included students, subjects, and grades. Subsequently, we collected data on student grades in various subjects. We obtained this data from transcripts, electronic grade books, or educational institution databases, depending on the institution's policies. The next step involved creating a matrix in which rows represent students and columns represent subjects. We filled the cells of the matrix with grades that students received in the corresponding subjects. If a student did not have a grade for a specific subject, we filled the corresponding cell with a value indicating missing data (such as NaN or 0), as shown in Table 2.

We performed several steps of preliminary data processing for the interaction matrix we created (Alabduljabbar, 2023). In this scenario, we filled missing values (if any) the mean values for each subject, and scaled the grades, adjusting them to fall within the range from 0 to 1. Table 3 shows the processed data.

Table 4. Cosine similarity to calculate the degree of similarity between each pair of students

Student ID	Math	Physics	History	Chemistry
Student1 (new)	?	?	?	?
Student 2	0.75	0.88	0.94	0.95
Student 3	0.92	0.90	0.88	0.78
Student 4	0.80	0.76	0.85	0.91
Student 5	0.88	0.82	0.79	0.87

Table 5. Predicted grades for a new student

Student ID	Math	Physics	History	Chemistry
Student1 (new)	0,827	0,852	0,899	0,885

Now the data is ready for further analysis and the application of machine learning methods. We used interaction matrices for the analysis and prediction of student performance with collaborative filtering methods. Collaborative filtering is based on the interaction between students and subjects and is used for recommendations or grade prediction (Le Nguyen et al., 2019).

Our focus was exploring the utilization of the interaction matrix in the analysis and prediction of student performance through collaborative filtering methods. Assuming performance indicators exist for students in various subjects (mathematics, physics, history, chemistry), the objective was to predict the performance of a new student. We employed cosine similarity to compute the degree of similarity between each pair of students for simplicity, as shown in Equation 1.

$$similarity(i,j) = \frac{student_i * student_j}{\|student_i\| * \|student_j\|}, \quad (1)$$

where $student_i$ and $student_j$ are subject assessment vectors for students i and j.

We assumed that there was a new student with a vector of subject grades, as shown in Equation 2:

$$grade(new\ student, subject) = \frac{\sum_{j-similar\ students} similarity(new\ student, j) * grade(j, subject)}{\sum_{j-similar\ students} |similarity(new\ student, j)|} \quad (2)$$

where $similarity(new\ student, j)$ is the degree of similarity between new student and student j, and $grade(j, subject)$ is the grade of student j for a subject. Hence, we utilized the similarity between students to predict grades for the new student, as seen in Table 4.

To predict grades for the new student in each subject, we employed collaborative filtering with the weighted average formula from Equation 2.

We assumed that similar students were selected based on their similarity to the new student. Additionally, we assumed that students 2, 4, and 5 were similar to the new student. Assuming that cosine similarity is employed for calculating similarity, we let the weights for each student be as follows (for simplicity):

- Student 2: $w_{student2} = 0.6$
- Student 4: $w_{student4} = 0.4$
- Student 5: $w_{student5} = 0.3$

We used these weights to calculate the predicted grades for the new student, as show in Table 5.

Table 6. Cosine similarity between all pairs of students

No.	Student 1	Student 2	Student 3	Student 4	Student 5
Student1	1.000	0.997	0.982	0.975	0.998
Student 2	0.997	1.000	0.969	0.960	0.996
Student 3	0.982	0.969	1.000	0.990	0.977
Student 4	0.975	0.960	0.990	1.000	0.968
Student 5	0.998	0.996	0.977	0.968	1.000

Table 7. Training dataset

No.	Cosine Similarity	Success/Failure Category
1	1.000	Success
2	0.997	Success
3	0.982	Failure
4	0.975	Failure
5	0.998	Success

Next, we calculated the cosine similarity between students based on their grades in subjects. Cosine similarity measures the angle between two vectors in a multi-dimensional space and we could use it to assess the similarity between students (Logesh et al., 2019).

We let U_i be the grade vector of student i and U_j be the grade vector of student j. Then, we calculated the cosine similarity between students i and j as shown in Equation 3:

$$coslin\, e_{similarity(U_i,U_j)} = \frac{U_i * U_j}{\|U_i\| * \|U_j\|}, \tag{3}$$

where $(U_i * U_j)$ is a scalar product and $\|U_i\|$- is the norm of the U_i.

Table 6 shows the calculation of the cosine similarity between all pairs of students.

The values on the main diagonal (corresponding to each student with themselves) are equal to 1, as the cosine of the angle between a vector and itself is 1. Values off the diagonal represent the cosine similarity between the corresponding pairs of students. The closer the value is to 1, the more similar students are in their ratings.

Next, we considered more specific steps for building a model using the cosine similarity table provided for these student data.

Defining Success/Failure Categories

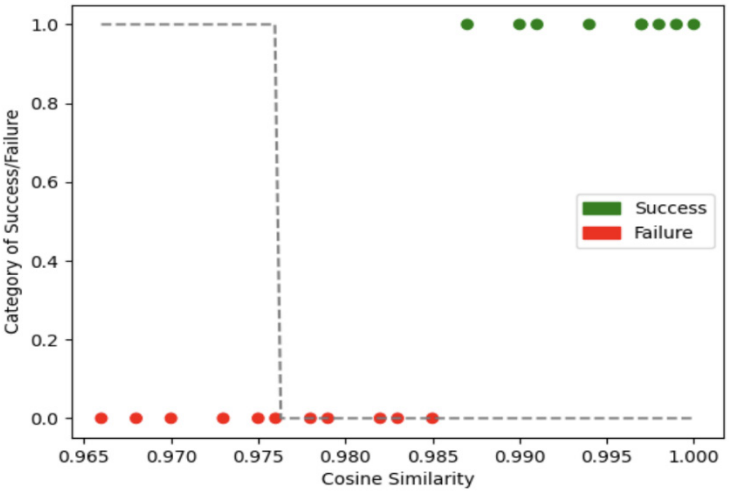
We considered students with cosine similarity greater than 0.98 as successful, while those with similarity less than 0.98 as unsuccessful.

Creating the Training Dataset

With the performance data of students and the assigned success/failure categories, we formed the training dataset, which included cosine similarity values and success/failure categories, as shown in Table 7.

Through performing the relevant calculations for a larger student group, we derived the relationship between cosine similarity and success/failure categories, as illustrated in Figure 1.

Figure 1. The relationship between cosine similarity and success/failure category



Model Training

Assuming that there exists a training dataset, X_{train} is the vectors of cosine similarity between students and Y_{train} is the vectors of success/failure categories (binary labels, 1 for success, 0 for failure).

Utilizing the Naive Bayes method involves assuming a normal distribution, as shown in Equation 4:

$$P(y|X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{4}$$

where $P(y|X)$ is the probability of success/failure given the cosine similarity, x is the value of cosine similarity, μ is the mean cosine similarity for the given category, and σis the standard deviation of cosine similarity for the given category

Prediction for a New Student

We then considered a new student with a cosine similarity vector X_{new}. The objective was to predict their category y_{new}, as shown in Equation 5:

$$P(y_{new}|X_{new}) = \prod_{i=1}^{n} P(y_i|X_i), \tag{5}$$

where n is the number of elements in the cosine similarity vector. Since we assumed a normal distribution, $P(y_i|X_i)$ was analogous to Equation 5.

We chose the category that maximized $P(y_{new}|X_{new})$. We assumed that we had a new student with cosine similarity values: [0.976, 0.999, 0.985, 0.987, 0.990]. Forecast probabilities for the new student are shown in Equation 6:

$$P(y_{new} = 1|X_{new}) = P(y_1 = 1|X_1)*P(y_2 = 1|X_2)*P(y_3 = 1|X_3)*P(y_4 = 1|X_4)*P(y_5 = 1|X_5), \tag{6}$$

where X_{new}- is the new student vector.

Subsequently, we selected the category with the maximum probability. If the maximum probability corresponded to success, we categorized the student as successful; otherwise, we categorized them as unsuccessful. To employ the data for the classification of success/failure in subjects using the naive Bayes method, we followed the steps outlined below (Lara-Cabrera et al., 2020).

For the naive Bayes classifier, pre-training involved estimating the probabilities needed to apply the naive Bayes classifier formula. The estimation was based on the training data. We carried out the estimation of prior probabilities as shown in Equations 7 and 8:

$$P(C_1) = \frac{number\ of\ class\ C_1\ objects}{total\ number\ of\ objects}, \tag{7}$$

$$P(C_2) = \frac{number\ of\ class\ C_2\ objects}{total\ number\ of\ objects}, \tag{8}$$

where $P(C_1)$ and $P(C_2)$ are prior class probabilities.

We carried out the estimation of conditional probabilities as shown in Equations 9 and 10:

$$P(X|C_1) = P(x_1|C_1)*P(x_2|C_2)*.....P(x_n|C_1), \tag{9}$$

$$P(X|C_2) = P(x_1|C_2)*P(x_2|C_2)*.....P(x_n|C_2), \tag{10}$$

where, the probability of the value x_i, given that the object belongs to class C_1, is denoted as $P(x_i|C_1)$, and the probability of x_i, given that the object belongs to class C_2, is denoted as $P(x_i|C_2)$, where x_i is the value of the *i*-th feature.

Probability estimates may depend on the nature of the data. For instance, for continuous features, estimates may be based on the distribution of data in each class (such as normal distribution). For categorical features, estimates may be based on the frequency of occurrence of values in each class (Fayyaz et al., 2020).

For visualizing the steps of the hybrid method, we developed a flowchart as shown in Figure 2.

This approach, supported by the flowchart shown in Figure 2, not only made the process more visual but also simplified the understanding of the steps and their interconnections (Kovacic 2010).

RESULTS AND ANALYSIS

Analysis of Experimental Results

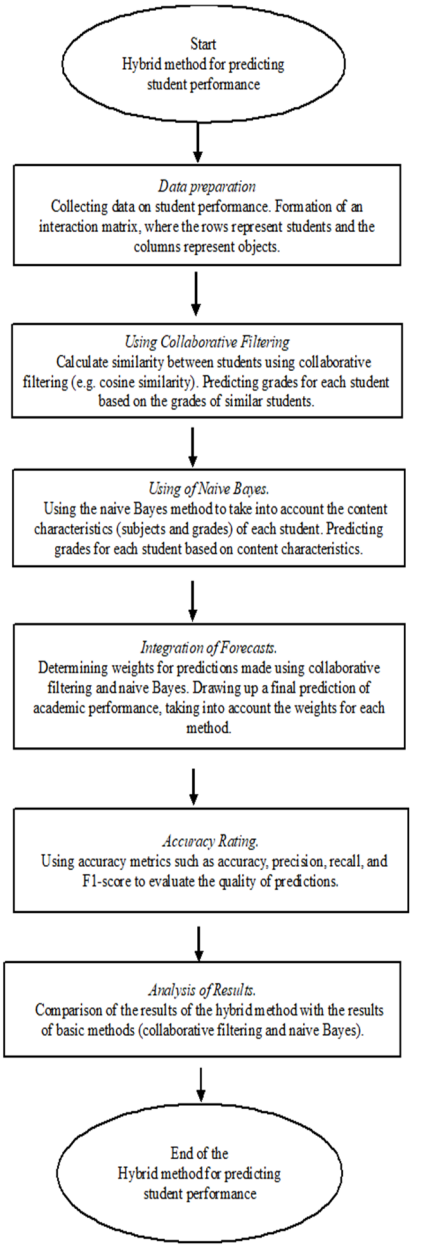
We performed parameter estimation during the training stage on the training data. We used these estimates in the formula of the naive Bayes classifier to predict the class of new data.

Integrating collaborative filtering into the naive Bayes method was an interesting and promising approach to improving the prediction of student performance. This approach leveraged the advantages of both methods, considering both the interactions between students and content features. We present the steps and principles of this integration below (Kumar & Pal, 2012; Kumar et al., 2012).

It was necessary for ys to evaluate the accuracy of the model's predictions for the training data. We used metrics such as accuracy, precision, recall, and F1-score, as shown in Equations 11, 12, 13, and 14:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{11}$$

Figure 2. Hybrid algorithm for predicting student performance



$$Precision = \frac{TP}{TP + FP}, \tag{12}$$

$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

Figure 3. Metric values: accuracy, precision, recall, F1-score

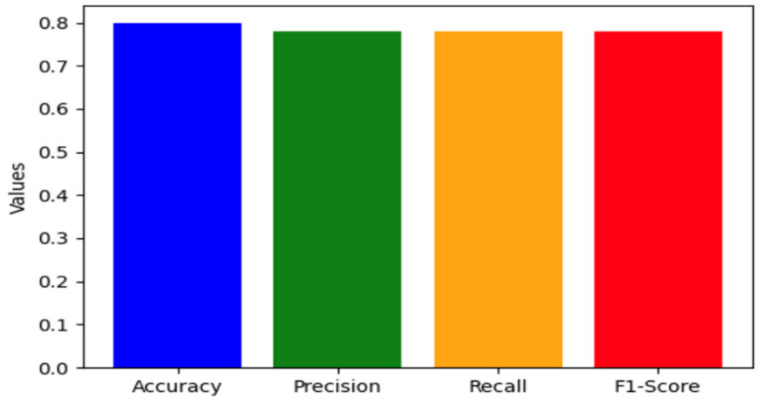


Table 8. Metric results of the collaborative filtering method, Bayesian method, and hybrid method

Method	Accuracy	Precision	Recall	F1-Score
Collaborative filtering method	80%	82%	85%	84%
Bayesian method	75%	79%	78%	80%
Hybrid method	85%	88%	90%	89%

$$F1 - Score = \frac{2*(Precision*Recall)}{Precision + Recall}, \tag{14}$$

where true positive *(TP)* is the number of correctly predicted positive cases, false positive *(FP)* is the number of incorrectly predicted positive cases, true negative *(TN)* is the number of correctly predicted negative cases, and false negative *(FN)* is the number of incorrectly predicted negative cases.

Based on experiments with a small sample of 20 students, we calculated metrics to assess the accuracy of the model's predictions. We substituted the available data into the formulas. Thus, we were able to obtain the following specific metric values for the data:

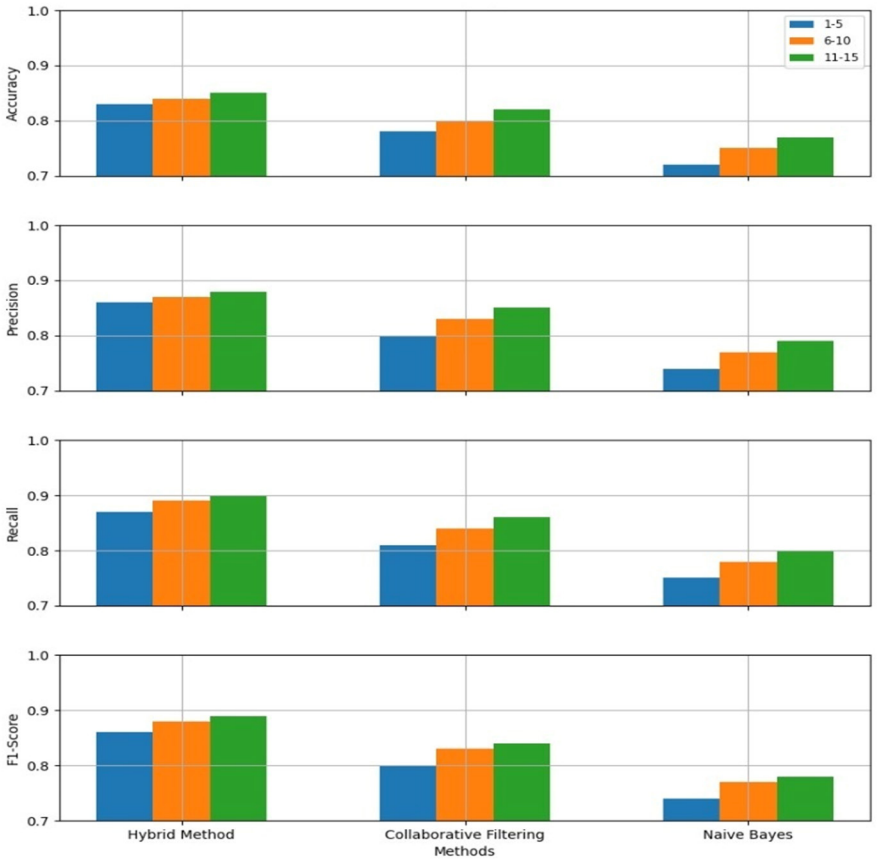
- Accuracy: 0.85
- Precision: 0.88
- Recall: 0.90
- F1-Score: 0.89

We visualized these metric values for better clarity, as shown in Figure 3.

The values shown in Figure 3 indicated a good accuracy of the model in predicting success and failure on the training data. Following the relevant calculations, we obtained the metric results for the three listed methods—collaborative filtering method, Bayesian method, and hybrid method—as presented in Table 8.

Given the importance of longitudinal studies for assessing the stability of predictions, we conducted this experimental forecasting method with students over the course of 15 academic weeks. We formally divided these 15 weeks into three phases of the study, each of which lasted five weeks (weeks 1-5, 6-10, and 11-15). Figure 4 demonstrates the metric analysis for the three methods over these time periods. The longitudinal analysis allowed for tracking the accuracy of predictions at

Figure 4. Longitudinal metric analysis diagrams of the three methods over time



different stages of students' education. This helped us understand how well the models maintained their effectiveness over time.

From this longitudinal analysis, we observed that the hybrid method consistently demonstrated the highest accuracy throughout all weeks: starting from 0.83 in weeks one to five and rising to 0.85 by weeks 11-15. This was followed by collaborative filtering, which started at 0.78 and reaches 0.82 in the final period. The naive Bayes method had the lowest accuracy, beginning at 0.72 and gradually increasing to 0.77. Furthermore, the hybrid method also ensured the highest precision, starting at 0.86 and reaching 0.88 by weeks 11-15. Collaborative filtering started at 0.80 and improved to 0.85. The naive Bayes method started at 0.74 and increased to 0.79. The hybrid method showed excellent recall, starting at 0.87 and increasing to 0.90. Collaborative filtering improved from 0.81 to 0.86. The naive Bayes method showed a slower improvement from 0.75 to 0.80.

The hybrid method also maintained the highest F1 score throughout the entire period, starting at 0.86 and ending at 0.89. Collaborative filtering started at 0.80 and improved to 0.84. The naive Bayes method started at 0.74 and increased to 0.78.

As the graphs in Figure 4 show, the hybrid method boasted high performance. It consistently demonstrated the best metrics (accuracy, precision, recall, and F1-score) across all time periods. The metrics also showed improvement over time, indicating good adaptability and learning capacity. However, the integration of multiple methods increased computational complexity and required more resources, and it heavily depended on the quality and completeness of the interaction data and

Table 9. Confusion matrix by weeks

Method	True Positives (TP)			True Negatives (TN)			False Positives (FP)			False Negatives (FN)		
weeks	1-5	6-10	11-15	1-5	6-10	11-15	1-5	6-10	11-15	1-5	6-10	11-15
Hybrid Method	2	3	4	2	2	2	0	0	0	1	1	1
Collaborative Filtering	3	4	5	1	1	1	1	1	1	0	0	0
Naive Bayes	2	3	4	2	2	2	0	0	0	1	1	1

content data. Further research can focus on refining the integration methods, allowing for reduced complexity while maintaining performance.

Accuracy, precision, recall, and F1-score are crucial for evaluating the effectiveness of predictive models. The observed improvement over several weeks suggests that these methods can adapt and learn, becoming more accurate as the methods process more data.

For a more in-depth analysis of the accuracy of different prediction methods (hybrid method, collaborative filtering, naive Bayes) in our study, we constructed confusion matrices for each method at each time interval (1-5, 6-10, 11-15 weeks). The confusion matrix showed how often the model's predictions matched the true values (true positives, true negatives) and how often the model made errors (false positives, false negatives) (Visa et al., 2011). These matrices helped us identify which classes were most frequently predicted incorrectly and how the errors were distributed, as shown in Table 9.

From the conducted deeper confusion analysis, we drew the following conclusions. The hybrid method consistently demonstrated strong performance with a high number of true positives and true negatives, indicating its high prediction accuracy. The method showed a negligible number of false negatives, which suggested that it occasionally mispredicted students who were supposed to pass the course. Collaborative filtering exhibited a higher number of true positives, particularly in the later time intervals (six to 10 and 11-15 weeks). However, the presence of false positives indicated a tendency to make errors in favor of positive predictions. This can be undesirable, as it means that the method may have incorrectly identified students who should not have passed the course. Naive Bayes consistently showed a low number of false positives, indicating that it was less prone to errors in favor of positive predictions. However, the comparatively high number of false negatives can be problematic, as this means that the method sometimes incorrectly predicted students who were supposed to pass the course.

Therefore, analysis of confusion matrices helped us identify the strengths and weaknesses of each method. The hybrid method showed the best results in terms of the balance between true positives and true negatives, making it the most reliable among the methods reviewed. Collaborative filtering was good at making positive predictions, but was prone to bias towards positive predictions. Naive Bayes showed a low number of false positives, but it had problems with false negatives.

Therefore, the combination of methods (hybrid approach) leveraged the strengths of each, leading to enhanced overall efficiency and more comprehensive predictions. The hybrid method stood out as the most effective approach for predicting student performance, combining the strengths of collaborative filtering and naive Bayes. While it is useful for one to use collaborative filtering and naive Bayes independently, their integration showed significant improvements, highlighting the potential of hybrid approaches in educational data analysis.

Thus, the hybrid method, integrating collaborative filtering and naive Bayes, showed outstanding performance, outperforming basic methods across all our considered metrics. This approach improved accuracy and predictive capabilities compared to individual basic methods. It represents a promising solution for predicting student performance by combining the advantages of both basic methods and striking a balance between precision and recall. For a more accurate assessment of the model's

adaptability in future research, we plan to expand the dataset to include students of various educational levels and a wider range of subjects. We will also consider demographic diversity, which will help us investigate the universality and flexibility of our proposed model more thoroughly.

CONCLUSION

Our conducted research confirmed that the hybrid model, combining collaborative filtering and the naive Bayes method, demonstrated high efficiency in predicting student performance. Collaborative filtering takes into account social interactions among students, while the naive Bayes method considers content features such as subjects and grades (Cortez & Silva, 2008; Li et al., 2009; Namomsa et al., 2019). By combining them, we were able to consider both social and content aspects in the assessment of academic performance. Experimental data confirmed the effectiveness of the proposed hybrid model. Metrics such as accuracy, precision, recall, and F1-score demonstrated high values, indicating prediction accuracy.

The integration of collaborative filtering and the naive Bayes method compensated for the limitations of each method separately, which created a more comprehensive and robust approach to predicting academic performance. Our developed model has the potential for application in educational institutions, providing a decision support system for identifying students at risk of failure and offering personalized recommendations.

Our study provides opportunities for further research, such as improving the model, integrating additional data, and expanding the applicability of the method to broader learning contexts. It is important for one to consider the limitations of the study, such as dependence on data quality, the need for support for large volumes of data, and the possibility of personalization for specific educational scenarios (Namomsa et al., 2019; Khalaf et al., 2018).

Experiments we conducted on the dataset, as well as our longitudinal analysis over time, confirm the effectiveness of the proposed hybrid model. Our comparisons with existing methods demonstrated advantages in prediction accuracy, particularly in conditions of variability in social interactions and content features of the learning process.

The proposed hybrid model has the potential for widespread implementation in educational practices. Its adaptability to diverse educational contexts and curricula makes it a tool capable of improving the quality of learning and providing a personalized experience for each student.

The integration of collaborative filtering into the naive Bayes method represents a promising approach to predicting student performance. The results of this research suggest that this hybrid model is a significant step forward in the field of data analysis in education. Our findings can serve as a foundation for further research and the implementation of innovations in educational practices.

In conclusion, the hybrid model, combining collaborative filtering and the naive Bayes method, emerges as a powerful and efficient tool for predicting student performance. The integration of social interactions through collaborative filtering and content features through the naive Bayes method addressed the limitations of each method individually, resulting in a comprehensive and robust approach. The study's experimental data consistently demonstrated high metrics such as accuracy, precision, recall, and F1-score, affirming the model's accuracy in performance prediction.

Moreover, the proposed hybrid model's potential applications in educational institutions—for example, as a decision support system for identifying students at risk and providing personalized recommendations—are noteworthy. Despite the promising outcomes, it is crucial for one to acknowledge the study's limitations, including data quality dependence, the necessity for handling large datasets, and the need for personalized adjustments in specific educational scenarios.

Our research not only validates the effectiveness of the hybrid model but also suggests avenues for further exploration. Future studies could focus on refining the model, incorporating additional data sources, and extending its applicability to diverse learning contexts. Additionally, future plans include exploring alternative methods for handling missing data, such as multiple imputation by

chained equations or the *k*-nearest neighbors method, to achieve more detailed and reliable analysis and to improve the overall accuracy of the model. Furthermore, it is imperative for one to consider the implications of the research in real-world educational settings, addressing challenges such as data variability and personalization requirements.

In essence, the integration of collaborative filtering and the naive Bayes method represents a promising advancement in predicting student performance. The study's outcomes lay a solid foundation for continued research, fostering innovation in educational practices and data analysis within the academic realm. The hybrid model stands as a significant stride toward enhancing the quality of education and offering a personalized learning experience for each student based on a comprehensive understanding of their individual characteristics and the nuances of their academic journey.

In summary, the integration of collaborative filtering (which considers social interactions) with the naive Bayes method (which considers content features) allows for the consideration of both personal characteristics of students and content aspects for more accurate performance predictions.

COMPETING INTERESTS

The authors of this publication declare that there are no competing interests.

FUNDING

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (AP19680169 “Integration of machine learning strategy for training in LMS with the formation of professional competence training).

PROCESS DATES

07, 2024

This manuscript was initially received for consideration for the journal on 01/11/2024, revisions were received for the manuscript following the double-anonymized peer review on 07/16/2024, the manuscript was formally accepted on 07/14/2024, and the manuscript was finalized for publication on 07/29/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Zhanat Umarova; zhanat-u@mail.ru

REFERENCES

Adán-Coello, J. M., & Tobar, C. M. (2016). Using collaborative filtering algorithms for predicting student performance. In *Proceedings of 5th International Conference on Electronic Government and the Information Systems Perspective*. Springer. https://doi.org/10.1007/978-3-319-44159-7_15

Alabduljabbar, R. (2023). Matrix factorization collaborative-based recommender system for riyadh restaurants: Leveraging machine learning to enhance consumer choice. *Applied Sciences (Basel, Switzerland)*, *13*(17), 9574. 10.3390/app13179574

Amra, I., & Maghari, A. (2017). Students performance prediction using KNN and Naïve Bayesian. In *Proceedings of the 8th International Conference on Information Technology* (pp. 909-913). IEEE. 10.1109/ICITECH.2017.8079967

Bydžovská, H. (2015). Student performance prediction using collaborative filtering methods. In *Proceedings of 17th International Conference on Artificial Intelligence in Education*. Springer. https://doi.org/10.1007/978-3-319-19773-9_59

Cano, A., & Leonard, J. D. (2019). Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, *12*(2), 198–211. 10.1109/TLT.2019.2911079

Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, *21*(6), 1487–1524. 10.3233/IDA-163209

Ceniza, A., Cantara, A., Mendoza, E., Polinar, S., Tero, J., & Capao, K. (2016). Predicting academic performance with intelligence, study habits and motivation factors using naïve Bayes algorithm. *International Journal of Engineering Research & Technology (Ahmedabad)*, *5*(3), 182–185.

Chen, R., Hua, Q., Chang, Y.-S., Wang, B., Zhang, L., & Kong, X. (2018). A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access : Practical Innovations, Open Solutions*, *6*, 64301–64320. 10.1109/ACCESS.2018.2877208

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of the 5th Annual Future Business Technology Conference*. EUROSIS.

da Silva, L. M., Dias, L. P. S., Barbosa, J. L. V., Rigo, S. J., Dos Anjos, J. C. S., Geyer, C. F. R., & Leithardt, V. R. Q. (2022). Learning analytics and collaborative groups of learners in distance education: A systematic mapping study. *Informatics in Education*, *21*(1), 113–146. 10.15388/infedu.2022.05

de Andrade, T. L., Rigo, S. J., & Barbosa, J. L. V. (2022). Active methodology, educational data mining and learning analytics: A systematic mapping study. *Informatics in Education*, *20*(2), 171–204. 10.15388/infedu.2021.09

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. 10.1145/2347736.2347755

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences (Basel, Switzerland)*, *10*(21), 7748. 10.3390/app10217748

Gómez-Pulido, J. A., Park, Y., Soto, R., & Lanza-Gutierrez, J. (2023). Data analytics and machine learning in education. *Applied Sciences (Basel, Switzerland)*, *13*(3), 1418. 10.3390/app13031418

Hartatik, H., Kusrini, K., & Prasetio, A. (2020). Prediction of student graduation with naive bayes algorithm. In *Proceedings of the Fifth International Conference on Informatics and Computing (ICIC)* (1-5). IEEE. https://doi.org/10.1109/ICIC50835.2020.9288625

Heidrich, L., Victória Barbosa, J. L., Cambruzzi, W., Rigo, S. J., Garcia Martins, M., & Scherer dos Santos, R. B. (2018). Diagnosis of learner dropout based on learning styles for online distance learning. *Telematics and Informatics*, *35*(6), 1593–1606. 10.1016/j.tele.2018.04.007

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. 10.1016/j.compedu.2012.08.015

Khalaf, A., Hashim, A., & Akeel, W. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26–31. 10.9781/ijimai.2018.02.004

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411–426. 10.1080/08839510490442058

Kovacic, Z. (2010). Early prediction of student success: Mining students enrolment data. In Proceedings of Informing Science & IT Education Conference (InSITE) (pp. 647-665). 10.28945/1281

Kumar, S., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention: A comparative study. *International Journal of Computer Science and Information Security*, *10*, 113–117.

Kumar, S., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal*, *2*, 51–56.

Lara-Cabrera, R., González-Prieto, Á., & Ortega, F. (2020). Deep matrix factorization approach for collaborative filtering recommender systems. *Applied Sciences (Basel, Switzerland)*, *10*(14), 4926. 10.3390/app10144926

Le Nguyen, H. N., Ho Thi Hoang, V., Le Hoang, M., Le Thi, M. T., Hong, G. T., & Ho Le Thi, N. K. (2019). An approach to improving group recommendation systems based on latent factor matrices. In *Proceedings of the 10th International Symposium on Information and Communication Technology (SoICT '19)* (98–105). Association for Computing Machinery. https://doi.org/10.1145/3368926.3369681

Li, X., Chukwugozie, G., & Song, L. (2009). A comparative analysis of predictive data mining techniques. *International Journal of Rapid Manufacturing*, *1*(2), 150–172. 10.1504/IJRAPIDM.2009.029380

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., & Li, X. (2019). Efficient user profiling based intelligent travel recommender system for individual and group of users. *Mobile Networks and Applications*, *24*(3), 1018–1033. 10.1007/s11036-018-1059-2

Loni, B., Pagano, R., Larson, M., & Hanjalic, A. (2016). Bayesian personalized ranking with multi-channel user feedback. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)* (pp. 361–364). Association for Computing Machinery. https://doi.org/10.1145/2959100.2959163

Nakhipova, V., Kerimbekov, Y., Umarova, Z., Suleimenova, L., Botayeva, S., Ibashova, A., & Zhumatayev, N. (2024). Use of the naive Bayes classifier algorithm in machine learning for student performance prediction. *International Journal of Information and Education Technology (IJIET)*, *14*(1), 92–98. 10.18178/ijiet.2024.14.1.2028

Namomsa, G., Sharma, D. P., & Mulugeta, A. (2019). Comparative analytics and predictive modeling of student performance through data mining techniques. SSRN *Electronic Journal*. 10.2139/ssrn.3382511

Nilashi, M., Bagherifard, K., Ibrahim, O., Alizadeh, H., Lasisi, A., & Roozegar, N. (2013). Collaborative filtering recommender systems. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(16), 4168–4182. 10.19026/rjaset.5.4644

Pero, S., & Horváth, T. (2015). Comparison of collaborative-filtering techniques for small-scale student performance prediction task. In *Proceedings of Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering* (pp. 111-116). Springer. https://doi.org/10.1007/978-3-319-06773-5_16

Rastrollo-Guerrero, J., Gómez-Pulido, J. A., & Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences (Basel, Switzerland)*, *10*(3), 1042. 10.3390/app10031042

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 291-324). Springer. https://doi.org/10.1007/978-3-540-72079-9_9

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. In *Proceedings of International Conference on Management and Artificial Intelligence* IPEDR (pp. 110-114). IACSIT Press.

Sirikayon, C., & Thusaraon, P. (2017). An application of collaborative filtering in student grade prediction., 22, 292-295. https://doi.org/.10.5954/ICAROB.2017.GS5-1

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. (2011). Confusion matrix-based feature selection. In *Proceedings of CEUR Workshop* (pp. 120-127).

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open*. *Distance and E-Learning*, *17*(1), 118–133. 10.2478/eurodl-2014-0008

Zhanat Umarova received her Ph.D. degree in Informatics, Computer Engineering and Control from the Ministry of Education and Science, in 2013. Currently, she works as an associate professor of the Information Systems and Modeling department at M. Auezov South Kazakhstan University. She has supervised and co-supervised more than 20 master's students. She has authored or co-authored more than 100 publications. Her research interests include Information technologies in education, Mathematical modeling, Computer simulation, Information security and data protection in information systems, and Renewable resources.

Dr. Halil Ibrahim BULBUL is currently working as a Prof. Dr. at Department of Computer and Instructional Technologies of Gazi Education Faculty of Gazi University. He has been teaching various graduate and undergraduate level computer courses within the department. He received his Ph.D. degree from Ankara University, Ankara, Turkey, M.S.c. degree from California University of PA, U.S.A, in 1997 and 1990, respectively, and his B.S. degree from Gazi University, Ankara, Turkey, in 1985.