International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 27, Number 3

August – 2026

Predicting Student Outcomes in Open High Schools Using Educational Data Mining

Ahmet Polat[1*] and Mehmet Barış Horzum[2]
*[1]Ministry of National Education, Sivas, Türkiye; [2]Department of Computer Education and Instructional Technology, Faculty of Education, Sakarya University, Adapazarı, Türkiye; *Corresponding Author*

Abstract

Open high schools fulfill a critical function by offering flexible educational pathways for students encountering diverse socioeconomic and personal challenges. Nevertheless, escalating dropout rates represent a significant concern, necessitating robust predictive models to facilitate early interventions. This study employed educational data mining (EDM) techniques to analyze the academic trajectories of 484,158 students enrolled in Turkish open high schools. We evaluated multiple classification algorithms—J48, decision tree, k-nearest neighbors (kNN), naïve Bayes, and random forest—across four distinct data preprocessing scenarios to predict student status (dropout/delayed graduation/graduation). The J48 algorithm demonstrated superior performance, achieving an accuracy of 80.47% and a kappa statistic of 0.61. Key findings reveal that academic and administrative features, notably total credit accumulation and initial enrollment type, are more important predictors than demographic variables. This research provides empirically grounded insights for the early identification of at-risk students. It offers data-driven recommendations to enhance student retention policies within open high school systems, contributing a large-scale, multi-class prediction analysis within a unique, under-researched national distance education context.

Keywords: educational data mining, dropout prediction, open high school, machine learning, distance education, student retention, early warning system

Introduction

Education is fundamental to human development, influencing individual opportunities and societal progress. The United Nations' (UN) 2030 Sustainable Development Goals reflect this priority, targeting universal primary and secondary access, vocational skills development, and equitable tertiary education (United Nations Department of Economic and Social Affairs [UNDESA], 2015), while the UN Human Development Index measures educational progress through average and expected years of schooling (United Nations Development Program [UNDP], 2021). As a UN member, Türkiye pursues these educational aims, yet faces persistent challenges in achieving these targets.

Equitable access to quality education remains a significant challenge in Türkiye (Ministry of National Education [MoNE], 2015). Although overall schooling rates vary across educational levels—68% in preschool, 91% in primary, 84% in secondary, and 44% in higher education (MoNE, 2019)—secondary education, in particular, exhibits notable regional disparities. For instance, the secondary school enrollment rate is 91% in western Anatolia, compared to 72% for males and 68% for females in southeastern Anatolia (MoNE, 2019). Furthermore, among adults aged 25–34, nearly half have not completed secondary education, with a significant gender gap in average years of schooling: 7.3 years for women versus 9 years for men (Organisation for Economic Co-operation and Development [OECD], 2018; UNDP, 2021).

Addressing educational challenges, Türkiye extended compulsory education to 8 years (Türkiye Cumhuriyeti Resmî Gazete [TC Resmî Gazete], 1997), then 12 years (TC Resmî Gazete, 2012). National development plans promoted open and distance education for equality and lifelong learning (TC Resmî Gazete, 2006, 2013). Subsequently, open high school (OHS) enrollment grew from 14% (940,268) of all secondary students in 2012 to 26% (1,554,938) in 2017 (MoNE, 2012, 2017). However, dropout rates increased dramatically; by 2017, OHS dropouts (N = 2,143,674) reached 137% of active enrollment (Polat, 2021), a persistence problem recognized in MoNE strategic planning (2015).

Secondary education dropout impacts national development through diminished human capital (Rosenzweig, 2010; Rumberger, 2020), reduced productivity (Aydın et al., 2012; Çalışkan et al., 2013), and increased social costs in crime, health, and life expectancy (Ereş, 2005; Kızmaz, 2004; Luy et al., 2019). Young adults without high school diplomas face higher unemployment rates (OECD, 2021). Parental education also significantly influences academic outcomes across generations (MoNE, 2022). Student dropout within Turkish OHSs represents an escalating problem affecting a student demographic facing socioeconomic and personal barriers to traditional schooling and reducing OHS dropout rates is thus crucial for both individual success and the country's socioeconomic advancement.

Open High Schools in Türkiye

Türkiye's first OHS (in Turkish, açık öğretim lisesi or AÖL) was established in 1992. Subsequently, the vocational open high school (in Turkish, mesleki açık öğretim lisesi or MAÖL) was established in 2006, followed by the open education imam hatip high school (açık öğretim imam hatip lisesi or AÖİHL) in 2016. While these institutions were initially part of AÖL, they later became separate entities. As of the 2020/2021 academic year, the total number of actively enrolled students across these OHSs reached 1,452,331 (MoNE, 2021).

Despite the size of the student population, research on the OHS dropout problem remains limited. Early studies focused on operational aspects, learning environments, materials, and support services

(Demiray & Sağlık, 2003). Later research examined reasons for transferring to open schools (Adıgüzel, 2016; Çuhadar Öncü, 2017; Şahin, 2017), student-perceived obstacles (Çiçek, 2005; Şahin, 2017; Sipahi, 2019; Soylu, 2014), academic achievement (Özkahveci, 2001), service effectiveness (Bedel, 2006; Sarıhan, 2010; Şentürk, 2009; Yavuz, 2014), student perspectives (Demirtaş et al., 2017; Dere, 2002), and individual differences (Randler et al., 2014).

However, much of this existing research is descriptive and localized, relies on small sample sizes (often focusing on MAÖL students due to face-to-face components), and lacks representativeness concerning the vast and diverse population enrolled across Türkiye. Consequently, there is limited understanding of the factors predicting persistence or dropout on a national scale, hindering the development of evidence-based interventions. While engaging large, representative samples present logistical challenges (labor, time, cost), the increasing availability of rich administrative data within student information systems offers a valuable opportunity. Educational data mining (EDM) techniques can leverage this data to uncover complex patterns and predictors of student outcomes that might otherwise remain obscured.

Educational Data Mining

Information technology advancements enable educational institutions to collect comprehensive student data through sophisticated management systems, creating valuable analytical opportunities. Data mining, extracting implicit and potentially useful information from large datasets (Han et al., 2011), provides tools for addressing complex educational challenges. EDM applies these methods to large-scale educational data, illuminating student behaviors and learning environments (International Educational Data Mining Society, n.d.). EDM applications primarily focus on predicting academic performance, identifying at-risk students, and understanding dropout factors (Romero & Ventura, 2010).

EDM holds promise for anticipating educational needs, personalizing learning experiences, and ultimately improving student outcomes. Prediction studies, a core component of EDM, assist educators in identifying critical success factors and their interrelationships, thereby highlighting areas for targeted intervention and systemic improvement. Within the Turkish context, the concurrent rise in demand for OHSs and the alarming increase in dropout rates create a compelling case for applying EDM. The existing research gap concerning dropout predictors in these specific institutions necessitates the use of data mining methods to uncover potentially complex patterns associated with success, delayed graduation, and dropout, with the ultimate goal of enhancing completion rates.

Identifying students at risk of dropout is a crucial prerequisite for implementing timely and effective interventions (Dekker et al., 2009; Heppen & Therriault, 2008; Márquez-Vera et al., 2016; Tanner, 2003). Dropout is a multifaceted phenomenon influenced by individual, academic, family, school, and community factors (Rumberger & Lim, 2008). However, collecting comprehensive data, particularly psychosocial or detailed background information, via traditional methods such as surveys is often impractical in large-scale distance education settings like Turkish OHSs. Therefore, leveraging readily available administrative data from student information systems presents a pragmatic and efficient approach to investigating dropout predictors, conserving resources while enabling analysis at scale.

A review of the EDM literature focused on predicting student performance and dropout reveals a predominant focus on higher education settings (Berens et al., 2019; Delen, 2010; Djulovic & Li, 2013; Kovacic, 2010; Lassibille & Navarro-Gomez, 2008). Studies examining dropout at the high school level are comparatively scarce (Lee & Chung, 2019; Márquez-Vera et al., 2013; Şara et al., 2015). Existing research spans diverse national contexts, educational systems (face-to-face, distance, e-learning), and data sources. While common predictors such as academic achievement (GPA, prior grades, exam scores), personal characteristics (gender, age, disability), and sometimes psychometric traits (motivation) emerge, the heterogeneity in study contexts, methodologies, sample sizes, and data types limits the generalizability of findings. This variability underscores the necessity for context-specific research.

This study is conceptually grounded in established models of student departure, primarily drawing upon Tinto's (1975) model of student integration, and Bean and Metzner's (1985) model tailored for non-traditional students. Tinto's framework emphasizes the critical roles of academic and social integration in fostering persistence. Bean and Metzner's model, particularly relevant to the OHS population, which often includes non-traditional learners, additionally highlights the influence of external environmental factors (e.g., employment, family responsibilities) alongside academic variables and background characteristics. This theoretical lens guides our variable selection, positing that a combination of factors—including academic progress (e.g., credits earned), background characteristics (e.g., enrollment type, age), and potentially external factors reflected in administrative data (e.g., employment status)—interact to shape student persistence within the unique context of OHSs.

Given research gaps—the lack of national-scale studies in the Turkish OHS context, the limited focus on the dropout problem in open secondary education, and the need for nuanced outcome prediction—an EDM study investigating dropout in this specific setting is warranted. Such research is essential not only for addressing existing knowledge gaps but also for guiding future research directions and informing the development of targeted, evidence-based policies and interventions aimed at improving student retention. This study directly addressed these gaps by applying predictive modeling to a comprehensive, nationwide dataset of Turkish OHS students, employing a multi-class outcome variable for a more nuanced and realistic representation of learner trajectories, and identifying the most important predictors specific to this unique distance secondary education environment.

Research Questions

This study aimed to predict the status (graduation, active/delayed graduation, or dropout) of students enrolled in Turkish OHSs at the end of their standard 4-year education period (defined as the official program duration starting from the student's initial enrollment date), using EDM methods applied to data available within the institutional student information system. In pursuit of this aim, the study posed the following research questions:

> RQ1: What is the classification performance of various machine learning models developed to predict the graduation, dropout, and active/delayed status of OHS students?
>
> RQ2: Which student features available in the administrative dataset are the most important predictors of these outcomes?

Methodology

Employing EDM techniques to analyze educational datasets, this study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) model (Chapman et al., 2000), a structured framework guiding the analysis from business understanding through deployment (Figure 1). This systematic methodology facilitated a rigorous approach to the data analysis process.

Data Analysis

Following the CRISP-DM framework, all data processing and modeling were carried out using RapidMiner Studio VM software (Version 9.0), a comprehensive data science platform facilitating visual workflow design and automation across the data science lifecycle (Altair Engineering, 2026).

Business Understanding

The core objective was predicting student status after the standard 4-year timeframe, with three target categories:

- graduated: students who completed all requirements and received their diploma within or by the end of the 4-year timeframe.
- active/delayed graduation: students who remained enrolled beyond the standard 4 years, indicating delayed progress towards graduation.
- dropout: students who discontinued their studies and were no longer enrolled, without having graduated.

Figure 1

CRISP-DM Model

Note. From *CRISP-DM Process Diagram,* by K. Jensen, 2012, Wikimedia Commons (https://commons.wikimedia.org/w/index.php?curid=24930610). CC BY-SA 3.0.

Data Understanding

The dataset was obtained with official permission from the MoNE Information Processing Department (Official Letter No: 65968543-605.01-E.86777), adhering to MoNE regulations. It comprised anonymized records for students who first enrolled in one of the three types of OHSs (AÖL, MAÖL, or AÖİHL) during the 2013 academic year. This year was chosen as it marked the implementation of 12-year compulsory education and saw a significant increase in OHS enrollment. The final analysis dataset, after handling exclusions for specific non-representative cases (e.g., deceased, returned to formal education), consisted of 484,158 student records.

The raw data contained various administrative, demographic, and academic features. Exploratory analysis using RapidMiner identified the potentially relevant attributes listed in Table 1.

Table 1

Features Initially Available in the Dataset

Category	Feature	Description
Administrative	School type	Type of OHS (AÖL/MAÖL/AÖİHL)
	Enrollment type	Educational status at enrollment (e.g., primary/secondary graduate, transfer from traditional HS)
	Field/department	Registered field or department (if applicable)
Demographic	Date of birth	Student's date of birth
	Gender	Student's gender
	Address city	City of residence
	Address district	District of residence
	Employment status	Employment status (employed/unemployed, type of employment)
	Disability status	Presence and type of disability
	Special status	Special circumstances (e.g., detained, child of a martyr/veteran, under court order)
	Military service	Military service status (relevant primarily for males)
Academic	Status	Target variable: Student's status (graduated, active, dropout)
	Previous system	The class transition system of the school student transferred from
	Previous school	Type of school transferred from (if applicable)
	Total credit count	Total course credits earned by the student

Note. OHS = open high school; AÖL = açık öğretim lisesi [open high school]; MAÖL = mesleki açık öğretim lisesi [vocational open high school]; AÖİHL = açık öğretim imam hatip lisesi [imam hatip open high school].

The majority of students (80%) were enrolled in AÖL, followed by MAÖL (13%) and AÖİHL (7%). Gender distribution varied by school type: we found approximately 60% were male in the AÖL, 62% were male in the MAÖL, whereas 63% were female in the AÖİHL.

Data Preparation

Several steps were taken to prepare the data for modeling:

- An age feature at the time of enrollment (calculated from date of birth) was added, aligning with theoretical propositions regarding the impact of age on student persistence.

- A region feature was derived using the European Union's nomenclature of territorial units for statistics (NUTS) level 1 classification (Eurostat, 2026) based on the address city to capture regional socioeconomic variations that might influence educational outcomes.

- The status feature was designated as the target variable. The distribution across the three classes was: graduated 11.1%, active/delayed graduation 26.2%, and dropout 62.7%.

The dataset exhibited significant class imbalance. We deliberately avoided resampling techniques to preserve the authentic data distribution reflecting the real-world phenomenon. Our strategy focused on using algorithms relatively robust to imbalance and employing evaluation metrics (precision, recall, F-measure, kappa) less sensitive to imbalance than simple accuracy (He & Garcia, 2009).

To systematically evaluate data representation impact on model performance, we implemented four progressive preprocessing stages: stage 1: minimal transformation; stage 2: removal of high-missingness features; stage 3: simplification of nominal features; and stage 4: discretization of continuous variables.

Modeling

Five classification algorithms were selected for this study: decision tree, J48, k-nearest neighbors (kNN), naïve Bayes, and random forest. These algorithms were chosen because they represent different methodological approaches to classification and have demonstrated effectiveness in EDM research (Kılınç, 2015; Shahiri & Husain, 2015; Umar, 2016; Yukselturk et al., 2014; Yurdakul, 2015). Testing multiple algorithms allowed us to identify which approach performed best for predicting student outcomes in the Turkish OHS context.

The selected algorithms operate using the following distinct methodologies:

- decision tree: constructs a flowchart-like model where branches represent decision nodes derived from specific student characteristics, such as credit accumulation.

- J48: an advanced decision tree that improves prediction by automatically pruning (trimming) unnecessary branches to prevent overfitting, when a model learns training data too specifically and performs poorly on new data. J48 works by examining student features, selecting the feature that best separates students into outcome groups, creating branches, and pruning branches that don't improve accuracy.

- random forest: generates an ensemble of decision trees and aggregates their predictions through a voting process to increase accuracy.

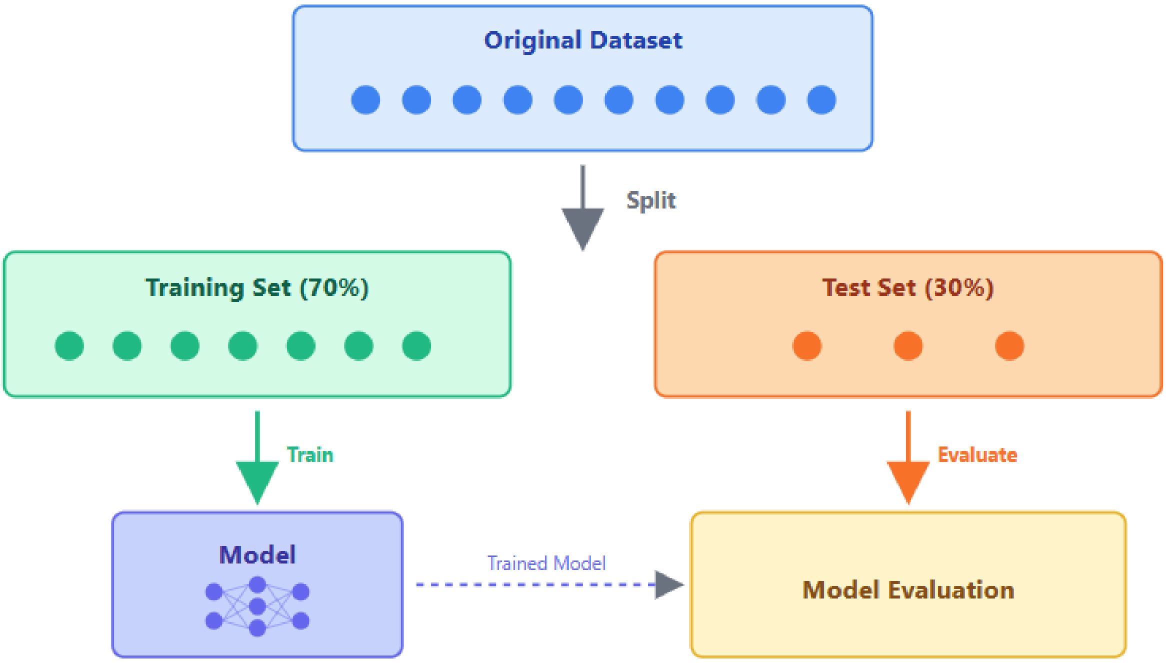
- k-nearest neighbors (kNN): uses instance-based learning, classifying students by identifying the prevailing outcome among similar data points.

- naïve Bayes: applies probabilistic principles to predict outcomes based on attribute likelihoods and the assumption of feature independence.

For model training and testing, the holdout method (70% training, 30% testing), depicted in Figure 2, was employed with stratified sampling, which builds random subsets to ensure the class distribution is the same as in the whole dataset (Altair Engineering, 2026).

Figure 2

Holdout Method Used for Model Training and Testing



All algorithms were run using their default parameter settings (without fine-tuning) within RapidMiner Studio (Altair Engineering, 2026). This approach was chosen to establish a baseline comparison of algorithm performance under standard conditions across different preprocessing stages (Romero & Ventura, 2020), rather than optimizing each algorithm individually, which was considered beyond the scope of this initial investigation but represents an avenue for future refinement (Witten et al., 2016).

Evaluation

Model performance was assessed using standard metrics derived from the confusion matrix on the test set (Hossin & Sulaiman, 2015): correct classification rate (CCR/accuracy), precision, recall (sensitivity), F-measure (F1-score), and kappa statistic (Cohen, 1960; Landis & Koch, 1977). Model performance metrics are explained in Figure 3, which also shows the confusion matrix from which we derived these metrics. (Further detail about this matrix is presented in Table 2).

Figure 3

Model Performance Metrics

Accuracy (CCR)

Percentage of all correct predictions

Example: 80 correct out of 100 students = 80% accuracy

Precision

When model predicts an outcome, how often is it correct?

Example: Predicted "Graduated" 20 times, 18 actually graduated = 90% precision

Recall (Sensitivity)

Of all students with an outcome, how many did model identify?

Example: 25 actually graduated, model found 20 of them = 80% recall

F-Measure (F1-Score)

Balance between Precision and Recall (harmonic mean)

Single metric combining both: 85% F-measure = good balance

Kappa Statistic

Performance compared to random guessing

Interpretation Scale:

≤ 0.00: No better than chance
0.01-0.20: Slight agreement
0.21-0.40: Fair agreement
0.41-0.60: Moderate agreement
0.61-0.80: Substantial agreement
0.81-1.00: Almost perfect

Confusion Matrix Example:

Actual \ Predicted	Dropout	Graduated	Active
Dropout	82,182	50	18,101
Graduated	634	15,459	2,112
Active	6,651	306	17,123

Correct predictions
Incorrect predictions

Using multiple metrics provides a comprehensive view of model performance. A model might achieve high CCR but low recall for a particular class, meaning it is correct most of the time but misses many cases of that outcome. In the Turkish OHS context, the ideal model would be one that not only achieves high overall CCR but also reliably identifies at-risk students (high recall for the dropout class) without excessive false alarms (high precision). By evaluating models across multiple metrics, we were able to identify which algorithm provided the best balance of performance characteristics for the specific goal of predicting student outcomes.

Deployment

This phase involved reporting the findings, discussing their implications, and formulating recommendations, as presented below.

Findings

The classification analyses were performed iteratively across the four preprocessing stages:

- stage 1: initial modeling using most available features with minimal transformation (baseline).
- stage 2: removal of features with high percentages of missing values or limited applicability (e.g., the features previous school, previous system, and military service status) to create a more complete and potentially robust dataset.
- stage 3: transformation and simplification of nominal features. For example, employment status and disability status were binarized (yes/no). Categories within enrollment type were consolidated. The region feature replaced detailed location data. This stage aimed to reduce complexity and potentially improve model generalizability.
- stage 4: discretization of continuous variables (age, total credit count) using different binning strategies (frequency-based, range-based) and varying numbers of bins (2, 4, 8). This explored whether converting continuous features into categorical ones improved performance for certain algorithms.

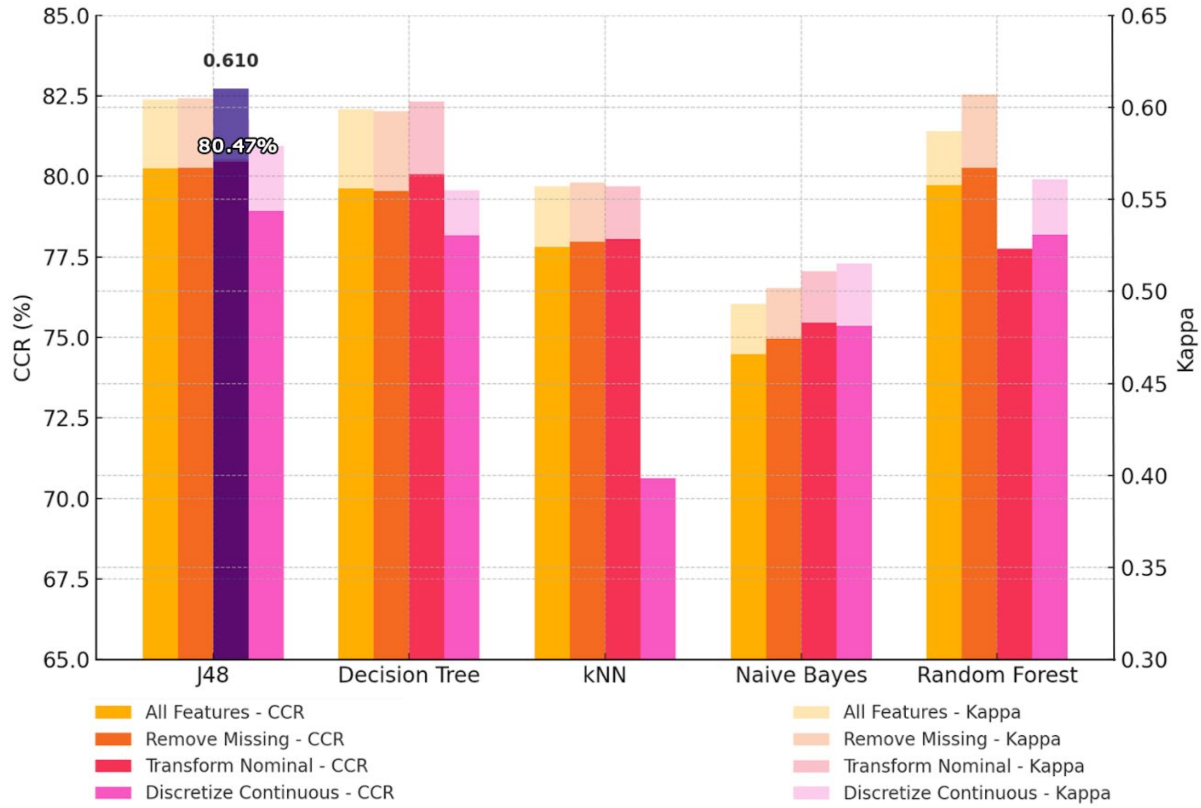
The performance of the five selected algorithms was evaluated at each stage using the test dataset.

Model Performance Comparison

The CCR and kappa metrics for the models developed using the aforementioned datasets are presented in Figure 4.

Figure 4

Comparison of Algorithm Performance Across Preprocessing Stages



Note. The purple bar highlights the most accurate algorithm, J48, and its performance metrics. CCR = correct classification rate.

Analysis across preprocessing stages showed accuracy (CCR) ranging from 73.36% to 80.47%. The highest performance was consistently achieved after the third preprocessing stage (simplifying nominal features). Within this stage, the J48 algorithm yielded the best results, achieving a CCR of 80.47% and a Kappa statistic of 0.61. This kappa value indicates substantial agreement (Landis & Koch, 1977), suggesting a robust model performing significantly better than chance.

Detailed Performance of the Best Model (J48—Stage 3)

The confusion matrix for the J48 model, the best performing on the test set (shown earlier in Figure 3), is presented in Table 2.

Table 2

Confusion Matrix for the J48 Model (N Instances in Test Set)

	Predicted			Total actual	Recall, %
	Dropout	Graduated	Active		
Actual					
Dropout	82.182	50	18.101	100.333	91.5
Graduated	634	15.459	2.112	18.205	97.7
Active	6.651	306	17.123	24.080	45.8
Total Predicted	89.467	15.815	37.336	142.618	
Precision, %	90.1	88.2	63.5		

The model demonstrates excellent performance for dropout and graduated classes (recall > 91.5%). Performance is notably lower for the active/delayed graduation class, particularly recall (45.8%). This difficulty in identifying students persisting beyond four years warrants further discussion.

The overall performance metrics for the best J48 model are shown in Table 3.

Table 3

Overall Performance Metrics for the Best J48 Model

Algorithm	Weighted avg recall, %	Weighted avg precision, %	Weighted avg F-measure, %	CCR, %	Kappa
J48	78.48	79.31	78.89	80.47	0.61

Note. CCR = correct classification rate.

These balanced overall metrics indicate a robust model that performed consistently across different facets of classification, minimizing bias toward predicting one class at the expense of others (Ben-David, 2008). This level of performance aligns well with benchmarks often considered effective for practical application in educational settings (Baker & Inventado, 2014; Romero & Ventura, 2010).

Feature Importance

Feature importance analysis for the best J48 model (Stage 3 data) revealed the relative importance of predictors (Figure 5).

Figure 5

Feature Importance Weights for the Best J48 Model

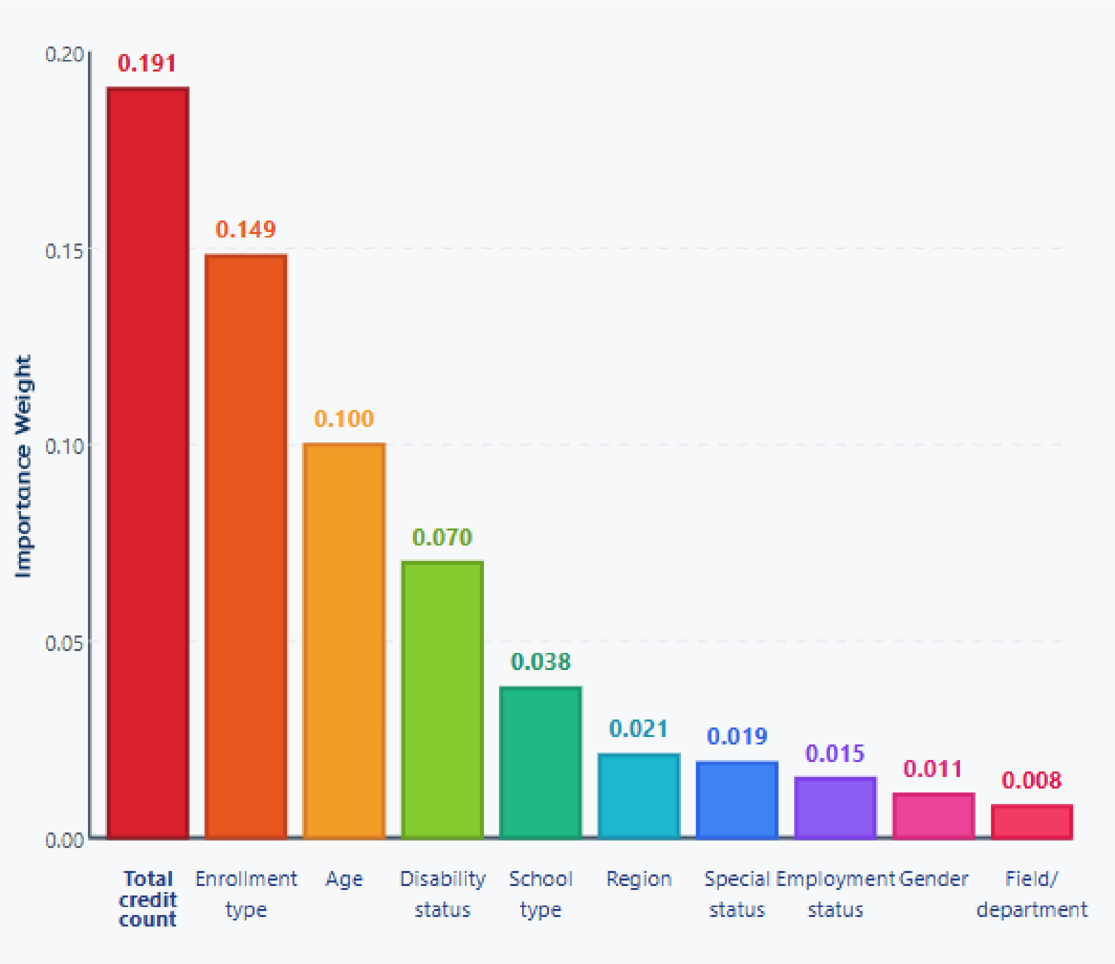


Figure 5 presents the feature importance weights, revealing each predictor's contribution to model performance. Academic progress indicators (total credit count, enrollment type) demonstrate greater predictive power than demographic variables.

Interpretation of Decision Rules

The J48 algorithm generated a complex decision tree (2,832 branches, 2,045 leaves) expressible as IF-THEN rules. A section from the tree structure of the J48 model is given in Figure 6.

Figure 6

A Section From the Tree Structure of the J48 Model

```
TOTALCREDIT > 48
|   REGISTRATION_TYPE = PRIMARY_SCHOOL_GRADUATE
|   |   TOTALCREDIT <= 91
|   |   |   AGE <= 18: ACTIVE (3693.0/941.0)
|   |   |   AGE > 18
|   |   |   |   TOTALCREDIT <= 66
|   |   |   |   |   FIELD/DEPARTMENT = NONE
|   |   |   |   |   |   AGE <= 19: ACTIVE (418.0/151.0)
|   |   |   |   |   |   AGE > 19
|   |   |   |   |   |   |   SCHOOL_TYPE = OPEN HIGH SCHOOL
|   |   |   |   |   |   |   WORKING = NO
|   |   |   |   |   |   |   SPECIAL_STATUS = NONE
|   |   |   |   |   |   |   |   REGION = TRNC: DROPOUT (1.0)
|   |   |   |   |   |   |   |   REGION = TRC: DROPOUT (599.0/235.0)
|   |   |   |   |   |   |   |   REGION = TRB
|   |   |   |   |   |   |   |   |   GENDER = FEMALE
|   |   |   |   |   |   |   |   |   |   TOTALCREDIT <= 64: DROPOUT (154.0/48.0)
|   |   |   |   |   |   |   |   |   |   TOTALCREDIT > 64: ACTIVE (26.0/11.0)
|   |   |   |   |   |   |   |   |   GENDER = MALE
|   |   |   |   |   |   |   |   |   |   AGE <= 29: ACTIVE (52.0/18.0)
|   |   |   |   |   |   |   |   |   |   AGE > 29: DROPOUT (42.0/11.0)
|   |   |   |   |   |   |   |   REGION = TRA
|   |   |   |   |   |   |   |   |   GENDER = FEMALE: ACTIVE (70.0/32.0)
```

While the full tree is impractical to display, representative rules highlight the model's logic as shown in Figure 7.

Figure 7

IF-THEN Decision Rule Examples of the J48 Model

These examples show the model combining multiple features, with total credit count often serving as a primary splitting criterion, reflecting its high importance.

Conclusion and Discussion

This study employed EDM classification analysis on a large dataset (N = 484,158) to predict the end-of-term status of students in Turkish OHSs. Departing from common binary classifications, this research's three-class approach (graduated, active, dropped out), despite complexities such as imbalance, offers a more nuanced representation of learner trajectories, contributing to an underdeveloped EDM area (Martins et al., 2023).

Following the CRISP-DM framework, we applied data transformation and manual feature selection to a dataset containing demographic, academic, and administrative features. Multiple classification algorithms were evaluated, acknowledging that no single algorithm is universally superior. The J48 algorithm emerged as the most effective, achieving an 80.47% correct classification rate (CCR) and a kappa value of 0.61, accurately predicting the status for four out of five students in the test set. This performance is notable given the three-class complexity and surpasses many binary classification studies. While J48 often performs well in EDM (Al-Barrak & Al-Razgan, 2016; Dekker et al., 2009; Kabakchieva, 2013; Kumar & Vijayalakshmi, 2011; Mishra et al., 2014; Sumitha & Vinothkumar, 2016), other algorithms such as random forest (Şara et al., 2015), approximate nearest neighbor (Ibrahim & Rusli, 2007), support vector machine (Delen, 2010), and k nearest neighbors (Yukselturk et al., 2014) have shown better results in specific contexts, likely due to differences in dataset size, feature structures, feature variety, and class distributions.

Through the lens of established student persistence theories, total credit count emerged as the most important predictor in this study, strongly aligning with Tinto's (1975) emphasis on academic integration. In OHSs, where traditional social integration is minimal, academic progress is the primary persistence anchor, likely reflecting achievement, motivation, and self-regulation crucial for distance learning (Zimmerman, 2002). This resonates with recent research showing academic progression's key role in distance learning retention (Colpo et al., 2024; Hannaford, 2021; Perez et al., 2018) and highlights its greater importance when social integration is limited. Shahiri and Husain (2015) also confirmed cumulative GPA (related to credits) as a reliable success indicator.

Enrollment type, the second key predictor, supports findings by Allensworth and Easton (2007) and Nagy and Molontay (2018) and aligns with Bean and Metzner's (1985) model emphasizing background and external factors for non-traditional students. This likely proxies for unobserved variables such as prior academic preparedness (students transferring after failure face higher risks; Bedel, 2013; Jimerson & Ferguson, 2007), socioeconomic status, and initial adaptation to distance learning. Its significance highlights the impact of pre-entry conditions (Astin, 1985) and life-course factors, especially for transfers from conventional schools due to academic challenges, facing academic challenges, and higher dropout risks (Bedel, 2013; Jimerson & Ferguson, 2007).

Age, disability status, region, employment, and gender were also important predictors, confirming learner characteristics' importance in distance education (Bozkurt et al., 2015; Durak et al., 2017). Intriguingly, academic/administrative variables (credits, enrollment type, age) were more predictive than demographics (gender, employment). This might suggest that in this large-scale distance setting, individual academic momentum and entry characteristics exert a more direct influence than demographic attributes, whose impact might be mediated differently than in face-to-face settings. It challenges assumptions about universal predictors, suggesting that in this context, Tinto's academic integration and Bean and Metzner's pre-entry/background factors are particularly influential due to reduced social integration and heightened demands for individual drive.

The model struggled to predict the active/delayed class (recall = 45.8%), due to group heterogeneity and the static nature of administrative data. Static data lacks dynamic indicators (engagement, momentum, intentions) needed to distinguish slow progressors from potential dropouts. Dynamic engagement data or qualitative methods could better illuminate this group's specific challenges and motivations (cf. Bowers, 2010).

Limitations include relying solely on administrative data, omitting variables such as family background, prior achievement (e.g., middle school GPA), psychometric factors, dynamic engagement metrics (e.g., login frequency), and nuanced socioeconomic indicators. Including such data improved accuracy elsewhere (Márquez-Vera et al., 2013; Şara et al., 2015) and could better differentiate risk levels, especially for the active/delayed-graduation group. Self-reported data reliability (e.g., employment) is another limitation. Despite limitations, this study contributes to a large-scale, multi-class predictive analysis within Türkiye's open secondary education system. It empirically shows academic progress and initial enrollment conditions outweigh demographics here, refining persistence theories (Bean & Metzner, 1985; Tinto, 1975) for large-scale distance learning. The findings offer an evidence base for targeted, data-driven interventions and policies, moving beyond generic assumptions. Future research should integrate comprehensive, dynamic data, explore real-time prediction, and use clustering to better understand student subgroups.

Recommendations

Based on the findings and discussion, the following recommendations are proposed:

For Practitioners

The implementation of data-driven early warning systems (EWS) is strongly recommended to identify at-risk students using key predictors such as low credit accumulation rates, high-risk enrollment types, and specific age profiles, with these systems integrated into existing infrastructure while addressing technical needs, data integration, staff training, and ethical guidelines. Targeted interventions should be developed based on specific predictors: for students with low credit accumulation, institutions should offer proactive academic counseling, study skills workshops, and review sessions for failed courses; for high-risk enrollment types (particularly transfers due to failure), mandatory enhanced orientation programs, mentor assignments, and early support access are essential. Academic momentum should be closely monitored by tracking credit accumulation rates over time, triggering interventions when progress slows relative to enrollment duration, and implementing system alerts for below-threshold credit counts.

Institutions should enhance onboarding and communication by tailoring activities to enrollment type and identified risks, using targeted communications for important reminders and motivational feedback. Systematic investigation of barriers through periodic surveys and focus groups, especially targeting at-risk students, can help understand specific obstacles and develop strategies to address issues such as exam absenteeism. Practitioners should advocate for enriching enrollment data with crucial information such as prior academic achievement, detailed socioeconomic indicators, and distance education readiness to improve predictive accuracy. Where available, course management system integration should be explored to leverage interaction data for finer-grained analysis of student progress and risk identification.

For Researchers

Future research should integrate diverse data sources while examining heterogeneity within student groups, especially the active/delayed graduation cohort, through mixed-methods approaches. Longitudinal modeling techniques such as survival analysis would track evolving risks, while advanced machine learning and explainable artificial intelligence could improve prediction accuracy without sacrificing interpretability (Er, 2023; Goran et al., 2024). Model refinement through hyperparameter optimization and class imbalance techniques offers performance enhancement opportunities as demonstrated by recent studies (Althaqafi et al., 2025; Sahlaoui et al., 2024). Implementation of early warning systems requires rigorous evaluation of their impact on retention outcomes. Cross-context comparisons across educational systems would identify universal versus context-specific predictors, while clustering analyses could reveal distinct student subgroups. Continued exploration of diverse algorithms and validation methods remains essential for robust research comparisons.

Limitations

This study's limitations include restricted feature availability from the administrative student information system, omitting potentially influential variables such as prior academic achievement, detailed socioeconomic factors, and behavioral indicators. These omissions may constrain the models' ability to fully capture persistence factors and affect predictor interpretation. Additionally, computational constraints with our large dataset (N = 484,158) necessitated using stratified holdout validation rather than k-fold cross-validation. These limitations suggest opportunities for future research leveraging more comprehensive datasets and enhanced computational resources.

Acknowledgements

This study is derived from the PhD dissertation by the first author, Ahmet Polat, submitted to Sakarya University, Institute of Educational Sciences, titled “Examining Dropout and Graduation Status of Open High School Students Using Educational Data Mining”, supervised by the second author, Professor Mehmet Barış Horzum.

References

Adıgüzel, M. U. (2016). *Öğrencilerin mesleki açık öğretim lisesini seçme nedenleri (Kayseri ili örneği)* [The reasons for selecting vocational open education high school by students (Kayseri sample)] (Thesis No: 446354) [Master's thesis, Erciyes University]. YÖK Thesis Center. https://tinyurl.com/38wupbwe

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, *6*(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year*. Consortium on Chicago School Research. https://consortium.uchicago.edu/sites/default/files/2018-10/07%20What%20Matters%20Final.pdf

Altair Engineering (2026). *RapidMiner* (Version 9) [Computer software]. https://docs.rapidminer.com/9.0/studio/getting-started/index.html

Althaqafi, T., Saleem, F., & AL-Ghamdi, A. A. M. (2025). Enhancing student performance prediction: The role of class imbalance handling in machine learning models. *Discover Computing*, *28*(1), Article 79. https://doi.org/10.1007/s10791-025-09576-4

Astin, A. W. (1985). *Achieving educational excellence*. Jossey-Bass.

Aydın, A., Sarıer, Y., & Uysal, Ş. (2012). Sosyoekonomik ve sosyokültürel değişkenler açısından PISA matematik sonuçlarının karşılaştırılması [The comparative assessment of the results of PISA mathematical literacy in terms of socioeconomic and sociocultural variables]. *Eğitim ve Bilim, 37*(164), 20–29. https://doi.org/10.15390/ES.2012.1024

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, *55*(4), 485–540. https://doi.org/10.3102/00346543055004485

Bedel, A. (2013). Sınıf tekrarı yapan ve yapmayan öğrencilerin akademik güdülenme ve kaygı düzeylerinin karşılaştırılması [Comparison of academic motivation and anxiety levels of students in classes repeated and not repeated]. *Milli Eğitim Dergisi*, *43*(200), 111–122. https://dergipark.org.tr/tr/download/article-file/457551

Bedel, Y. (2006). *Açıköğretim lisesi mesleki açıköğretim programı akademik danışmanlık hizmetlerinin etkililiğinin değerlendirilmesi* [Evaluation of the effectiveness of open education high school vocational curriculum academic advisory services] (Thesis No: 204612) [Master's thesis, Ankara University]. YÖK Thesis Center. https://tinyurl.com/4r72kf9v

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications*, *34*(2), 825–832. https://doi.org/10.1016/j.eswa.2006.10.022

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk—Predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research, and Evaluation*, *15*(1), Article 7. https://doi.org/10.7275/r4zq-9c31

Bozkurt, A., Akgun-Ozbek, E., Yilmazel, S., Erdogdu, E., Ucar, H., Guler, E., Sezgin, S., Karadeniz, A., Sen-Ersoy, N., Goksel-Canbek, N., Dincer, G. D., Ari, S., & Aydin, C. H. (2015). Trends in distance education research: A content analysis of journals 2009–2013. *The International Review of Research in Open and Distributed Learning*, *16*(1), 330–363. https://doi.org/10.19173/irrodl.v16i1.1953

Çalışkan, Ş., Karabacak, M., & Meçik, O. (2013). Türkiye'de eğitim-ekonomik büyüme ilişkisi: 1923–2011 (Kantitatif bir yaklaşım) [Relationship between education and economic growth in Turkey: 1923–2011 (A quantitative approach)]. *Yönetim Bilimleri Dergisi*, *11*(21), 29–48. https://dergipark.org.tr/tr/download/article-file/46273

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS Inc. https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf

Çiçek, S. (2005). *Kız meslek lisesi açık lise programı uygulamalarında karşılaşılan sorunlar ve çözüm önerileri (Elâzığ, Malatya ve Diyarbakır İlleri örneği)* [Problems in practice in girls' open vocational secondary school and recommendations for solutions (Elazığ, Malatya, and Diyarbakır samples)] (Thesis No: 188405) [Master's thesis, Fırat University]. YÖK Thesis Center. https://tinyurl.com/4j6wep9t

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Colpo, M. P., Primo, T. T., de Aguiar, M. S., & Cechinel, C. (2024). Educational data mining for dropout prediction: Trends, opportunities, and challenges. *Revista Brasileira de Informática na Educação, 32*, 220–256. https://doi.org/10.5753/rbie.2024.3559

Çuhadar Öncü, E. (2017). *Ortaöğretim öğrencilerinin açık öğretim lisesine geçiş nedenlerine ilişkin öğrenci, okul yöneticisi, öğretmen ve veli görüşleri* [Opinions of students, principals, teachers, and parents about factors leading to high school students enrolling in open high school] (Thesis No: 468228) [Master's thesis, Ankara University]. YÖK Thesis Center. https://tinyurl.com/mtdzbyje

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Proceedings of the International Working Group on Educational Data Mining* (pp. 41–50). International Working Group on Educational Data Mining. https://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Demiray, U., & Sağlık, M. (2003). Açıköğretim fakültesi ve Açıköğretim lisesi uygulamalarını içeren araştırmalara ilişkin bir değerlendirme [General evaluation of the research on open education faculty and open high school applications]. *The Turkish Online Journal of Educational Technology*, *2*(4), 50–59. https://tojet.net/articles/v2i4/248.pdf

Demirtaş, Z., Tutkun, Ö. F., & Arslan, A. (2017). Mesleki açık öğretim lisesi (MAÖL) öğrencilerinin mesleki eğitime yönelik görüşleri [Vocational open education high school students' opinions on vocational education]. *PESA Uluslararası Sosyal Araştırmalar Dergisi*, *3*(4), 231–240. https://dergipark.org.tr/tr/download/article-file/454515

Dere, S. (2002). *Açıköğretim lisesine kayıtlı engelli öğrencilerin sisteme ilişkin görüş ve beklentilerinin değerlendirilmesi* [The evaluation of expectations and opinions of disabled students who are enrolled at open high school] (Thesis No: 117551) [Master's thesis, Ankara University]. YÖK Thesis Center. https://tez.yok.gov.tr

Djulovic, A., & Li, D. (2013). Towards freshman retention prediction: A comparative study. *International Journal of Information and Education Technology*, *3*, 494–500. https://www.ijiet.org/papers/324-K045.pdf

Durak, G., Çankaya, S., Yünkül, E., Urfa, M., Topraklıklıoğlu, K., Arda, Y., & İnam, N. (2017). Trends in distance education: A content analysis of master's thesis. *TOJET: The Turkish Online Journal of Educational Technology*, *16*(1), 203–218. https://files.eric.ed.gov/fulltext/EJ1124887.pdf

Er, E. (2023). An explainable machine learning approach to predicting and understanding dropouts in MOOCs. *Kastamonu Education Journal*, *31*(1), 143–154. https://doi.org/10.24106/kefdergi.1246458

Ereş, F. (2005). Eğitimin sosyal faydaları: Türkiye-AB karşılaştırması [Social contribution of education in Turkey: Comparison of Turkey and European Union]. *Milli Eğitim Dergisi*, *33*(167). https://dhgm.meb.gov.tr/yayimlar/dergiler/Milli_Egitim_Dergisi/167/orta3-eres.htm

Eurostat. (2026). *NUTS—Nomenclature of territorial units and statistical regions—Level 1*. European Union. https://ec.europa.eu/eurostat/documents/345175/17779945/2021-NUTS-1-map.pdf/0d90fc54-903c-347b-8a12-aa6ced7a4d25?t=1698686269555

Goran, R., Jovanovic, L., Bacanin, N., Stanković, M. S., Simic, V., Antonijevic, M., & Zivkovic, M. (2024). Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques. *IEEE Access*, *12*, 122377–122400. https://doi.org/10.1109/ACCESS.2024.3446653

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Hannaford, L., Cheng, X., & Kunes-Connell, M. (2021). Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. *Nurse Education Today*, *99*, Article 104784. https://doi.org/10.1016/j.nedt.2021.104784

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heppen, J. B., & Therriault, S. B. (2008). *Developing early warning systems to identify potential high school dropouts* [Issue Brief]. National High School Center. https://files.eric.ed.gov/fulltext/ED521558.pdf

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Ibrahim, Z., & Rusli, D. (2007, September 5). *Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression* [Presentation]. 21st Annual SAS Malaysia Forum, Kuala Lumpur. https://www.researchgate.net/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression

International Educational Data Mining Society. (n.d.). *Home*. https://educationaldatamining.org/

Jensen, K. (2012). *CRISP-DM process diagram*. Wikimedia Commons. https://commons.wikimedia.org/w/index.php?curid=24930610

Jimerson, S. R., & Ferguson, P. (2007). A longitudinal study of grade retention: Academic and behavioral outcomes of retained students through adolescence. *School Psychology Quarterly*, *22*(3), 314–339. https://doi.org/10.1037/1045-3830.22.3.314

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61–72. https://doi.org/10.2478/cait-2013-0006

Kılınç, Ç. (2015). *Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi* [Examining the effects on university student success by data mining techniques] (Thesis No: 415460) [Master's thesis, Eskişehir Osmangazi University]. YÖK Thesis Center. https://tinyurl.com/2yvzw3xu

Kızmaz, Z. (2004). Öğrenim düzeyi ve suç: Suç-okul ilişkisi üzerine sosyolojik bir araştırma [Educational level and crime: A sociological research on the relation between school and crime]. *Fırat University Journal of Social Science*, *14*(2), 291–319. https://dergipark.org.tr/en/download/article-file/71991

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

Kumar, S. A., & Vijayalakshmi, M. N. (2011). Efficiency of decision trees in predicting student's academic performance. In D. C. Wyld & M. Wozniak (Eds.), *First International Conference on Computer Science, Engineering and Applications: CCSEA proceedings* (pp. 335–343). AIRCC Publishing. https://airccj.org/CSCP/vol1/csit1230.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lassibille, G., & Navarro-Gomez, L. (2008). Why do higher education students drop out? Evidence from Spain. *Education Economics*, *16*(1), 89–105. https://doi.org/10.1080/09645290701523267

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, *9*(15), Article 3093. https://doi.org/10.3390/app9153093

Luy, M., Zannella, M., Wegner-Siegmundt, C., Minagawa, Y., Lutz, W., & Caselli, G. (2019). The impact of increasing education levels on rising life expectancy: A decomposition analysis for Italy, Denmark, and the USA. *Genus*, *75*, Article 11. https://doi.org/10.1186/s41118-019-0055-0

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, *13*(8), Article 4702. https://doi.org/10.3390/app13084702

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. In J. P. Mittal (Chair), *Proceedings: Fourth International Conference on Advanced Computing & Communication Technologies* (pp. 255–262). IEEE. https://doi.org/10.1109/ACCT.2014.105

Ministry of National Education. (2012). *Millî eğitim İstatistikleri: Örgün eğitim 2011–2012* [National education statistics: Formal education 2011-2012].

https://sgb.meb.gov.tr/meb_iys_dosyalar/2025_06/02135804_meb_istatistikleri_orgun_egitim_2011_2012.pdf

Ministry of National Education. (2015). *Millî Eğitim Bakanlığı: 2015–2019 stratejik planı* [Ministry of National Education: 2015-2019 Strategic plan]. http://sgb.meb.gov.tr/meb_iys_dosyalar/2015_09/10052958_10.09.2015sp17.15imzasz.pdf

Ministry of National Education. (2017). *Millî eğitim İstatistikleri: Örgün eğitim 2016/17* [National education statistics: Formal education 2016/17]. https://sgb.meb.gov.tr/meb_iys_dosyalar/2017_09/08151328_meb_istatistikleri_orgun_egitim_2016_2017.pdf

Ministry of National Education. (2019). *Milli eğitim İstatistikleri: Örgün eğitim 2018/19* [National education statistics: Formal education 2018/19]. http://sgb.meb.gov.tr/meb_iys_dosyalar/2019_09/30102730_meb_istatistikleri_orgun_egitim_2018_2019.pdf

Ministry of National Education. (2021). *Millî eğitim İstatistikleri: Örgün eğitim 2020/21* [National education statistics: Formal education 2020/21]. https://sgb.meb.gov.tr/meb_iys_dosyalar/2021_09/10141326_meb_istatistikleri_orgun_egitim_2020_2021.pdf

Ministry of National Education. (2022). *Ortaöğretim kurumlarına ilişkin merkezi sınav* [The central exam for secondary education institutions]. https://cdn.eba.gov.tr/icerik/2022/06/2022_LGS_rapor.pdf

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* (pp. 000389–000394). IEEE. https://doi.org/10.1109/INES.2018.8523888

Organisation for Economic Co-operation and Development (2018). *Education at a glance 2018: OECD indicators*. OECD. https://doi.org/10.1787/eag-2018-en

Organisation for Economic Co-operation and Development. (2021). *Education at a glance 2021: OECD indicators*. OECD. https://doi.org/10.1787/b35a14e5-en

Özkahveci, Ö. (2001). *Açıköğretim lisesi mesleki açık öğretim programı öğrencileri ile kız meslek liseleri öğrencilerinin akademik başarılarının karşılaştırılması* [Comparison of academic achievement of open high school vocational education program students and girls' vocational high school students] (Thesis No: 117810) [Master's thesis, Gazi University]. YÖK Thesis Center. https://tez.yok.gov.tr

Perez, B., Castellanos, C., & Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. In A. Orjuela-Canon, J. Figueroa-Garcia, & J. Arias-Londono (Eds.), *Applications of computational intelligence, ColCACI 2018* (pp. 111–125). Springer. https://doi.org/10.1007/978-3-030-03023-0_10

Polat, A. (2021). *Açık öğretim liseleri öğrencilerinin okul terki ve mezuniyet durumlarının eğitsel veri madenciliği ile incelenmesi* [Examining dropout and graduation status of open high school students using educational data mining] (Thesis No: 681548) [Doctoral dissertation, Sakarya University]. YÖK Thesis Center. https://tinyurl.com/22k97c8s

Randler, C., Horzum, M. B., & Vollmer, C. (2014). The influence of personality and chronotype on distance learning willingness and anxiety among vocational high school students in Turkey. *The International Review of Research in Open and Distributed Learning*, *15*(6), 93–110. https://doi.org/10.19173/irrodl.v15i6.1928

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

Rosenzweig, M. R. (2010). Microeconomic approaches to development: Schooling, learning, and growth. *Journal of Economic Perspectives*, *24*(3), 81–96. https://doi.org/10.1257/jep.24.3.81

Rumberger, R. W. (2020). The economics of high school dropouts. In S. Bradley & C. Green (Eds.), *The economics of education: A comprehensive overview* (2nd ed., pp. 149–158). Academic Press. https://doi.org/10.1016/B978-0-12-815391-8.00012-4

Rumberger, R. W., & Lim, S. A. (2008). *Why students drop out of school: A review of 25 years of research*. California Dropout Research Project. https://ej.issuelab.org/resources/11658/11658.pdf

Şahin, B. (2017). *Açık öğretim lisesi öğrenci ve mezunlarının katılım örüntüleri (Ankara İli Çankaya İlçesi örneği)* [Participation patterns of open high school students and graduates (Ankara city Çankaya district sample)] (Thesis No: 468273) [Doctoral dissertation, Ankara University]. YÖK Thesis Center. https://tinyurl.com/4kk3s7jx

Sahlaoui, H., Alaoui, E. A. A., Agoujil, S., & Nayyar, A. (2024). An empirical assessment of SMOTE variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models. *Education and Information Technologies*, *29*(5), 5447–5483. https://doi.org/10.1007/s10639-023-12007-w

Şara, N.-B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In M. Verleysen (Ed.), *Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 319–324). ESANN. https://core.ac.uk/download/pdf/269280541.pdf

Sarıhan, Ş. (2010). *Mesleki açık öğretim lisesindeki öğrenci hizmetlerinin etkililiği konusunda öğrenci görüşleri (Ankara ili örneği)* [Student opinions on effectiveness of student service at open vocational high school (Ankara sample)] (Thesis No: 279946) [Master's thesis, Ankara University]. YÖK Thesis Center. https://tinyurl.com/bjupprjn

Şentürk, E. (2009). *Mesleki açık öğretim lisesi bilişim teknolojileri alanında verilen eğitimin etkililiğine yönelik öğrenci görüşlerinin değerlendirilmesi* [Evaluation of students' opinions towards effectiveness of education in vocational open high school's information and communication technologies department] (Thesis No: 235277) [Master's thesis, Uludağ University]. YÖK Thesis Center. https://tinyurl.com/dsrpwx4r

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Sipahi, K. B. (2019). *Mesleki açık öğretim lisesi öğrencilerinin uzaktan eğitime yönelik algıladıkları engeller ile tutumları arasındaki ilişkinin incelenmesi* [Researching the relationship between barriers and attitude perceived by vocational open high school students towards distance education] (Thesis No: 538965) [Master's thesis, Sakarya University]. YÖK Thesis Center. https://tinyurl.com/5n8u23a4

Soylu, S. (2014). *Mesleki açık öğretim lisesi muhasebe eğitiminde karşılaşılan sorunlar ve çözüm önerileri* [Problems encountered in accounting education at vocational open high school and proposed solutions] (Thesis No: 366242) [Master's thesis, Gazi University]. YÖK Thesis Center. https://tinyurl.com/c6x5czw9

Sumitha, R., & Vinothkumar, E. S. (2016). Prediction of students outcome using data mining techniques. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *2*(6), 132–139. https://ijseas.com/volume2/v2i6/ijseas20160615.pdf

Tanner, D. E., Newbold, B. L., & Johnson, D. B. (2003). *Academic achievement as a dropout predictor* (ED478173). ERIC. https://files.eric.ed.gov/fulltext/ED478173.pdf

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Türkiye Cumhuriyeti Resmî Gazete [*Republic of Türkiye official gazette*], No. 23084. (1997, August 18). Law no. 4306 (pp. 2–6). https://www.resmigazete.gov.tr/arsiv/23084.pdf

Türkiye Cumhuriyeti Resmî Gazete [*Republic of Türkiye official gazette*], No. 26215. (2006, July 1). Dokuzuncu kalkınma planı [Ninth development plan]. https://www.resmigazete.gov.tr/eskiler/2006/07/20060701M1-2.pdf

Türkiye Cumhuriyeti Resmî Gazete [*Republic of Türkiye official gazette*], No. 28261. (2012, April 11). Law no. 6287. http://www.resmigazete.gov.tr/eskiler/2012/04/20120411-8.htm

Türkiye Cumhuriyeti Resmî Gazete [*Republic of Türkiye official gazette*], No. 28699. (2013, July 6). Onuncu kalkınma planı [Tenth development plan]. https://www.resmigazete.gov.tr/eskiler/2013/07/20130706M1-1-1.doc

Umar, U. (2016). *Evaluation of school administrators by using data mining techniques* (Thesis No: 436503) [Master's thesis, Fırat University]. YÖK Thesis Center. https://tinyurl.com/3scyjutj

United Nations Department of Economic and Social Affairs. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda

United Nations Development Program. (2021). *Human development report 2020: The next frontier: Human development and the Anthropocene*. https://hdr.undp.org/content/human-development-report-2020

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://doi.org/10.1016/C2015-0-02071-8

Yavuz, H. (2014). *Mesleki açık öğretim lisesi öğrencilerinin sunulan hizmetlerin etkililiği konusunda görüşleri ve motivasyon düzeyleri* [Student opinions on effectiveness of student services and motivation levels at vocational open high school] (Thesis No: 367965) [Master's thesis, Fatih University]. YÖK Thesis Center. https://tinyurl.com/3jpe74k4

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

Yurdakul, S. (2015). *Veri Madenciliği ile lise öğrenci performanslarının değerlendirilmesi* [Assessment of high school students' performance by means of data mining] (Thesis No: 418479) [Master's thesis, Kırıkkale University]. YÖK Thesis Center. https://tinyurl.com/3t3syc6k

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2