Düzce University Journal of Science and Technology
Düzce Üniversitesi Bilim ve Teknoloji Dergisi
14 (3), 2026, 759–770.
https://doi.org/10.29130/dubited.1830199
e-ISSN: 2148-2446
https://dergipark.org.tr/en/pub/dubited
Research Article

Career Satisfaction Prediction Using Machine Learning: Feature Importance and Performance Analysis

Murat Urfalıoğlu[1]*

[1] Bolu Abant İzzet Baysal University, Gerede Faculty of Applied Sciences, Department of Management Information Systems, Bolu, Türkiye.
ror.org/01x1kqx83

*Corresponding author:
murat.urfalioglu@ibu.edu.tr

Received: 25.11.2025
Revision: 12.04.2026
Accepted: 19.04.2026
Online: 24.07.2026

Abstract: Career satisfaction is a dynamic and multifaceted phenomenon shaped by professional development, individual experiences, and career-related outcomes. This study aims to predict individuals' career satisfaction using a machine learning-based regression framework with multidimensional features. The dataset consists of 400 observations, including demographic, academic, professional, and social attributes. Six machine learning algorithms, namely Linear Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting (GB), were employed and comparatively evaluated. To enhance methodological rigor, potential data leakage risk was assessed through correlation analysis, and a restricted feature set was constructed by excluding variables highly correlated with the target variable. Model performance was evaluated using both cross-validation and train–test split approaches. Hyperparameter optimization was performed using GridSearchCV, and model performance was evaluated with coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) metrics. The results indicate that tree-based methods tend to outperform other models. After optimization, the GB model achieved the best performance on the held-out test set, with an R^2 value of 0.9213. Feature importance analysis using SHapley Additive exPlanations (SHAP) revealed that individual and career-related variables had a stronger influence on predictions. Overall, this study provides a methodologically transparent framework for predicting career satisfaction and supports data-driven decision-making in career-related contexts.

Keywords: *Career satisfaction, Machine learning, Regression analysis, Feature importance, SHAP*

I. INTRODUCTION

Career planning and management is a multifaceted process that includes theoretical frameworks, individual initiative, and organizational support. Career is defined as an evolutionary sequence of work experiences that develop over time throughout an individual's life (Arthur, 2008). This definition does not view career as limited to a specific job or profession but encompasses the entirety of sequential and parallel work experiences acquired throughout an individual's life. Unlike the terms "job" or "profession," the concept of career implies temporal continuity and encompasses the processes of learning, change, and development in an individual's working life. This approach considers career as a multidimensional phenomenon in both individual and societal contexts. It is based on an interdisciplinary understanding that combines psychological, sociological, and organizational perspectives. Thus, career is positioned as a dynamic process that reflects not only an individual's professional advancement but also their experiential transformation within the life cycle (Arthur, 2008; Arthur et al., 1989; Gunz & Peiperl, 2007).

Career satisfaction is an indicator of an individual's perception of the level of success and fulfillment achieved throughout their professional life (Greenhaus et al., 1990). This definition considers career satisfaction not only as an emotional response to the current job but also as a general assessment of the individual's entire career process, including success, career goals, promotions received, certifications obtained, and income. In this context, career satisfaction is closely related to factors such as the degree to which the individual achieves their career goals, the progress they make, work-life balance, professional development, income, and sense of achievement throughout their career (Çakmak Otluoğlu, 2014; Greenhaus et al., 1990; Judge et al., 1995; Lounsbury et al., 2004; Ng et al., 2005). Therefore, career satisfaction is a comprehensive assessment that reflects the extent to which the experiences and gains obtained by the individual during their career process match their personal expectations and values.

Machine learning refers to computational methods that enable systems to learn patterns from data and make predictions or decisions without being explicitly programmed. By identifying statistical relationships within data, machine learning models can improve their performance over time and provide flexible and scalable solutions for complex problems. In this sense, machine learning is recognized as an interdisciplinary field at the intersection of

Cite as: Urfalıoğlu, M. (2026). Career satisfaction prediction using machine learning: Feature importance and performance analysis. *Düzce University Journal of Science and Technology, 14*(3), 759–770. https://www.doi.org/10.29130/dubited.1830199

This is an open access paper distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.

statistics, artificial intelligence, and computer science (Géron, 2017), and is often associated with concepts such as predictive analytics and statistical learning (Müller & Guido, 2017). A central objective of machine learning is to develop models that can generalize beyond the training data and accurately predict unseen observations. Rather than memorizing existing data, these models aim to capture underlying patterns that enable reliable prediction in new situations (Kelleher et al., 2015). This capability makes machine learning particularly suitable for analyzing complex and multidimensional phenomena such as career satisfaction.

Machine learning systems are classified and studied according to specific criteria due to their wide variety of features. Géron (2017) states that this classification is generally divided based on three different criteria. First, systems are divided according to whether they require a training process based on human supervision. Accordingly, supervised, unsupervised, semi-supervised, and reinforcement learning methods are specified. The second criterion relates to how systems process data. Some systems learn by processing data in batches, while others adopt an online learning approach that receives data in real time and updates in real time. The third criterion focuses on the structural learning form of the model. According to this criterion, it is decisive whether systems use instance-based learning, which relies on directly comparing new data points with past examples, or model-based learning, which generalizes by discovering statistical patterns in the training data. This classification provides an important framework for understanding the suitability of machine learning techniques for different types of problems.

In this study, a supervised learning approach was adopted to model and predict career satisfaction. Supervised learning involves learning the relationship between input variables and a target variable using labeled data (Beyeler, 2017). Within this framework, the present study focuses on regression tasks, where the objective is to predict continuous target values. Accordingly, machine learning algorithms capable of modeling continuous outcomes were employed to estimate individuals' career satisfaction levels (Beyeler, 2017; Géron, 2017; Müller & Guido, 2017). Common algorithms used in supervised learning include K-Nearest Neighbors (KNN), Linear Regression (LR), Support Vector Machines (SVR), Decision Trees (DT), Random Forest (RF), and Gradient Boosting (GB). In this study, multiple machine learning algorithms were employed and compared to evaluate their predictive performance in modeling career satisfaction.

One study that addresses career prediction using machine learning is that of LeBlond et al. (2025), which examined the effect of academic experiences on long-term career and job satisfaction among adults with disabilities using the Random Forest algorithm and achieved an accuracy rate of 68.6% for job satisfaction and 72.5% for career satisfaction. Ramos-Pulido et al. (2024) examined the career satisfaction of university graduates using various data science models. In their study, they applied five different machine learning algorithms: decision tree, Random Forest, Gradient Boosting, support vector machines, and artificial neural networks. The SHAP (SHapley Additive exPlanations) method was then used to identify the salient features of the best predictive model. The study indicated that career satisfaction is related to academic education, graduate satisfaction, employment status, published scientific studies, and other factors. In literature, there are also studies that examine job satisfaction as well as career satisfaction using machine learning methods. Celbiş et al. (2023) used various machine learning techniques to understand the individual factors affecting job satisfaction among employees in the Netherlands. In their study, they applied sequential and tree-based methods for prediction and clustering. The findings of the study show that satisfaction with the social atmosphere among coworkers, salary satisfaction, and the feeling of being appreciated are the key determinants of job satisfaction. Another study related to machine learning and career is Urfalıoğlu and Aşantuğrul (2026), which focused on data-driven career prediction and faculty placement.

When the literature on career satisfaction is examined, it can be observed that machine learning algorithms are capable of achieving high levels of predictive accuracy in modeling complex educational and career-related outcomes (Celbiş et al., 2023; Holgado-Apaza et al., 2023; LeBlond et al., 2025; Ramos-Pulido et al., 2024). However, it also appears that many studies tend to rely on a limited set of algorithms and do not comprehensively address aspects such as hyperparameter optimization or advanced interpretability techniques such as SHAP. In addition, limited attention has been given to evaluating potential data leakage through alternative modeling scenarios or to assessing the stability of model performance under different feature configurations. Therefore, there is a need for more comprehensive studies that not only compare multiple algorithms but also incorporate robust validation procedures, hyperparameter optimization strategies, and interpretability approaches.

This study differs from existing approaches in the literature by providing a comparative analysis of multiple machine learning algorithms on a comprehensive dataset that includes multidimensional demographic, academic, professional, and social factors influencing career satisfaction. The primary aim of this study is to predict individuals' career satisfaction levels using these multidimensional characteristics and to identify the most influential features through a machine learning-based modeling approach. Within this framework, LR, DT, KNN, SVR, RF, and GB algorithms were employed and comparatively evaluated. In addition, the impact of hyperparameter optimization on model performance was examined in detail using GridSearchCV, and model performance was rigorously evaluated through validation strategies such as train-test splitting and cross-validation. By using a comprehensive set of performance metrics, the study evaluates not only predictive accuracy but also error behavior and generalization capability.

Furthermore, the contribution of individual features to model predictions was analyzed using SHAP-based

interpretability methods, and alternative modeling scenarios were constructed to assess potential data leakage and the stability of model performance under different feature configurations. In this respect, the study contributes to the literature by combining feature-set comparison, leakage-risk assessment, and model interpretability within the context of career satisfaction prediction.

II. MATERIAL AND METHOD

A. *Research Model*

This study is based on a supervised machine learning regression framework designed to predict individuals' career satisfaction levels. Within this framework, multiple machine learning algorithms are comparatively evaluated using a comprehensive set of features, including demographic, academic, professional, social, and career-related attributes.

The research model is structured around a systematic machine learning workflow. In this sequential process, data preprocessing procedures are first applied to prepare the dataset for modeling. Subsequently, multiple regression-based machine learning models are trained and validated using both cross-validation and train–test split approaches. To enhance the reliability and generalizability of the results, hyperparameter optimization is performed using GridSearchCV. In addition, alternative modeling scenarios are incorporated into the research design to assess the potential impact of data leakage by comparing full and restricted feature sets. Finally, model performance is evaluated using multiple performance metrics, and feature importance is analyzed to support the interpretability of the model outcomes.

B. *Dataset*

The dataset used in this study is an open-access dataset published on the Kaggle platform under the title "Education and Career Success Dataset" (Shamim, 2025). The dataset includes features related to individuals' educational background, academic performance indicators, professional experiences, skill levels, and career outcomes. Table 1 presents the attribute names and their corresponding descriptions included in the dataset.

Table 1. Dataset features and their descriptions.

Feature Name	Description	Data Type	Value Range / Category
Student_ID	Unique alphanumeric identifier assigned to each student to ensure confidentiality.	Categorical	1–400
Age	Age of the student.	Numerical	21–25
Gender	Self-reported gender identity of the student.	Categorical	"Male", "Female"
High_School_GPA	High school graduation GPA (4-point scale).	Numerical (Decimal)	2.0–4.0
SAT_Score	Total score on the SAT university entrance exam.	Numerical (Integer)	1160–1580
University_GPA	Undergraduate graduation GPA (4-point scale).	Numerical (Decimal)	2.0–4.0
Field_of_Study	Academic major or field of study.	Categorical	"Computer Science", "Medicine", "Business", etc.
Internships_Completed	Number of internships completed during undergraduate education.	Numerical (Integer)	0–4
Projects_Completed	Number of major academic projects completed (e.g., thesis, research, applied projects).	Numerical (Integer)	0–9
Certifications	Number of industry-recognized certifications earned (e.g., AWS, PMP).	Numerical (Integer)	0–5
Soft_Skills_Score	Score (1–10) reflecting communication, teamwork, and leadership skills.	Numerical (Integer)	1–10
Networking_Score	Score representing professional networking activity (e.g., LinkedIn engagement, events attended).	Numerical (Integer)	1–10
Job_Offers	Number of full-time job offers received within six months after graduation.	Numerical (Integer)	0–5
Starting_Salary	Annual starting salary (USD) for the first full-time job.	Numerical (Decimal)	$31,000–$152,000
Career_Satisfaction	Satisfaction level in the first job (1 = very dissatisfied, 10 = very satisfied).	Numerical (Integer)	5–10
Years_to_Promotion	Number of years between job entry and the first promotion.	Numerical (Integer)	1–5
Current_Job_Level	Career level at the time of data collection.	Categorical	"Entry", "Mid", "Senior", "Executive"
Work_Life_Balance	Perceived work–life balance level (1 = poor, 10 = excellent).	Numerical (Integer)	5–9
Entrepreneurship	Whether the individual has founded or co-founded a business.	Categorical (Binary)	"No"

The dataset includes individual's demographic information (student ID, age, gender), academic performance indicators (high school and university GPAs, SAT score, field of study), professional development indicators (number of internships, projects, and certifications), social skill levels (communication and networking scores), career outcomes (number of job offers, starting salary, job level, years to promotion, work–life balance, and career satisfaction), as well as entrepreneurship status. Descriptive statistical summaries for the numerical features in the dataset are presented in Table 2.

Table 2. Descriptive statistics of the dataset.

Feature	Mean	Standard Deviation (SD)	Minimum	Maximum
Age	23.00	1.41	21	25
High_School_GPA	3.62	0.18	3.20	3.90
SAT_Score	1388.60	114.31	1160	1580
University_GPA	3.44	0.28	2.80	4.00
Internships_Completed	2.50	1.06	0.00	4.00
Projects_Completed	6.23	1.90	2.00	9.00
Certifications	2.62	1.27	0.00	5.00
Soft_Skills_Score	7.93	1.31	5.00	10.00
Networking_Score	6.93	1.31	4.00	9.00
Job_Offers	2.74	1.35	0.00	5.00
Starting_Salary (USD)	87,562.50	29,685.41	31,000.00	152,000.00
Career_Satisfaction	7.79	1.37	5.00	10.00
Years_to_Promotion	2.91	1.34	1.00	5.00
Work_Life_Balance	6.41	0.93	5.00	9.00

According to the explanation provided in the "Provenance" section on the Kaggle platform, this dataset was compiled from multiple verified sources as part of a longitudinal study examining the educational background, career development, and professional outcomes of university students who graduated between 2018 and 2022. The data were collected from academic transcripts and standardized test scores obtained from university registrar offices; online surveys administered by the research team; anonymized professional networking activity retrieved via the LinkedIn API; job offer and salary information provided by university career services offices; and follow-up telephone interviews conducted at six- and twelve-month intervals after graduation (Shamim, 2025).

To assess the risk of data leakage, correlations between the predictor variables and the target variable (Career_Satisfaction) were analyzed. Features exceeding the threshold of $r>0.80$ were considered as potential sources of leakage, as they may directly encode or strongly represent the target variable. Accordingly, model performance was evaluated using features with a lower risk of leakage in order to obtain more reliable results. Correlation-based feature removal was used only as a partial safeguard against data leakage. Since the dataset lacked sufficient temporal information, this transparent screening approach was preferred to reduce, but not completely eliminate, leakage risk.

C. Data Preprocessing

Comprehensive data preprocessing procedures were carried out to enhance the reliability and performance of the machine learning model. First, the data types of all variables in the dataset were examined, revealing the presence of 5 nominal (categorical) and 14 numerical features. Second, the dataset was inspected for missing values. The analysis indicated that no missing data were present in the dataset.

In regression analysis, independent variables must be expressed in numerical form. As emphasized by Kelleher et al. (2015), when a feature is categorical, multiplying its values directly by a numerical coefficient in a regression equation is conceptually meaningless. Therefore, the nominal/categorical features in the dataset were transformed into binary/numerical representations before being included in the model. For this purpose, appropriate encoding techniques (e.g., one-hot encoding) were applied. This transformation converts categorical variables into numerical format by representing each category as a separate column. Through this transformation, all features in the dataset were converted into numerical form, thereby increasing the applicability and consistency of the machine learning models. In addition, feature scaling was applied using the StandardScaler method to ensure that variables with different ranges do not disproportionately influence model performance.

During the data cleaning process, two variables were removed from the dataset:

- Student_ID: This feature serves solely as an identifier for individual observations and does not contribute to the predictive power of the model. Since identifier variables may introduce meaningless variation into the learning process, they were excluded from the analysis.
- Entrepreneurship: This feature contains the same value ("No") for all observations. Because it shows no variability, it does not provide any informational value to the model. Removing variables with a single constant value helps reduce model complexity and prevents potential decreases in model performance.

As a result of these procedures, the dataset was prepared for modeling with 400 observations and 29 features. These features enable the analysis of the relationship between individuals' educational backgrounds and their career outcomes, as well as the development of machine learning models aimed at predicting career satisfaction.

D. Modeling and Analysis Procedure

Data processing and modeling steps were carried out using the Python programming language. The pandas and NumPy libraries were utilized for data preprocessing and numerical operations. The implementation, training, and evaluation of machine learning models were carried out using the scikit-learn library. To enhance feature importance analysis and model interpretability, the SHAP library was employed. Visualization processes were conducted using the Matplotlib library.

In this study, multiple machine learning algorithms were employed to address the regression problem of predicting individuals' career satisfaction. The selected algorithms were determined based on their ability to model both linear and non-linear relationships, their performance across different data structures, and their widespread use in the literature for predictive modeling tasks. Accordingly, six algorithms were evaluated:

RF is a tree-based ensemble learning method that constructs multiple decision trees using different subsets of the data and features, and combines their predictions to improve model stability and accuracy. In regression problems, the final prediction is obtained by averaging the outputs of individual trees, which helps reduce overfitting and enhances generalization performance (Breiman, 2001).

LR is a fundamental regression method that models the relationship between the dependent variable and one or more independent variables using a linear function. It is used to estimate continuous target values and serves as a baseline approach in predictive modeling (Raschka et al., 2022).

DT is a tree-based learning method that partitions the data space into smaller subsets through recursive splitting. It can be applied to both classification and regression tasks and is effective in capturing non-linear relationships between variables (Rokach & Maimon, 2015).

KNN is an instance-based learning method that predicts the value of a new observation based on the similarity to its K nearest neighbors in the dataset. It commonly uses distance metrics such as Euclidean distance to measure similarity between observations (Kelleher et al., 2015).

SVR is a regression method that aims to make predictions within a specified error margin (ε) while minimizing model complexity. It relies on support vectors and can capture non-linear relationships through the use of kernel functions (Géron, 2017).

GB is an ensemble learning method that builds a strong predictive model by sequentially training weak learners, typically decision trees. Each new model focuses on reducing the errors of the previous one, thereby gradually improving overall prediction performance (Géron, 2017).

To evaluate the risk of data leakage, the analyses were conducted under two different modeling scenarios. Model-1 (Full Model) includes all predictor variables, whereas Model-2 (Restricted Model) was constructed by excluding features that are highly correlated with the target variable.

In both scenarios, identical data preprocessing steps and modeling procedures were applied to ensure comparability. The dataset was split into training and test sets with an 80%–20% ratio, using a fixed random state (random_state = 42) to ensure reproducibility. During hyperparameter optimization, only the training data were used, and 10-fold cross-validation was applied to prevent data leakage and improve model generalizability.

Hyperparameter optimization was performed using the GridSearchCV method. For each algorithm, appropriate hyperparameter ranges were defined, and the optimal parameter combinations were selected based on the coefficient of determination(R^2). The best-performing models obtained through this process were then evaluated on the test dataset.

In the performance evaluation stage, five key metrics were considered to comprehensively assess the predictive capability of the models: the R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). These metrics jointly provide a comprehensive evaluation of model accuracy, error magnitude, and generalization performance.

R^2 is a fundamental performance metric in regression problems that measures how well the model explains the variability of the target variable and its ability to predict new observations (Zheng & Casari, 2018). The formula for R^2 is given in Eq. (1).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

- y_i: The actual (observed) value of the dependent variable for the i-th observation
- $\hat{y}_i$: The predicted value for the i-th observation
- $\bar{y}$: The mean of the observed values
- n: The total number of observations

RMSE is a commonly used performance metric for regression problems. It indicates the standard deviation of the prediction errors made by the model. The closer the RMSE value is to 0, the better the prediction performance of the model (Géron, 2017). The formula for RMSE is given in Eq. (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

- y_i: The actual value of the i-th observation
- $\hat{y}_i$: The predicted value of the i-th observation
- n: The total number of observations

MAE is an error metric commonly used when the goal is to reduce the influence of outliers. It represents the average of the absolute differences between the predicted values and the actual values (Géron, 2017; Kelleher et al., 2015). The formula for MAE is given in Eq. (3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (3)$$

MAE values range from $[0, \infty)$. A lower MAE indicates better model performance.

RAE is defined as the ratio of the absolute difference between the predicted and actual values to the deviation of the actual values from their mean. It is a normalized error metric used to evaluate the performance of numerical prediction (regression) models. Lower RAE values indicate better predictive accuracy. The formula for RAE is given in Eq. (4).

$$RAE = \frac{\sum_{i=1}^{n} | \hat{\boldsymbol{y}}_i - \boldsymbol{y}_i |}{\sum_{i=1}^{n} | \boldsymbol{y}_i - \bar{\boldsymbol{y}} |} \quad (4)$$

- $\hat{\mathbf{y}}_\mathbf{i}$**:** The value predicted by the model.
- $\mathbf{y}_\mathbf{i}$**:** The actual (true) value.
- $\bar{\mathbf{y}}$**:** The mean of the actual values.
- **n:** The number of observations.

RRSE is computed by taking the square root of the ratio between the model's total squared error and the total squared error of a baseline predictor (i.e., a predictor that always outputs the mean of the actual values in the training data). Although this metric is not particularly meaningful for classification tasks (categorical prediction), it represents a reasonable criterion for regression tasks where the goal is to predict numerical values (Bouckaert et al., 2018; Witten et al., 2017). The formula for RRSE is given in Eq. (5).

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5)$$

- $\hat{y}_i$: The value predicted by the model.
- y_i: The actual (true) value.
- $\bar{y}$: The mean of the actual values.
- n: The number of observations.

E. Feature Importance Analysis

In this study, the SHAP method was used to evaluate feature importance. SHAP values quantify the marginal contribution of each feature by considering all possible feature combinations, thereby enabling a clearer interpretation of model outputs (Lundberg & Lee, 2017). This approach allows for the identification of the most influential variables in predicting career satisfaction and provides a more transparent understanding of the model's decision-making process.

F. Ethical Issues

The dataset used in this study is openly accessible, and all participant information has been anonymized. Therefore, the study does not require approval from an institutional ethics committee. All analyses were conducted solely for scientific and research purposes.

III. RESULTS AND DISCUSSIONS

In this study, the selected machine learning algorithms were applied to the dataset to predict career satisfaction, and their performances were evaluated using a 10-fold cross-validation approach. During the modeling process, LR, DT, KNN, SVR, RF, and GB algorithms were employed. The performance of each algorithm was assessed using R^2, MAE, RMSE, RAE, and RRSE metrics.

In the initial stage, the algorithms were trained and evaluated using their default parameter settings. The performance results obtained at this stage are presented in Table 3.

Table 3. 10-fold cross-validation performance results obtained using all features.

Model	R^2	MAE	RMSE	RAE	RRSE
RF	0.996048	0.018275	0.073081	0.015794	0.052848
GB	0.994686	0.026738	0.088709	0.023103	0.063965
DT	0.992141	0.015000	0.076344	0.012803	0.055018
SVR	0.984608	0.108973	0.159715	0.096421	0.118620
LR	0.982776	0.117621	0.176218	0.103665	0.129782
KNN	0.976749	0.061000	0.193062	0.054938	0.144211

As shown in Table 3, all algorithms produced relatively high performance values when all features were included in the model. The highest explanatory power was observed in the RF model (R^2 = 0.996048), followed by GB (R^2 = 0.994686) and DT (R^2 = 0.992141). In contrast, the lowest R^2 value was obtained from the KNN model (R^2 = 0.976749). Numerous studies in literature also emphasize that the Random Forest algorithm demonstrates strong performance, particularly in modeling complex, multivariate structures, with high accuracy and low error rates (Breiman, 2001; Liaw & Wiener, 2002; Megersa et al., 2025; Nguyen et al., 2025).

Following the 10-fold cross-validation results presented in Table 3, an additional analysis was conducted to further evaluate the generalizability of the models. In this context, the dataset was split into training and test sets with an 80%–20% ratio, and all models were retrained using their default hyperparameter settings.

After the training process, the performance of the models on both the training and test datasets was calculated and compared using R^2, MAE, RMSE, RAE, and RRSE metrics. The results are presented in Table 4.

Table 4. Comparison of training and test performance results using all features.

Model	Train R^2	Test R^2	Train MAE	Test MAE	Train RMSE	Test RMSE	Train RAE	Test RAE	Train RRSE	Test RRSE
RF	0.9997	0.9898	0.0055	0.0305	0.0244	0.1405	0.0048	0.0258	0.0178	0.1010
KNN	0.9871	0.9793	0.0431	0.0550	0.1553	0.2000	0.0373	0.0465	0.1137	0.1439
GB	0.9999	0.9884	0.0062	0.0379	0.0130	0.1496	0.0054	0.0321	0.0095	0.1076
SVR	0.9906	0.9888	0.0951	0.1079	0.1325	0.1471	0.0824	0.0913	0.0970	0.1058
LR	0.9864	0.9805	0.1100	0.1173	0.1596	0.1940	0.0953	0.0992	0.1168	0.1395
DT	1.0000	0.9741	0.0000	0.0500	0.0000	0.2236	0.0000	0.0423	0.0000	0.1609

As shown in Table 4, all models achieved very high performance on the training data. In particular, the DT model produced an R^2 value of 1.00 with zero error on the training set, indicating a perfect fit to the training data. Similarly, RF and GB models also yielded near-perfect results on the training data. However, when the test performance is considered, slight decreases in performance can be observed across all models. This suggests that some models, particularly DT, may be prone to overfitting.

Overall, when all features are included, the models demonstrate strong predictive performance on both training and test datasets, although differences between training and test results highlight the importance of evaluating generalization capability.

Following the training and test performance results presented in Table 4, a correlation analysis was conducted to examine the extent to which model performance may be influenced by relationships among variables in the dataset. In this context, a correlation matrix was computed to reveal linear relationships among all variables, and the results were visualized using a heatmap. In addition, variables showing high correlation with the target variable (Career_Satisfaction) were identified to assess potential data leakage risks. The resulting correlation structure is presented in Figure 1.

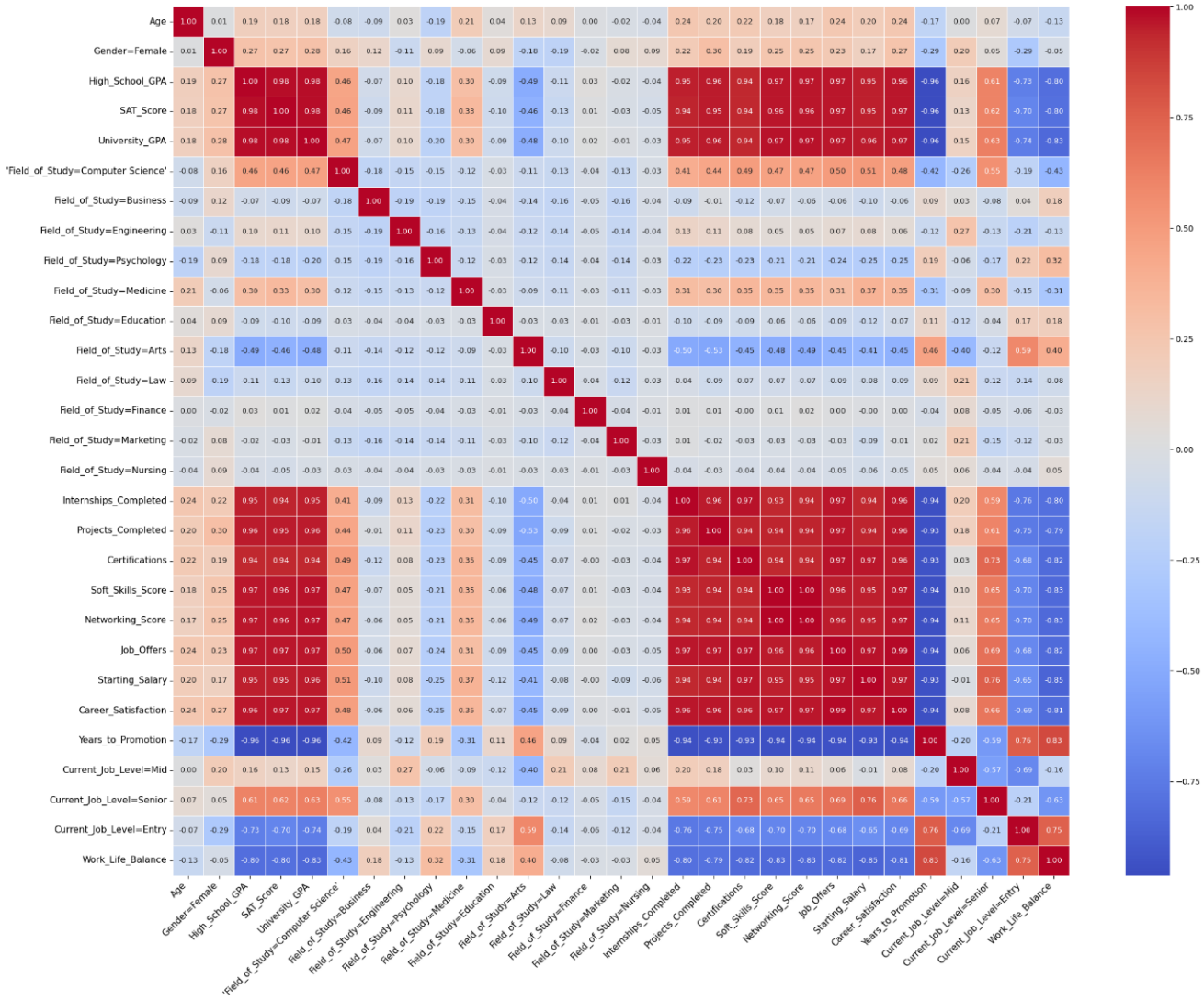


Figure 1. Correlation structure among features (heatmap).

As shown in Figure 1, some features exhibit relatively high correlations with the target variable (Career_Satisfaction). Based on the correlation analysis, these strong associations may indicate a potential risk of data leakage or overly optimistic model performance. To address this concern and to provide a more realistic evaluation of predictive capability, a restricted feature set was constructed by excluding variables that are conceptually close to the target variable or highly associated with it. In this context, Job_Offers, Soft_Skills_Score, Networking_Score, SAT_Score, Starting_Salary, University_GPA, Projects_Completed, High_School_GPA, Internships_Completed, Certifications, Work_Life_Balance, and Years_to_Promotion were excluded from the analysis.

As a result, a "Restricted Feature Set" consisting of 16 features was obtained. Subsequently, the same machine learning algorithms were re-applied to this dataset using a 10-fold cross-validation procedure, and model performances were evaluated using R^2, MAE, RMSE, RAE, and RRSE metrics. The results are presented in Table 5.

Table 5. 10-fold cross-validation performance results obtained using the restricted feature set.

Model	R^2	MAE	RMSE	RAE	RRSE
RF	0.8946	0.2329	0.4276	0.2080	0.3189
DT	0.8774	0.2335	0.4556	0.2083	0.3400
GB	0.8585	0.3621	0.4960	0.3232	0.3702
KNN	0.8380	0.3265	0.5342	0.2920	0.3968
SVR	0.8261	0.3668	0.5509	0.3272	0.4103
LR	0.8165	0.4683	0.5683	0.4170	0.4237

As shown in Table 5, model performance generally decreased when the restricted feature set was used. This reduction suggests that the excluded variables had a substantial contribution to the predictive performance of the models. Among the models, RF achieved the highest performance (R^2 = 0.8946), followed by DT and GB. In contrast, LR, KNN, and SVR exhibited relatively lower performance levels.

Following the cross-validation results presented in Table 5, an additional analysis was conducted to further evaluate the generalizability of the models using the restricted feature set. In this context, the dataset was split into training and test sets with an 80%–20% ratio, and the models were trained using their default hyperparameter settings. The results are presented in Table 6.

Table 6. Comparison of training and test performance results using the restricted feature set.

Model	Train R^2	Test R^2	Train MAE	Test MAE	Train RMSE	Test RMSE	Train RAE	Test RAE	Train RRSE	Test RRSE
RF	0.9451	0.9181	0.1684	0.2375	0.3202	0.3978	0.1458	0.2008	0.2344	0.2862
DT	0.9502	0.8976	0.1444	0.2442	0.3049	0.4446	0.1250	0.2065	0.2232	0.3198
SVR	0.8588	0.8922	0.3409	0.3102	0.5134	0.4563	0.2952	0.2624	0.3758	0.3282
KNN	0.8906	0.8900	0.2700	0.2975	0.4519	0.4610	0.2338	0.2516	0.3308	0.3316
GB	0.9056	0.8744	0.3138	0.3426	0.4197	0.4926	0.2717	0.2897	0.3072	0.3543
LR	0.8341	0.8529	0.4626	0.4303	0.5564	0.5330	0.4005	0.3639	0.4073	0.3834

As shown in Table 6, the highest performance on the test dataset was achieved by the RF model (R^2 = 0.9181), followed by DT and SVR. In contrast, the LR model exhibited the lowest performance.

Following the training and test performance results presented in Table 6, a hyperparameter optimization process was conducted to further improve model performance. In this context, optimal hyperparameter combinations for each model were identified using the GridSearchCV method on the restricted feature set. During the optimization process, 5-fold cross-validation was applied, and model performance was evaluated based on the R^2 metric. The test performance results of the optimized models, along with the selected hyperparameter configurations, are presented in Table 7.

Table 7. Test performance results after hyperparameter optimization.

Model	R^2	MAE	RMSE	RAE	RRSE
GB	0.9213	0.2610	0.3898	0.2207	0.2804
RF	0.9191	0.2394	0.3951	0.2025	0.2842
KNN	0.9169	0.2340	0.4007	0.1979	0.2882
DT	0.9077	0.2267	0.4223	0.1917	0.3038
SVR	0.8614	0.2996	0.5173	0.2534	0.3721
LR	0.8529	0.4303	0.5330	0.3639	0.3834

As shown in Table 7, the highest performance after hyperparameter optimization was achieved by the GB model (R^2 = 0.9213), followed by RF and KNN. In contrast, the LR model exhibited the lowest performance. Overall, model performance improved after the optimization process, and tree-based methods continued to rank among the top-performing models.

In terms of error metrics, different models demonstrated varying strengths. Considering the MAE metric, the lowest error value was obtained by the DT model (MAE = 0.2267), followed by KNN (0.2340) and RF (0.2394). In contrast, the LR model showed the highest MAE value (0.4303).

A similar pattern is observed for RMSE, where the GB model achieved a relatively low value (RMSE = 0.3898). The RF (0.3951) and KNN (0.4007) models also produced competitive results. However, the DT model exhibited a higher RMSE (0.4223), suggesting the presence of larger prediction errors despite its low MAE. The LR model again showed the highest RMSE (0.5330).

When RAE is considered, the DT model achieved the lowest RAE value (0.1917), followed by KNN (0.1979) and RF (0.2025). The GB model produced a slightly higher RAE (0.2207), consistent with its MAE results. Similarly, the LR model demonstrated the highest RAE (0.3639), indicating weaker performance compared to other models.

In terms of RRSE, the GB model achieved the lowest value (0.2804), followed closely by RF (0.2842) and KNN (0.2882). The DT model showed a slightly higher RRSE (0.3038), while the LR model again exhibited the highest value (0.3834), confirming its comparatively weaker predictive performance.

Following the hyperparameter optimization results presented in Table 7, an explainable artificial intelligence (XAI) analysis was conducted to interpret the decision structure of the best-performing model. In this context, the optimized GB model was retrained using the restricted feature set, and the contribution of features to model predictions was analyzed using SHAP.

The relative impact and direction of each feature on the predictions were visualized using a SHAP summary plot, and the results are presented in Figure 2.

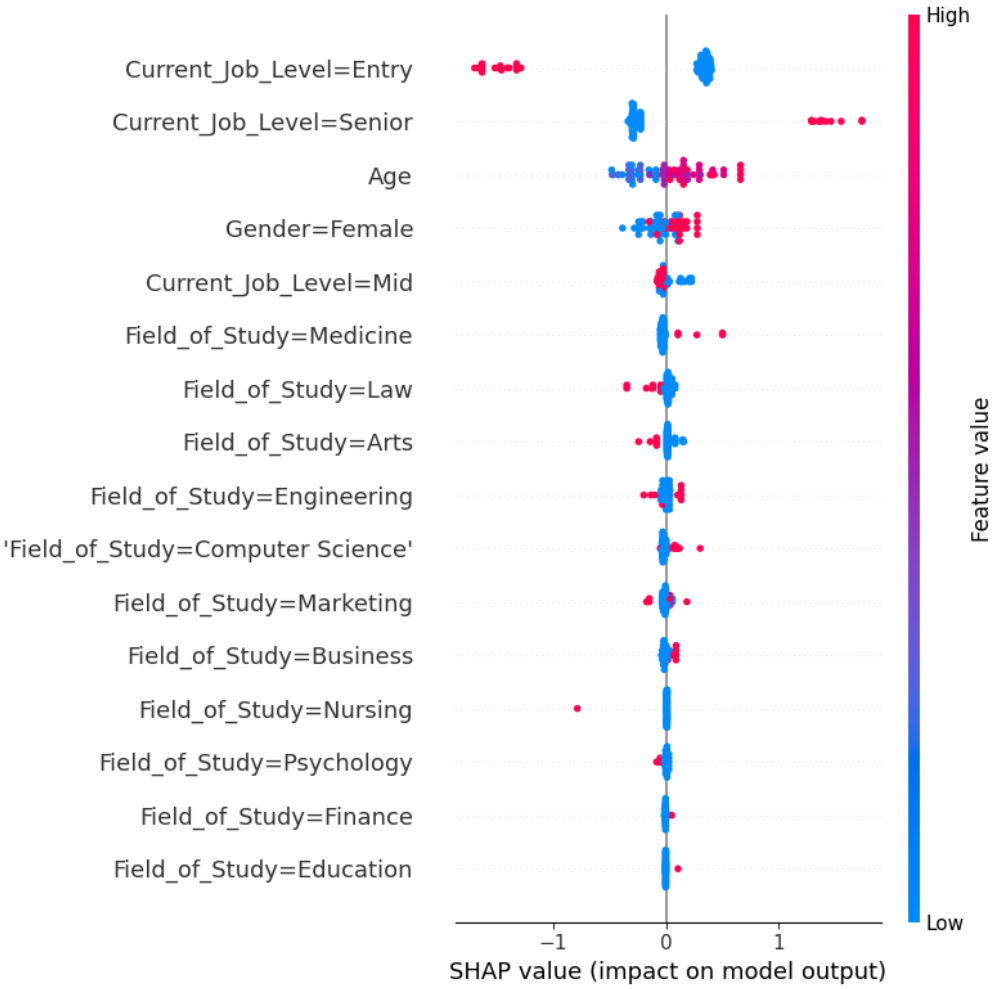


Figure 2. SHAP summary plot showing feature importance.

As shown in Figure 2, the most influential features in the model predictions include Current_Job_Level categories (particularly Entry and Senior) and Age. These features exhibit a wide range of SHAP values, indicating a stronger and more variable impact on the model output. This result is consistent with prior studies emphasizing the significant effects of demographic characteristics and job position on job satisfaction (Kılıç & Saygılı, 2021).

In contrast, features related to the field of study (Field_of_Study categories) generally show lower SHAP values, suggesting a more limited contribution to the model predictions. Additionally, both low and high values of certain features appear to influence the model output in different directions. Overall, the results suggest that model predictions are more strongly associated with individual and career-related features, while field-related attributes contribute relatively less to the predictive outcomes.

IV. CONCLUSION

In this study, it was aimed to predict career satisfaction using machine learning methods, and the performances of different regression algorithms were comparatively analyzed. In the initial stage, very high performance values were obtained when all features were included in the analysis. However, correlation analysis revealed that some features exhibited strong relationships with the target variable. This situation was considered as a potential indication of data leakage or overfitting risk.

In this context, in order to achieve a more realistic evaluation of model performance, a restricted feature set was constructed by excluding features that exhibited strong relationships with the target variable. Analyses conducted using the restricted dataset showed that model performance decreased to some extent, but provided more balanced and generalizable results.

The findings suggest the importance of considering data leakage risk in the evaluation of model performance and supporting the analysis with alternative modeling scenarios. When different modeling approaches were compared, tree-based methods (particularly RF and GB) tended to demonstrate higher performance than the other models. Following hyperparameter optimization, the GB model achieved the highest predictive performance.

In addition, the decision-making mechanism of the model was examined using SHAP analysis, and the results indicated that the most influential features were those related to individual characteristics and career level. This analysis enhances the interpretability of the model outputs and provides insight into which features are more influential in the prediction process.

This study has several limitations. First, the dataset used was obtained from secondary sources, and researcher control over data quality and measurement processes is limited. In addition, the absence of explicit temporal information in the dataset makes it difficult to establish a clear temporal separation between the independent variables and the target variable. This limitation motivated the development of alternative modeling approaches aimed at reducing the potential risk of data leakage. Furthermore, the findings are based on a single dataset, which may limit the generalizability of the results. Therefore, future studies are encouraged to validate the proposed models using real-world and longitudinal datasets.

In this context, future studies are recommended to utilize different datasets, extend the analysis with time-series or longitudinal data, and explore alternative modeling techniques. In addition, the use of different approaches to enhance model interpretability may contribute to a more comprehensive understanding of the results.

In conclusion, this study suggests that machine learning methods can be effective tools for predicting career satisfaction and provides a valuable foundation for future research in this field.

DECLARATIONS

Acknowledgements: The author declares that there are no acknowledgements to be made regarding any individual or institution.

Author Contributions: All parts of the study were carried out by the author.

Conflict of Interest: The author declares no conflict of interest.

Supporting Institutions: This research received no external funding.

Ethical Approval: This study used a publicly available and anonymized secondary dataset. The author did not directly collect data from human participants; therefore, institutional ethics committee approval was not required.

Plagiarism Statement: This article has been evaluated for plagiarism and no instances of plagiarism were detected.

Use of AI Tools: The author/authors declare that no Artificial Intelligence (AI) tools were used in the creation of this article.

REFERENCES

Arthur, M. B. (2008). Examining contemporary careers: A call for interdisciplinary inquiry. *Human Relations, 61*(2), 163–186. https://doi.org/10.1177/0018726707087783

Arthur, M. B., Hall, D. T., & Lawrence, B. S. (Eds.). (1989). *Handbook of career theory*. Cambridge University Press.

Beyeler, M. (2017). *Machine learning for OpenCV: A practical introduction to the world of machine learning and image processing using OpenCV and Python*. Packt Publishing.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2018). *WEKA manual for version 3.8.3*. University of Waikato.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Çakmak Otluoğlu, K. Ö. (2014). Kariyer bağlılığının kariyer başarısı üzerindeki etkisinin incelenmesi üzerine bir araştırma. *Yönetim ve Ekonomi Araştırmaları Dergisi, 12*(23), 349–363. https://izlik.org/JA58KX82RE

Celbiş, M. G., Wong, P. H., Kourtit, K., & Nijkamp, P. (2023). Job satisfaction and the "Great Resignation": An exploratory machine learning analysis. *Social Indicators Research, 170*, 1097–1118. https://doi.org/10.1007/s11205-023-03233-3

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal, 33*(1), 64–86. https://doi.org/10.5465/256352

Gunz, H., & Peiperl, M. A. (Eds.). (2007). *Handbook of career studies*. SAGE Publications.

Holgado-Apaza, L. A., Carpio-Vargas, E. E., Calderon-Vilca, H. D., Maquera-Ramirez, J., Ulloa-Gallardo, N. J., Acosta-Navarrete, M. S., Barrón-Adame, J. M., Quispe-Layme, M., Hidalgo-Pozzi, R., & Valles-Coral, M. (2023). Modeling job satisfaction of peruvian basic education teachers using machine learning techniques. *Applied Sciences, 13*(6),

Article 3945. https://doi.org/10.3390/app13063945
Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz, R. D., Jr. (1995). An empirical investigation of the predictors of executive career success. *Personnel Psychology, 48*(3), 485–519. https://doi.org/10.1111/j.1744-6570.1995.tb01767.x
Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT Press.
Kılıç, T., & Saygılı, İ. (2021). İş tatmini ile alt boyutları arasındaki ilişkiler ve demografik faktörlerin bu değişkenler üzerindeki etkisi: Görgül bir araştırma. *Çukurova Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 25*(1), 1–17. https://doi.org/10.51945/cuiibfd.588027
LeBlond, B., Christ, B. R., Ertman, B., Chapman, O., Pillai, R., & Perrin, P. B. (2025). Using machine learning to identify educational predictors of career and job satisfaction in adults with disabilities. *Disabilities, 5*(2), Article 56. https://doi.org/10.3390/disabilities5020056
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.
Lounsbury, J. W., Park, S.-H., Sundstrom, E., Williamson, J. M., & Pemberton, A. E. (2004). Personality, career satisfaction, and life satisfaction: Test of a directional model. *Journal of Career Assessment, 12*(4), 395–406. https://doi.org/10.1177/1069072704266658
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.
Megersa, A. G., Negash, F., Hailu, A., Melak, A., Assefa, A., Aseged, T., & Sinkie, S. (2025). Prediction of body weight in indigenous sheep using random forest regression, support vector regression, and classification and regression trees algorithms. *Veterinary and Animal Science, 30*, Article 100517. https://doi.org/10.1016/j.vas.2025.100517
Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media.
Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology, 58*(2), 367–408. https://doi.org/10.1111/j.1744-6570.2005.00515.x
Nguyen, D.-K., Nguyen, Q.-T., & Dinh, V.-P. (2025). Integrating artificial intelligence with kinetic studies for Cr(vı) removal using young durian fruit biochar: A random forest regressor approach. *RSC Advances, 15*(49), 42238–42253. https://doi.org/10.1039/d5ra05229g
Ramos-Pulido, S., Hernández-Gress, N., & Torres-Delgado, G. (2024). Exploring the relationship between career satisfaction and university learning using data science models. *Informatics, 11*(1), Article 6. https://doi.org/10.3390/informatics11010006
Raschka, S., Liu, Y. H., & Mirjalili, V. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python.* Packt Publishing.
Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: Theory and* applications (2nd ed.). World Scientific Publishing.
Shamim, A. (2025). *Education & career success* [Data set]. Kaggle. Retrieved July 19, 2025, from https://www.kaggle.com/datasets/adilshamim8/education-and-career-success
Urfalıoğlu, M., & Aşantuğrul, N. (2026). Veri madenciliği teknikleri kullanılarak lise öğrencilerinin mesleki kişilik profilleri ve kişilik özelliklerine dayalı kariyer yönlendirme: Makine öğrenmesi. *Bilişim Teknolojileri Dergisi, 19*(1), 11–24. https://doi.org/10.17671/gazibtd.1719617
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists.* O'Reilly Media.