OKU Fen Bilimleri Enstitüsü Dergisi
8(3): 1100-1119, 2025

OKU Journal of The Institute of Science and Technology, 8(3): 1100-1119, 2025

Osmaniye Korkut Ata Üniversitesi
Fen Bilimleri Enstitüsü
Dergisi

Osmaniye Korkut Ata University
Journal of The Institute of Science
and Technology

Performance Comparison of Neural Networks: A Case of Data Scientists' Job Change Prediction

Aslı ÖRGERİM[1*], Tuğba TUNÇ ABUBAKAR[2], Mahmut TOKMAK[3]

[1,2,3]Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay Applied Technology and Business High School, Management Information Systems Department, Burdur, Turkiye

[1]https://orcid.org/0000-0001-7785-6200
[2]https://orcid.org/0000-0002-5447-2391
[3]https://orcid.org/0000-0003-0632-4308
*Corresponding author: agode@mehmetakif.edu.tr

Research Article

Article History:
Received: 10.05.2024
Accepted: 04.03.2025
Published online: 16.06.2025

Keywords:
Artificial neural network
XGBoost
Classification
Data scientists' job change request

ABSTRACT

In today's world, the era of big data, companies in every sector have to deal with huge amounts of data generated. Such data must be processed, analyzed, and interpreted to be used in making business decisions. Businesses employ data scientists for this purpose. These people have great costs to businesses. For this reason, it is a significant issue for businesses to predict the employee who intends to change jobs in people working as data scientists in enterprises. In this study; the job change thoughts of data scientists were predicted by artificial neural networks. Data cleaning, missing data completion with linear regression-based iterativelmputer method, data balancing with SMOTE (Synthetic Minority Oversampling Technique) algorithm, data normalization with standard scaler method were performed on the dataset used, respectively. The dataset was then trained with a multilayer perceptron algorithm and a deep neural network model. The trained models were tested and an accuracy of 84.2% was obtained with the multilayer perceptron algorithm and 87.5% with the deep neural network model. To compare the performance of artificial neural network models, analyses were performed with the frequently used Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, Extra Trees, Gradient Boosting, and XGBoost algorithms. As a result of these tests, an accuracy of 91.1% was obtained with the XGBoost algorithm and performance metrics were presented.

Sinir Ağlarının Performans Karşılaştırması: Veri Bilimcilerinin İş Değişikliği Tahmini Örneği

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 10.05.2024
Kabul tarihi: 04.03.2025
Online Yayınlanma: 16.06.2025

Anahtar Kelimeler:
Yapay sinir ağları
XGBoost
Sınıflandırma
Veri bilimcilerinin iş değişikliği isteği

ÖZ

Büyük veri dönemi olarak adlandırılan günümüz dünyasında, her sektördeki firmaların üretilen çok büyük miktarda veriyle uğraşması gerekmektedir. Bu tür verilerin iş kararları vermede kullanılabilmesi için işlenmesi, analiz edilmesi, yorumlanması gerekir. İşletmeler bu amaçla veri bilimcileri istihdam etmektedirler. Bu kişilerin işletmelere büyük maliyetleri bulunmaktadır. Bu nedenle işletmelerde veri bilimcisi olarak çalışan kişilerde iş değişikliği düşüncesi olan çalışanın tahmin edilmesi işletmeler açısından çok önemli bir konudur. Bu çalışmada; veri bilimcilerin iş değişikliği düşüncelerinin yapay sinir ağları ile tahmini yapılmıştır. Kullanılan veri seti üzerinde, sırasıyla, veri temizleme, lineer regresyon tabanlı iterativimputer yöntemiyle eksik veri tamamlama, SMOTE algoritmasıyla veri dengeleme, standart scaler metodu ile veri normalizasyonu yapılmıştır. Daha sonra veri seti çok katmanlı algılayıcı algoritması ve derin sinir ağı modeliyle eğitilmiştir. Eğitilen modeller test edilip çok katmanlı algılayıcı algoritması ile %84.2 derin sinir ağı modeli ile %87.5 doğruluk değeri elde edilmiştir. Yapay sinir ağları modellerinin performansını karşılaştırabilmek amacıyla sıkça kullanılan Naive Bayes, Destek Vektör Makineleri, Karar Ağaçları, Rastgele Ormanlar, Ekstra Ağaçlar ile gradyan artırma modellerinden Gradient

Boosting ve XGBoost algoritmaları ile analizler yapılmıştır. Bu testler sonucunda ise XGBoost algoritmasıyla %91.1 doğruluk değeri elde edilmiş ve performans metrikleri ortaya konmuştur.

To Cite: Örgerim A., Tunç Abubakar T., Tokmak M. Performance Comparison of Neural Networks: A Case of Data Scientists' Job Change Prediction. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2025; 8(3): 1100-1119.

1. Introduction

Although the amount of data generated worldwide daily constantly changes, it has been increasing rapidly in recent years. Technological advances such as the Internet, mobile devices, sensors, and social media are effective in increasing data generation and storage capacity. In numerical terms, approximately 500 hours of video, 500 million tweets, 4 million hours of YouTube video, and data of 4.3 billion internet users are produced worldwide in a minute. This means that approximately 2.5 quintillion (2.5 x 10^18) bytes of data are generated daily (Domo, 2023; Techjury, 2023).

In parallel with this, the rapid development of technology and digitalization has increased the amount of data obtained by businesses. These data provide information about customer preferences, product performance, market trends, and many other issues. However, these data need to be analyzed correctly in order to make them valuable for businesses. This is where data science and artificial intelligence (Artificial Intelligence: AI) come into play. In today's rapidly digitalizing world, data science and artificial intelligence are becoming increasingly important. The two disciplines complement each other, helping businesses gain valuable insights from large datasets and use this information to make strategic decisions (Górriz et al., 2020). With the advent of artificial intelligence introduced during Industry 4.0, organizations have gained the ability to simulate various variables in advance and make more accurate decisions as a result. Thanks to this technological advancement, which provides the ability to take action a few steps ahead, organizations make the future predictable (Karagöz, 2022).

Artificial intelligence is defined as giving a computers or machines the ability to think, reason, generalize, make decisions, and produce solutions that typically require human intelligence can do. The development of artificial intelligence has provided great benefits to the economy, promoted social development, and contributed to almost every field (Zhang and Lu, 2021). In addition, artificial intelligence has changed many aspects of organizations, from production styles, supply chains, management styles, customer relations, and competition styles to strategic planning (Karagöz, 2022). Expert systems, robotics, natural language processing, speech understanding, computer vision, computer science, control theory, biology, psychology, philosophy, and mathematics constitute the scope and foundations of artificial intelligence. Artificial intelligence technologies can be listed as machine learning, artificial neural networks (ANN), deep learning, fuzzy logic, genetic algorithms and expert systems (Bingöl et al., 2020; Metlek, 2021). ANN, one of these artificial intelligence techniques, is a computational model inspired by biological neural networks that process information in the human brain. Neural networks studies can learn and associate large datasets obtained from experiments, simulations and the real world. A trained neural network serves as an analytical tool for qualified predictions of actual results. Effective neural network models for their training and validation exist and continue to be developed, while at the same time achieving high accuracy scores in their predictions (Asteris and Mokos, 2020).

Data science helps businesses make strategic decisions to increase their productivity, reduce costs and increase profitability by analyzing trends and patterns in large datasets. Data science is a field that involves the collection, storage, processing, analysis and interpretation of data using disciplines such as statistics, data mining, artificial intelligence and machine learning. Data scientists work on large datasets of businesses by analyzing data using these disciplines. By collecting, storing, processing, analyzing, and interpreting data, data scientists provide important information used in the decision-making processes of businesses. Data scientists develop strategies to increase the growth potential of businesses by analyzing customer behavior, market trends, product performance and many other issues. Data scientists use programming languages such as Python, R, SAS, and SQL to analyze data. They use tools such as Apache Hadoop, Apache Spark, Apache Cassandra, and NoSQL databases to work on large datasets. In addition, data scientists utilize artificial intelligence technologies to process large amounts of unstructured information, and automate or solve complex tasks (Altunışık, 2015; Baumeister et al., 2020; Górriz et al., 2020; Mohamed et al., 2020).

Data scientists help businesses make strategic decisions to increase their productivity, reduce their costs and increase their profitability by working on large datasets. Data scientists are important employees for companies. Hiring, training and developing data scientists represent a huge investment for companies. Businesses spend great efforts to hire data scientists and these efforts are often very costly. However, the departure of the data scientist may cause this investment to be wasted. The intention to change jobs of data scientists working in the enterprise is defined as "the movement of workers in and out of employment in a given company" (Sexton et al., 2005). The reasons for job change of data scientists can be divided into organizational, individual, and environmental reasons. Organizational job change can be given as an example of data scientists being dissatisfied with their work environment (Aras, 2016). Individual reasons include the fact that the intention to change jobs increases as the education level of data scientists increases (Yener, 2016). The low level of education and economic development of the city or country can be given as examples for environmental reasons (Varol, 2017). Data scientist turnover can be costly for businesses in several different ways. Typically, these companies receive multiple candidate registrations for their training programs. Therefore, they want to know which of these candidates, after the training period, might looking for a new job in other companies. This forecast is extremely useful because it helps reduce the cost and time, improves training quality, and aids in planning courses and categorizing candidates (Tran and Nguyen, 2021). Businesses may face several costs due to the departure of data scientists, such as re-hiring costs, data loss, and costs arising from project delays. In addition to these costs, the turnover or resignation of data scientists leads to a loss of productivity. On the other hand, as productivity decreases, corporate profit also decreases. Reducing the rate of turnover in businesses is vital for organizational success. It is of great importance for companies to predict whether data scientists have such an intention or not and to ensure the continuity of data scientists by human resources through reward, incentive, etc. methods. Therefore, businesses want to predict the intention of data scientists to change jobs to minimize the risk of data scientists leaving their jobs.

For this purpose, various studies have been conducted in the literature: Kyalkond et al. (2022)'s study, Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), Random Forests (RF) methods were used. As a result of the study, 86.2%, 88.1%, 86.9%, 89.5% accuracy were obtained from the methods respectively. In this direction, the RF algorithm gave the highest accuracy rate. Conlon (2021) RF, AVG Blender, and Light Gradient Boosted Trees Classifier machine learning methods with Early Stopping. On average, they obtained an accuracy value of over 78%. In his study, he revealed that the characteristics that have the most influence on whether a data scientist wants to change his job are; city development index, company size, and company type. In their study, Tran and Nguyen (2021) used DT, Naive Bayes (NB), KNN, SVM, RF, XGBoost and LightGBM (LGBM) methods to predict whether employees intend to change jobs. As a result of the study, they obtained the best accuracy of 80% with the LGBM algorithm. There are studies on different datasets in the literature. Dutta and Bandyopadhyay (2020) used a feed-forward neural network and a 10-fold cross-validation procedure to predetermine the employee's job change process. Oliveira et al. (2019), based on a dataset containing time series data on 3952 managers' e-mail communications for 12 months and using a recurrent neural network technique, predicted the rate of job change. As a result of the study, 80% accuracy was obtained.

The difference of this study from the literature lies in use of boosting machine learning algorithms as well as known machine learning algorithms and their detailed comparison. In addition, the use of the SMOTE method and iterativelmputer methods distinguishes the study from the existing literature.

In this study, a prediction was made to determine whether data scientists, who work in a very important position for businesses, are changing jobs or looking for a new job. Multi-Layer Perceptron (MLP) and Deep Neural Network (DNN) were used as prediction models. To compare the performance metrics obtained in the models used, analyses were made with NB, SVM, Decision Tree (DT), RF, Extra Trees (ET), which are frequently used in the literature, and Gradient Boosting and XGBoost algorithms, which have increased in popularity in recent years. The proposed study aims to contribute to the reduction of labor and cost losses in the entire process of recruiting, training, and integrating data scientists into positions of sensitive importance for businesses. In addition, it aims to compare the performances of MLP and DNN algorithms with existing artificial intelligence methods.

2. Material and Method

2.1. Dataset

This study used the Kaggle dataset "HR Analytics: Data Scientists' Job Change" Kaggle dataset (Kaggle, 2022). This dataset is open access and no permission is required to use it. This dataset comprises 19.158 data with 14 features, including the target feature. In the target feature, a value of 0 indicates that data scientists do not intend to change jobs, while a value of 1 indicates that data scientists are looking to change jobs. The dataset in question is designed for studies to predict the probability of a data scientist changing to a new job. The features in the dataset and their descriptions are shown in Table 1.

Table 1. Features of the dataset

Features	Description
enrollee_id	Data scientist identification number
city	City code
city_development _index	City development index
gender	Gender distribution of data scientists
relevent_experience	Relevant experience of data scientists
enrolled_university	Type of university program, if any, in which the data scientists are enrolled
education_level	Education level of data scientists
major_discipline	Training ground for data scientists
experience	Total experience of data scientists (in years)
company_size	Number of data scientists in the current employer's company
company_type	Current company type
Last_new_job	Difference between previous job and current job (in years)
training_hours	Completed training hours
target	Output 0-Not seeking a job change, 1-Seeking a job change

When the dataset is analyzed, it is observed that there is up to 38% missing data for some features. The features, numbers, and percentages of missing data are given in Table 2.

Table 2. Number and percentage of features with missing data

Features	Number of Missing Data	Percentage of Missing Data
enrollee_id	0	0.00
city	0	0.00
city_development_index	0	0.00
gender	4508	23.53
relevent_experience	0	0.00
enrolled_university	386	2.01
education_level	460	2.40
major_discipline	2813	14.68
experience	65	0.34
company_size	7409	38.67
company_type	6140	32.05
last_new_job	423	2.21
training_hours	0	0.00
target	0	0.00

Of the 19.158 individuals in the dataset corresponding to the information obtained from the data scientists, 4.777 indicated that they were looking for a new job, and 14.381 indicated that they were not looking for a job. The number of people in the target feature is shown in Figure 1.

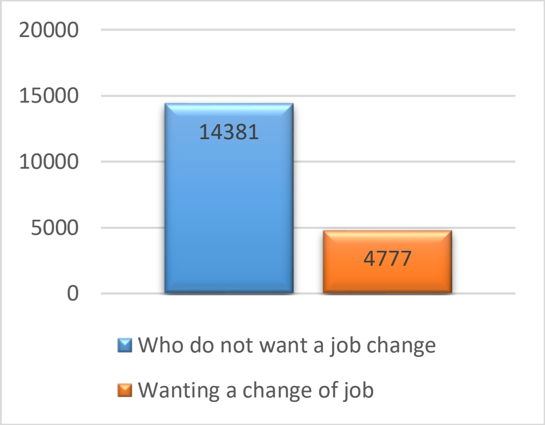


Figure 1. Numbers of target class variable

2.2.Dataset Pre-Processing

Data preprocessing, one of the processes applied to datasets, has critical importance in data analysis and machine learning projects. Data preprocessing refers to a series of processes to improve the quality and consistency of data (Ramkumar et al., 2023). Data preprocessing starts with data collection and cleaning processes. It checks the accuracy, completeness, relevance, and consistency of the data and performs the necessary preprocessing operations to obtain meaningful data. Data preprocessing steps such as data cleaning, data transformation, data scaling, feature selection, and data merging are important to obtain more accurate results for data analysis and the training of machine learning models (Ajakwe et al., 2024). In this study, firstly, data cleaning was performed from the data preprocessing steps, and the "enrolle_id" column, which expresses the number of the data scientists, was deleted. Secondly, since the dataset contains 38% missing data when all fields are taken into account, as shown in Table 2, the fields containing missing data were completed with the Linear Regression-based IterativeImputer method. IterativeImputer is a machine learning technique used to complete missing data, provided by the very popular scikit-learn Python software package (Goh et al., 2023). This technique creates a model using techniques such as LR, DT, KNN, and SVM to predict the missing data and then completes the missing data using this model. IterativeImputer takes into account relationships with other features when estimating missing data. It estimates missing values as a function of other features using a model such as linear regression and updates these estimates in an iterative process. In this way, the filled values are compatible with the natural structure of the dataset. It offers a more advanced approach compared to simple methods (e.g. filling with mean or median). Simple methods often reduce the variance of the feature with missing data, resulting in a loss of information. IterativeImputer overcomes this disadvantage and produces more consistent results. In the third step, categorical features are converted into numerical representations, and the values of different numerical features are converted to the interval [0,1] using the StandardScaler function. In the fourth step, it was observed that the data was unbalanced when looking at the target variable of the dataset as shown in Figure 1. The SMOTE algorithm was used to eliminate data imbalance. SMOTE algorithm is a synthetic sampling method used in datasets with unbalanced class distributions. The SMOTE algorithm selects the samples around the samples in the minority class in the dataset and creates synthetic samples by randomly combining them. This process is performed based on the measured distances between samples. The SMOTE algorithm aims to reduce the class imbalance in the dataset by increasing the number of samples in the minority class in the dataset (Kartal and Özen, 2017). The number of target class variables in the dataset after the class imbalance is removed is given in Figure 2.

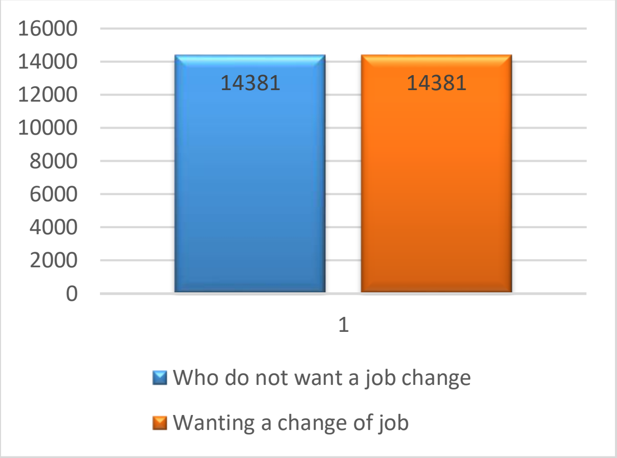


Figure 2. Target class variable numbers after the dataset class ımbalance is removed

2.3. Artificial Neural Networks

ANN is a subfield of artificial intelligence that models and imitates the information processing and learning abilities of the human brain. The main goal of ANN is to make decisions by thinking like a human and to be able to learn on its own. ANN is inspired by the human brain as a result of the mathematical modeling of the learning process (Sadanand and Bhosale, 2023).

Due to the inadequacy of hardware developments, the number of hidden layers and nodes in ANNs increased in the early 2000s, but they were not used. However, afterward, with the GPU and other hardware developments, ANNs with many hidden layers have started to be used again since their computational costs have been reduced (Şeker et al., 2017). Today, ANN is an important technology used in many application areas. Especially the developments in deep learning and neural network architectures make ANN models more and more successful. Research areas in this field have been increasing in recent years. The use of ANN models in areas such as image processing, natural language processing, robotics, gaming and music is becoming widespread. In addition, the use of ANN models in critical areas such as health, finance and security is also increasing (LeCun et al., 2015). ANN applications are used in data interpretation, data filtering, data association, prediction and classification problems (Yılmaz, 2020). In general, ANNs are used for face recognition, license plate recognition (Göde and Doğan, 2023) voice recognition, image processing, system modeling, weather forecasting, fault detection, disease diagnosis, robotics, finance, banking, space sciences, and automotive fields can be given as examples (Öztürk and Şahin, 2018).

ANN mimics the human biological nervous system. Neurons in the biological nervous system consist of four parts: nucleus, dendrite, axon and connections. Each nerve cell receives the incoming information through the dendrite and transmits it to the nucleus. This information is collected in the soma. The collected information is transmitted to the axon and processed here. The processed information is transferred to the other side of the neuron. The task of the connections is to transmit the processed information to other neurons in the relationship (Zhang et al., 2019). Figure 3 shows the biological nerve cell structure.

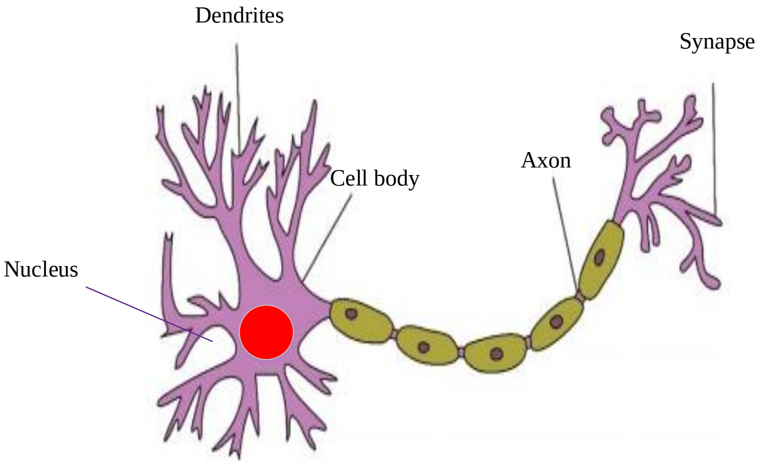


Figure 3. Biological nerve cell (neuron) structure (Fiorelli et al., 2015)

The corresponding elements of biological nerve cell and ANN cell are given in Table 3.

Table 3. Corresponding elements of biological nerve Cell and ANN cell (Tang et al., 2019; Güner, 2021)

Biological nerve cell elements	ANN elements
Dendrite	Aggregation function
Cell Body	Activation function
Axon	Output value
Neuron	Input value
Synapse	Weight

ANN has 5 basic elements. (1) Inputs; information coming from outside or from a different cell. (2) Weights; the importance and effect of the incoming information for the neuron is expressed by weights. Incoming information is first multiplied by the weight. If the weight value is small, it is insignificant, if it is large, it is important. If the weight value is zero, it has no effect. (3) Summation Function; After the information is multiplied by the weights, it is summed in the summation function. Thus, the net input of the relevant cell is calculated. (4) Activation Function; processes the net input information and produces the output value corresponding to this information. (5) Outputs; transfer the output values produced in the activation function to the outside (Maind and Wankar, 2014). Figure 4 shows the ANN structure.

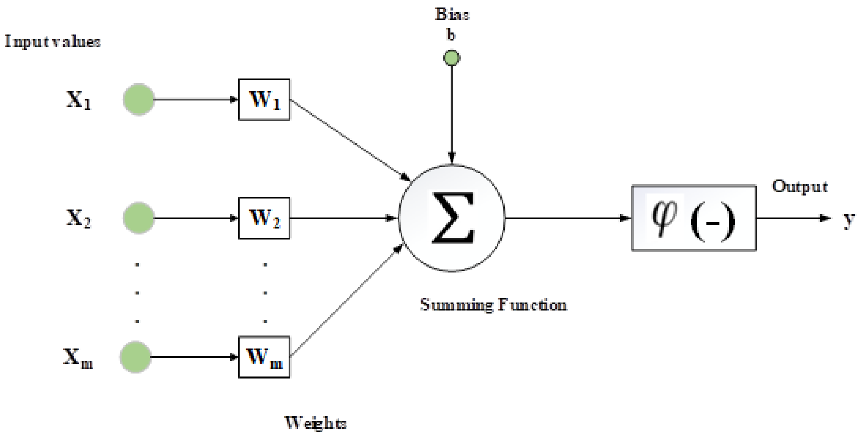


Figure 4. ANN structure (Martínez-Álvarez et al., 2015)

Xi values represent the input elements. Xi input values are multiplied by Wi. The b value is added to the result obtained as a result of multiplication in the summation function. In the next step, the activation function is applied. As a result, y output value is obtained (Oliveira et al., 2017).

2.4. Artificial Neural Network Models

ANN models differ according to architecture, connection direction, and flow direction of network signals. In this context, ANN models are named single-layer perceptrons, multilayer perceptrons, feed-forward networks, and feedback networks (Öztürk and Şahin, 2018). In this section of the paper, the models used in the study are explained.

2.4.1. Multilayer Perceptrons

MLPs consist of the input layer, hidden layer and output layer. MLPs update the error level until the difference between the expected output value and the actual output value is minimized (Zhang et al., 2023). Figure 5 shows the MLP model.

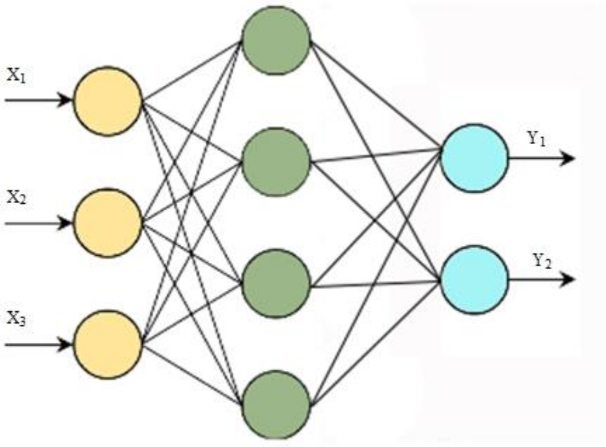


Figure 5. MLP model (Zhang et al., 2023)

2.4.2. Deep Neural Networks

Today, the increase in the amount of data, the diversity of variables, and the studies to reveal the complex relationships between these variables have increased the interest in ANN-based deep learning approach. DNNs are defined as a wider and deeper form of ANNs. These networks contain at least two hidden layers and the number of neurons in these layers is high. The deep learning approach is used in many fields such as speech recognition, computer vision, and natural language processing. There are many different deep network architectures such as deep neural networks (DNN), deep belief network (DBN), convolutional neural networks (CNN), and recurrent neural networks (RNN) (Gazel and Bati, 2019).

DNNs consist of multiple levels of non-linear processing with numerous intermediate layers. DNNs are mostly used to deal with unstructured and unlabeled data. Data enters from the input layer and is processed through the layers. At each layer, the input data is multiplied by weights and passed to an activation function. This is done by optimizing the weights. Optimized weights allow the network to provide more accurate results. Today, DNNs are used in many areas such as computer image and video analysis, natural language processing, speech recognition, and games (Lee, 2021). Figure 6 shows the DNN network structure.

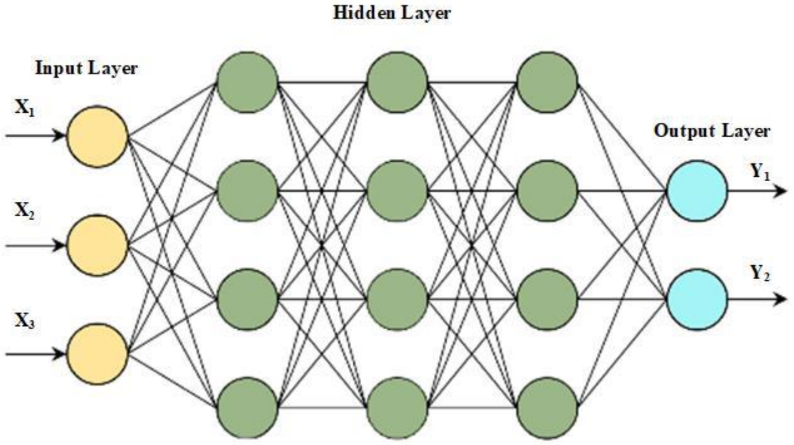


Figure 6. *DNN network structure (Tekerek and Bay, 2019)*

2.5. Performance Evaluation

The dataset consisting of 28.762 samples obtained after the data preprocessing steps was divided into 80% training and 20% testing. Accuracy, precision, sensitivity, and F1-scores were calculated to evaluate the training and test performance of the models. At the same time, the results obtained in the target class prediction with the Confusion Matrix, which contains the numerical data of the actual and predicted classes used to evaluate the performance of classifiers in classification problems, are expressed both numerically and visually (Çavuşoğlu and Kaçar, 2019). The general expression of the confusion matrix is given in Table 4.

Table 4. Confusion matrix

Estimated values	Actual values	
	Negative	Positive
Negative	True Negative (TN)	False Negative (FN)
Positive	False Positive (FP)	True Positive (TP)

Performance evaluation metrics used in the study are explained below (Dutta and Bandyopadhyay, 2020; Nkongolo and Tokmak, 2023):

- Accuracy: It is calculated as the ratio of the correctly predicted areas in the model to the total dataset. The accuracy formula is shown in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

- Precision: The accuracy of positive predictions is found by the Precision formula. The precision formula is shown in Equation 2.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

- Sensitivity **(**Recall**):** The proportion of positive samples correctly detected by the model. The sensitivity formula is shown in Equation 3.

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

- F1-Score: When calculating the F1 score value, a new value is calculated by using the precision and sensitivity values together. Equation 4 shows the F1-Score formula.

$$F1\text{-}Score = 2 \ x \ \frac{precision \ x \ sensitivity}{precision+sensitivity} \tag{4}$$

The classification models proposed in this study were trained and tested on Google Colab Notebook platform using Python (Colab, 2021). This section describes the methodology of the study. The flowchart of the study is shown in Figure 7.

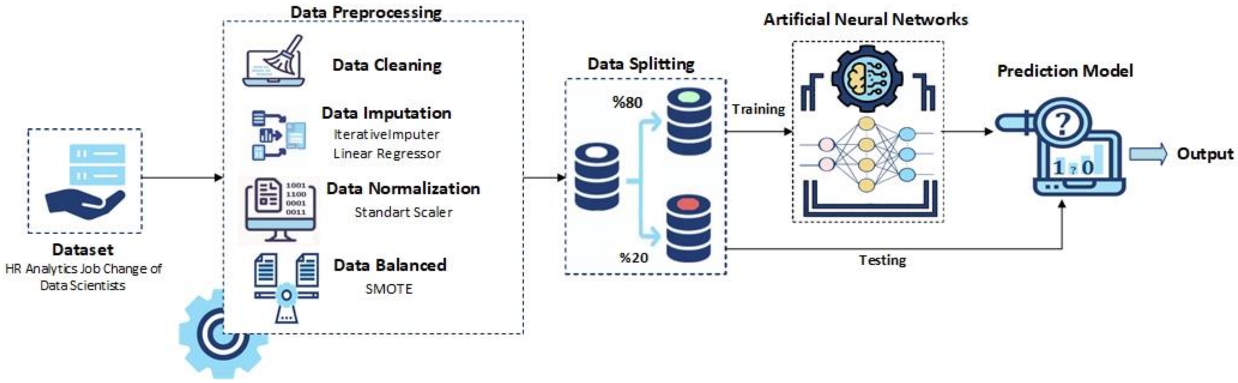


Figure 7. Proposed model

Among the artificial intelligence methods used in the study, Python sklearn library was used for NB, SVC, DT, MLP, RF, Gradient Boosting, ET, classifiers, xgboost library for XGBoost classifier and Tensorflow keras library for DNN classifier. The model parameters used to obtain the experimental results of the artificial intelligence methods employed in this study are presented in Table 5. "Default" parameters were adopted in NB, SVC, DT, ET, XGBoost algorithms. While determining these parameters, the content of the dataset, number of features, and data type were taken into consideration.

Table 5. Model hyperparameters

Model Name	Parameters
RF	max_depth=6, n_estimators=100, max_features=1
Gradient Boost	n_estimators=100, learning_rate=1,0, max_depth=1, random_state=0

MLP	max_iter=500, alpha=1e-4, solver="adam", activation='relu'
	random_state=40, n_iter_no_change=50, learning_rate_init=0,01, momentum=0,3

DNN	Hidden Layer activation='relu'; Output Layer activation='sigmoid';
	loss='binary_crossentropy'; optimiser='adam'; metrics=['accuracy']

3. Results and Discussion

In this study, Google Colab Notebook was used for training and testing the classification models. GPU runtime was selected in the notebook and the models were run with Python 3.10.11 version. The neural network models were trained by systematically increasing the number of hidden layers and the number of neurons in the hidden layers, and then the performances of different network models were tested. The performance metrics of the trained MLP and DNN models obtained as a result of the testing process are given in Table 5. In the test of the MLP network model with two hidden layers, each containing 1000 neurons, the highest accuracy was obtained with an 84.2% accuracy compared to other MLP models. Similarly, in the test of the DNN network model with two hidden layers and 250 neurons in each hidden layer, the highest accuracy was obtained with an 87.5% accuracy compared to other DNN models. The values obtained are shown in Table 6.

In order to compare the performances of the ANN methods used in this study, the dataset was trained with some artificial intelligence methods commonly used in the literature, and forecasting was performed. In order to make a comparison, the classical DT, RF, ET, NB, SVM algorithms and Gradient boosting XGBoost machine learning algorithms, which are among the gradient boosting models which have become popular in recent years, were selected. The performance metrics obtained are given in Table 6, the lowest accuracy performance was 69.2% with the NB algorithm and the highest accuracy performance was 91.1% with the XGBoost algorithm. Table 7 shows the performance comparison.

Table 6. MLP and DNN hidden layer numbers and performance metrics

Hidden Layer	Accuracy		Precision		Recall		F1- Score		Support
	MLP	DNN	MLP	DNN	MLP	DNN	MLP	DNN	
5-5	0.763	0.766	0.7667	0.77	0.7633	0.767	0.7622	0.765	5753
10-10	0.786	0.766	0.787	0.766	0.786	0.766	0.786	0.766	5753
50-50	0.793	0.844	0.794	0.844	0.793	0.844	0.793	0.844	5753
100-100	0.795	0.858	0.796	0.858	0.796	0.858	0.795	0.858	5753
250-250	0.808	**0.875**	0.815	0.875	0.809	0.874	0.807	0.874	5753
500-500	0.831	0.862	0.835	0.862	0.831	0.862	0.83	0.862	5753
1000-1000	**0.842**	0.861	0.845	0.861	0.842	0.861	0.842	0.861	5753
5-5-5	0.771	0.765	0.776	0.765	0.772	0.765	0.771	0.765	5753
10-10-10	0.789	0.775	0.79	0.777	0.789	0.776	0.789	0.775	5753
50-50-50	0.795	0.83	0.795	0.832	0.795	0.83	0.794	0.83	5753
100-100-100	0.801	0.854	0.801	0.854	0.8	0.854	0.8	0.854	5753
250-250-250	0.831	0.863	0.832	0.863	0.831	0.863	0.831	0.863	5753
500-500-500	0.839	0.858	0.839	0.858	0.839	0.858	0.839	0.858	5753
1000-1000-1000	0.84	0.862	0.842	0.862	0.841	0.862	0.84	0.862	5753

In Figure 8, the accuracy values obtained as a result of the test after the training according to the number of hidden layers and neurons of the neural network models are presented graphically.

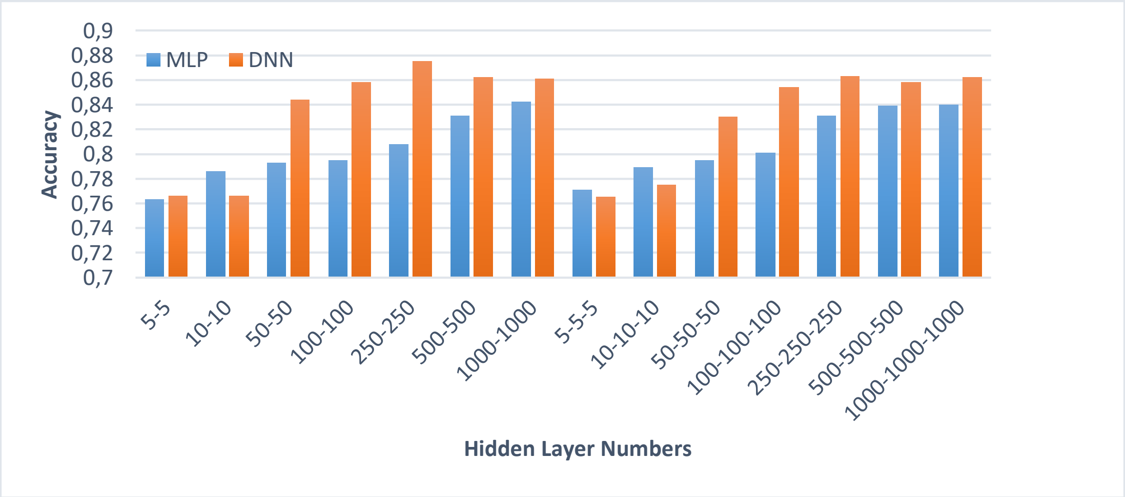
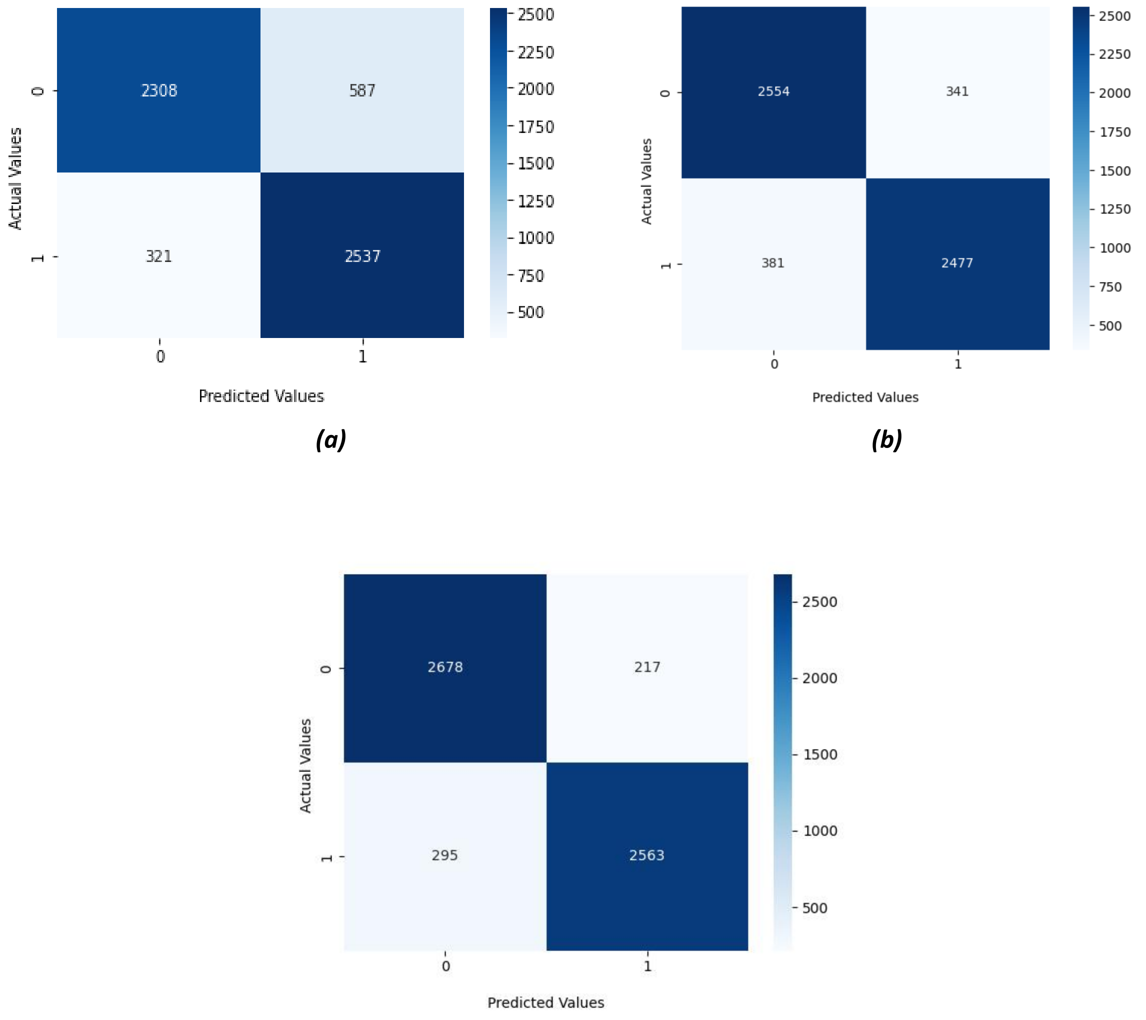


Figure 8. MLP-DNN accuracy rates according to the number of hidden layers

The confusion matrix of the XGBoost algorithm, which obtained the highest accuracy among the artificial intelligence methods used with MLP and DNN algorithms, the motivation source of the proposed study, is presented in Figure 9.

Table 7. Performance comparison of ANN algorithms and other machine learning methods

Models	Accuracy	Precision	Recall	F1-Score	Support
NB	0.692	0.693	0.692	0.692	5753
SVM	0.678	0.683	0.677	0.675	5753
DT	0.806	0.809	0.805	0.805	5753
MLP	0.842	0.845	0.842	0.842	5753
RF	0.86	0.86	0.86	0.86	5753
Gradient boosting	0.865	0.866	0.865	0.865	5753
ET	0.871	0.872	0.871	0.871	5753
DNN	0.875	0.875	0.874	0.874	5753
XGBoost	0.911	0.911	0.911	0.911	5753



(a)

(b)

(c)

Figure 9. **(a)** MLP confusion matrix (1000-1000) **(b)** DNN confusion matrix (250-250) **(c)** XGBoost confusion matrix

In this study, the critical impact of the SMOTE and IterativeImputer data preprocessing techniques on model performance is highlighted. The dataset used in the study contained challenges such as missing data and an imbalanced class distribution. Proper handling of these challenges played a key role in improving classification performance. An imbalanced class distribution can significantly affect the ability of machine learning models to accurately predict minority classes. In this study, the SMOTE method was employed to address the class imbalance issue, and minority class samples were augmented by synthesizing new examples from existing ones. By providing a more balanced structure to the dataset, SMOTE enabled the models to effectively learn from both minority and majority class samples.

Missing data is a common issue in machine learning projects, which can degrade model performance and lead to inaccurate results. In the study, the IterativeImputer method was utilized to overcome the problem of missing data. This method allowed for the estimation and completion of missing data using a linear regression-based approach. IterativeImputer stood out as an effective method, particularly for features with a high level of missing data. Its ability to complete missing data in a manner consistent with the overall structure of the model was a significant factor in improving the accuracy rates obtained.

The proper application of SMOTE and IterativeImputer methods not only enhanced the accuracy of classification models but also improved their generalization capabilities. Specifically, SMOTE's augmentation of minority class samples and IterativeImputer's completion of missing data increased the quality and representativeness of the dataset, contributing to the model's ability to produce more consistent results. This outcome made it possible for data-sensitive models, such as the XGBoost algorithm, to achieve a high accuracy rate of 91.1%.

XGBoost generally performs well on smaller and well-structured data sets, while Neural networks tend to perform better on large, complex, and high-dimensional data sets (Zhang and Lu, 2021). The structure of the dataset used in this study allowed XGBoost to take advantage of its advantages but limited the DNN's capacity to learn complex data patterns. Although the DNN model exhibited a slightly lower accuracy rate (87.5%) compared to XGBoost, it demonstrated the strengths of deep learning. The layered structure of DNNs has the capacity to learn more complex patterns in the data. However, the performance of DNN models depends on the hyperparameter settings (e.g., number of layers and neurons, learning rate) and a large amount of computational power. In this study, model performance is improved by determining the optimal number of layers and neurons. However, it was realized that the size of the dataset should be increased to achieve higher performance.

The study differs from most studies in the literature by addressing both the unbalanced class problem and the missing data problem simultaneously. While most studies focus on one of these problems, this study offers a comprehensive solution by addressing both problems together. For example, Conlon (2021) solved the missing data problem by deleting data and ignored the class imbalance. Tran and Nguyen (2021) addressed the data imbalance with Smote's method and completed the missing data using "fillna". Considering the results we obtained in the study, the study makes a difference with the proposed data preprocessing steps and the proposed machine learning methods.

4. Conclusion

In today's world, companies in different sectors need to deal with large amounts of data from the physical world, such as sensors, RF-IDs, GPS, as well as human-generated data such as social networks and the internet. By collecting, storing, processing, analyzing, and interpreting the results of such data, businesses employ data scientists to provide important information used in business decision-making processes. These people have great costs to businesses in the period from recruitment to integration into the department they work in. For this reason, predicting employees who intend to change their jobs or who are looking for a new job is a very important issue for businesses.

This paper provides significant benefits to businesses in retaining talented employees, improving recruitment strategies, and reducing organizational risks. Additionally, it offers opportunities to enhance the employee experience by analyzing the reasons for turnover. In this way, businesses can anticipate potential losses, achieve cost savings, and maintain their data science capabilities in the long term.

In this paper, a study was conducted to predict the probability of job changes for data scientists who are of great importance to the companies. Using demographic information, experience, etc. that affect the decision of data scientists, the probability of continuing to work in the company or changing jobs was predicted by ANN methods. In this context, the dataset was trained and tested with MLP and DNN. MLP and DNN have a layered structure due to their structure. Therefore, the layers and the number of neurons they contain were determined from low to high, and tests were performed. As a result of these tests, the MLP model with two hidden layers and 1000 neurons in each hidden layer obtained the best results compared to other MLP models with 84.2% accuracy and 84.2% F1-score. Similarly, the DNN model with two hidden layers, each containing 250 neurons, achieved 87.5% accuracy and an 87.4% F1-score. When the results are analyzed, it is seen that keeping the number of neurons in the hidden layers low or increasing them too much negatively affects the performance metrics.

Since the motivation of the study is to demonstrate the performance of ANN algorithms, other machine learning methods in the literature such as DT, RF, ET, NB, SVM, Gradient boosting, and XGBoost were also trained and tested. As a result of the tests, the best accuracy rate was obtained with XGBoost algorithm with 91.1%. The MLP algorithm outperformed the NB, SVM, DT algorithms in terms of accuracy, but underperformed the RF, ET, Gradient boosting, and XGBoost algorithms. Similarly, the DNN algorithm outperformed NB, SVM, DT, RF, ET, and Gradient boosting algorithms in terms of accuracy, while XGBoost showed lower performance.

In future studies, it is planned to collect more data and add it to the dataset to improve our performance results obtained with algorithms. In addition, it is thought that it may be useful to conduct an analysis not only for data scientists but also for all employees working in the company and to conduct sentiment analysis to analyze the opinions of employees about their jobs.

Statement of Conflict of Interest

The authors have declared no conflict of interest.

Author's Contributions

The contribution of the authors is equal.

References

Ajakwe SO., Deji-Oloruntoba O., Olatunbosun SO., Duorinaah FX., Bayode IA. Multidimensional perspective to data preprocessing for model cognition verity: data preparation and cleansing-approaches for model optimal feedback validation. In Recent Trends and Future Direction for Data Analytics 2024; 15-57. IGI Global. doi: 10.4018/979-8-3693-3609-0.ch002

Altunışık R. Büyük veri: Fırsatlar kaynağı mı yoksa yeni sorunlar yumağı mı? Yıldız Social Science Review 2015; 1(1): 45–76.

Aras M. İşveren markasının örgütsel bağlılık ve işten ayrılma niyetine etkisi: Katılım bankacılığı örneği. Doktora Tezi. Sakarya Üniversitesi, 2016.

Asteris PG., Mokos VG. Concrete compressive strength using artificial neural networks. Neural Computing and Applications 2020; 32(15): 11807–11826.

Baumeister F., Barbosa MW., Gomes RR. What is required to be a data scientist?: Analyzing job descriptions with centering resonance analysis. International Journal of Human Capital and Information Technology Professionals (IJHCITP) 2020; 11(4): 21–40.

Bingöl K., Akan EA., Örmecioğlu HT., Er A. Depreme dayanıklı mimari tasarımda yapay zeka uygulamaları: Derin öğrenme ve görüntü işleme yöntemi ile düzensiz taşıyıcı sistem tespiti. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi 2020; 35(4): 2197–2210.

Çavuşoğlu Ü., Kaçar S. Anormal trafik tespiti için veri madenciliği algoritmalarının performans analizi. Academic Platform-Journal of Engineering and Science 2019; 7(2): 205–216.

Colab. Google Colaboratory 2021. https://colab.research.google.com/ (accessed January 21, 2021).

Conlon SJ. Why do data scientists want to change jobs: Using Machine Learning Techniques to Analyze Employees' Intentions in Switching Jobs. IJMIT 2021; 16: 59–71. https://doi.org/10.24297/ijmit.v16i.9058.

Domo. Data Never Sleeps 80 2023. https://www.domo.com/learn/data-never-sleeps-8 (accessed April 24, 2023).

Dutta S., Bandyopadhyay SK. Employee attrition prediction using neural network cross validation method. International Journal of Commerce and Management Research 2020; 6(3): 80–85.

Fiorelli FAS., Campoleone ET., Neto AH. Artificial neural network for predicting energy consumption 2015. https://doi.org/10.13140/RG.2.1.2758.0320.

Gazel SER., Bati CT. Derin sinir ağları ile en iyi modelin belirlenmesi: mantar verileri üzerine Keras uygulaması. Yuzuncu Yıl University Journal of Agricultural Sciences 2019; 29(3): 406–417.

Goh WWB., Hui HWH., Wong L. How missing value imputation is confounded with batch effects and what you can do about it. Drug Discovery Today 2023; 28(9): 103661.

Göde A., Doğan A. License plate recognition system based on artificial ıntelligence with different approach. ECJSE 2023. https://doi.org/10.31202/ecjse.1172426.

Górriz JM., Ramírez J., Ortíz A., Martinez-Murcia FJ., Segovia F., Suckling J. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. Neurocomputing 2020; 410: 237–270.

Güner OÖ. Toplu yemek hizmetlerinde makine öğrenmesi algoritmaları ile talep planlama. Yüksek Lisans Tezi. İstanbul Üniversitesi-Cerrahpaşa, Lisansüstü Eğitim Enstitüsü, Endüstri Mühendisliği Ana Bilim Dalı, 2021.

Kaggle. Kaggle 2022. https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists (accessed May 10, 2023).

Karagöz H. Dijital çağda yönetimde yeni bir anlayış: dijital ikiz entegrasyonu. In: Ceyhan S, Ada S, editors. Dijital Çağda Yönetim Üzerine Güncel Konular ve Araştırmalar, Ankara: Nobel Yayınevi; 2022.

Kartal E., Özen Z. Dengesiz veri setlerinde sınıflandırma. Mühendislikte Yapay Zekâ ve Uygulamaları, 1st Ed, O Torkul, S Gülseçen, Y Uyaroğlu, G Çağıl, and MK Uçar, Eds Sakarya: Sakarya Üniversitesi Kütüphanesi Yayınevi 2017; 109: 131.

Kyalkond SA., Manikanta Sanjay V., Manoj Athreya H., Aithal SS., Rajashekar V., Kushal BH. Data scientist job change prediction using machine learning classification techniques. Ubiquitous Intelligent Systems: Proceedings of Second ICUIS 2022, Springer; 2022, 211–219.

LeCun Y., Bengio Y., Hinton G. Deep learning. Nature 2015; 521(7553): 436–444.

Lee KJ. Architecture of neural processing unit for deep neural networks. Advances in Computers, vol. 122, Elsevier; 2021, 217–145.

Maind MSB., Wankar MP. Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication 2014; 2(1): 96–100. https://doi.org/10.17762/ijritcc.v2i1.2920.

Martínez-Álvarez F., Troncoso A., Asencio-Cortés G., Riquelme JC. A survey on data mining techniques applied to electricity-related time series forecasting. Energies 2015; 8(11): 13162–13193.

Metlek S. Disease detection from cassava leaf ımages with deep learning methods in web environment. International Journal of 3D Printing Technologies and Digital Industry 2021; 5(3): 625–644.

Mohamed A., Najafabadi MK., Wah YB., Zaman EAK., Maskat R. The state of the art and taxonomy of big data analytics: view from new big data framework. Artificial Intelligence Review 2020; 53: 989–1037.

Nkongolo M., Tokmak M. Zero-Day threats detection for critical infrastructures. In: Gerber A, Coetzee M, editors. South African Institute of Computer Scientists and Information Technologists, Cham: Springer Nature Switzerland; 2023, p. 32–47. https://doi.org/10.1007/978-3-031-39652-6_3.

Oliveira JM., Zylka MP., Gloor PA., Joshi T. Mirror, mirror on the wall, who is leaving of them all: predictions for employee turnover with gated recurrent neural networks. Collaborative Innovation Networks: Latest Insights from Social Innovation, Education, and Emerging Technologies Research 2019: 43–59.

Oliveira R., Araújo RC., Barros FJ., Segundo AP., Zampolo RF., Fonseca W., Dmitriev V., Brasil F. A system based on artificial neural networks for automatic classification of hydro-generator stator windings partial discharges. Journal of Microwaves, Optoelectronics and Electromagnetic Applications 2017; 16: 628–645.

Öztürk K., Şahin ME. Yapay sinir ağları ve yapay zekâ'ya genel bir bakış. Takvim-i Vekayi 2018; 6(2): 25–36.

Ramkumar M., Malathi K., Pavithra K. Optimizing machine learning model accuracy via OBNT algorithm: Advanced Data Preprocessing Technique 2024; 1-6: doi: 10.1109/icses60034.2023.10465344

Sadanand D., Bhosale S. Basic of artificial neural network. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) 2023; 3(3): doi:10.48175/IJARSCT-8159

Şeker A., Diri B., Balık HH. Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme. Gazi Mühendislik Bilimleri Dergisi 2017; 3(3): 47–64.

Sexton RS., McMurtrey S., Michalopoulos JO., Smith AM. Employee turnover: a neural network solution. Computers & Operations Research 2005; 32(10): 2635–2651.

Tang J., Yuan F., Shen X., Wang Z., Rao M., He Y., Sun Y., Li X., Zhang W., Li Y., Gao B., Qian H., Bi G., Song S., Yang JJ., Wu H. Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges. Advanced Materials 2019; 31(49): 1902761.

Techjury. Big data statistics 2023. https://techjury.net/blog/big-data-statistics/ (accessed April 6, 2023).

Tekerek A., Bay OF. Design and implementation of an artificial intelligence-based web application firewall model. Neural Network World 2019; 29(4): 189–206.

Tran OT., Nguyen LP. Trainee churn prediction using machine learning: A case study of data scientist job. Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021). CEUR Workshop Proceedings, Da Nang, Vietnam (Oct 2021), 2021.

Varol F. Çalışanların örgütsel bağlılık ve iş tatminlerinin işten ayrılma niyetlerine olan etkisi: İlaç sektörü örneği. Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 2017; 38): 200–208.

Yener S. İşten ayrılma niyetinin belirleyeni olarak psikolojik rahatlık. Anadolu Üniversitesi Sosyal Bilimler Dergisi 2016; 18(3): 169–192.

Yılmaz A. Yapay zeka. İstanbul: KODLAB Yayın Dağıtım Yazılım ve Eğitim Hizmetleri San. ve Tic. Ltd. Şti.; 2020.

Zhang C., Lu Y. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration 2021; 23: 100224.

Zhang Jinghua Li C., Yin Y., Zhang Jiawei, Grzegorzek M. Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. Artificial Intelligence Review 2023; 56(2): 1013–1070.

Zhang Q., Yu H., Barbiero M., Wang B., Gu M. Artificial neural networks enabled by nanophotonics. Light: Science & Applications 2019; 8(1): 42.