
**POLİTEKNİK DERGİSİ**

***JOURNAL of POLYTECHNIC***

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

# Evaluation of profession predictions for today and the future with machine learning methods: emperical evidence from Turkey

## *Makine öğrenmesi yöntemleri ile günümüz ve geleceğe yönelik meslek tahminlerinin değerlendirilmesi: Türkiye'den ampirik deliller*

***Yazar(lar) (Author(s))**: Ebru KARAAHMETOĞLU[1], Süleyman ERSÖZ[2], Ahmet Kürşat TÜRKER[3] , Volkan ATEŞ[4], Ali Fırat İNAL[5]*

*ORCID[1]: 0000-0003-4381-7865*

*ORCID[2]: 0000-0002-7534-6837*

*ORCID[3]: 0000-0001-6686-9241*

*ORCID[4]: 0000-0002-7534-6837*

*ORCID[5]: 0000-0001-6686-9241*

**To cite to this article**: Karaahmetoğlu E., Ersöz S., Türker A.K., Ateş V. ve İnal A.F., "Evaluation of profession predictions for today and the future with machine learning methods : emperical evidence from Turkey",***Journal of Polytechnic***, 26(1): 107-124, (2023).

**Bu makaleye şu şekilde atıfta bulunabilirsiniz**: Karaahmetoğlu E., Ersöz S., Türker A.K., Ateş V. ve İnal A.F., "Evaluation of profession predictions for today and the future with machine learning methods : emperical evidence from Turkey",***Politeknik Dergisi***, 26(1): 107-124, (2023).

**Erişim linki (To link to this article):** http://dergipark.org.tr/politeknik/archive

**DOI:** **10.2339/politeknik.985534**

# Evaluation of Profession Predictions for Today and the Future with Machine Learning Methods : Emperical Evidence From Turkey

## *Highlights*

- *Analyzing documents with text mining methods.*
- *Applying machine learning algorithms with the proposed models.*
- *Performance comparisions with the popular evaluation metrics.*

## *Graphical Abstract*

*In the study, the machine learning algorithms were applied to the documents analyzed by text mining methods. It was tried to predict current and future professions of Turkey with the proposed model.*

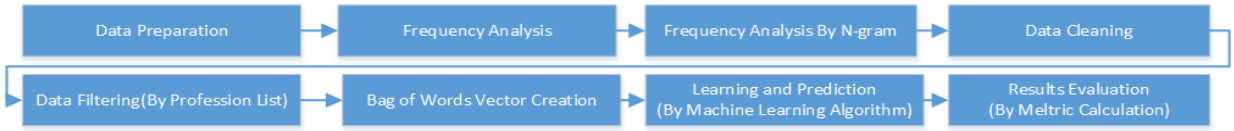


**Figure.** Text Mining Machine Learning Algorithm

## *Aim*

*The aim of this study is to investigate the professions of the future and current in Turkey by the application of supervised learning algorithms and clustering methods to various Turkish data including documents belonging to Turkey's institutions.*

## *Design & Methodology*

*The files obtained from the internet were analyzed by text mining methods and the machine learning algorithms were applied to the documents to produce predictions about the popular jobs of today and future.*

## *Originality*

*The originality of this study is to predict current and future directions of professional job roles as a response to Industry 4.0 impact on digitally disrupted business enrironment in Turkey.*

## *Findings*

*The popular professions were predicted with an accuracy rate between ≅0.81 and ≅0.93. Public jobs were highlighted in terms of popularity because of the economic imbalances in recent periods. Data science will become important role in all occupations in the industry 4.0 age.*

## *Conclusion*

*For analyzing labor trends, text mining approach could be an alternative to more traditional approaches such as employer surveys The compatible results were obtained with the priority occupational areas announced by YOK. SGD and Perceptron algorithm showed superioty to the other algorithms.*

## *Declaration of Ethical Standards*

*The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.*

# Evaluation of Profession Predictions for Today and the Future with Machine Learning Methods : Emperical Evidence From Turkey

*Araştırma Makalesi / Research Article*

**Ebru KARAAHMETOĞLU[1*], Süleyman ERSÖZ[1], Ahmet Kürşat TÜRKER[1]**

**Volkan ATEŞ[2], Ali Fırat İNAL[1]**

[1]Mühendislik ve Mimarlık Fakültesi, Endüstri Mühendisliği Bölümü, Kırıkkale Üniversitesi, Türkiye
[2]Enformatik Bölümü, Kırıkkale Üniversitesi, Türkiye

(Geliş/Received : 20.08.2021 ; Kabul/Accepted : 12.09.2021 ; Erken Görünüm/Early View : 21.09.2021)

**ABSTRACT**

For the purpose of evaluating present and future trends of professions within the labor market, text mining approach could be an alternative to more traditional approaches such as employer surveys. Specifically, machine learning algorithms are used for making accurate predictions about the future directions of the professions which consequently will influence professional development of labour force. The aim of this study is to investigate the professions of the future and current in Turkey by the application of supervised learning algorithms and clustering methods to various Turkish data including documents belonging to Turkey's institutions. In this study, the popular professions were predicted with an accuracy rate between ≅0.81 and ≅0.93 thorough various machine learning algorithms. It was discovered that methodologically perceptron and stochastic gradient descent algorithms demonstrated superiority over other algorithms thanks to their intelligence functions. Furthermore, the analysis of current professions in Turkey revealed that the class of "Professional occupations", "Managers" and "Technicians and assistant professional members" were popular, and according to the analysis of the future, information technology-based occupations will be important. Although limited Turkish data sources for the analysis of future, results with an accuracy of nearly 1 were produced.

**Keywords: Text mining, machine learning, professions, supervised learning.**

# Makine Öğrenmesi Yöntemleri ile Günümüz ve Geleceğe Yönelik Meslek Tahminlerinin Değerlendirilmesi : Türkiye'den Ampirik Deliller

**ÖZ**

İşgücü piyasasındaki mesleklerin mevcut ve gelecekteki eğilimlerini belirlemede metin madenciliği yaklaşımı, işveren anketleri gibi geleneksel yöntemlere alternatif olarak kullanılabilir. Teknik olarak, iş gücünün mesleki gelişimini etkileyecek mesleklerin, gelecekteki eğilimleri hakkında doğru tahminlerde bulunmak için makine öğrenme algoritmaları kullanılmaktadır. Bu çalışmanın amacı, Türkiye'deki kurumlara ait belgeler de dahil olmak üzere, çeşitli Türkçe verilere denetimli öğrenme algoritmaları ve kümeleme yöntemleri uygulanarak, Türkiye'deki geleceğin ve şimdiki mesleklerin araştırılmasıdır. Çalışmada, çeşitli makine öğrenme algoritmaları aracılığıyla ≅0.81 ve ≅0.93 arasında bir doğruluk oranıyla, popüler meslekler tahmin edilmiştir. Metodolojik olarak Perceptron ve Stokastik Gradyan İniş algoritmalarının, içerdiği zeka fonksiyonları sayesinde diğer algoritmalara göre üstünlük gösterdiği keşfedilmiştir. Ayrıca, Türkiye'deki mevcut mesleklerin analizi, "Profesyonel meslekler", "Yönetici" ve "Teknisyen ve meslek mensubu yardımcıları" sınıfının popüler olduğu ve gelecek analizine göre bilgi teknolojisi tabanlı mesleklerin önemli olacağı çıkarımı yapılmıştır. Geleceğin analizi için sınırlı Türkçe veri kaynakları olmasına rağmen, yaklaşık 1 doğrulukta sonuçlar üretilmiştir.

**Anahtar Kelimeler: Metin madenciliği, makine öğrenmesi, meslekler, denetimli öğrenme.**

## 1. INTRODUCTION

With the advances in artificial intelligence technologies, the development of robot technologies has accelerated mechanization in professions. The automation of some of the business processes, now done by machines and robots, and ever-increasing usage of artificial intelligence may cause the partial or complete disappearance of some professions in the future. Nevertheless, new professions will also begin to emerge due to the changing business models and processes within digitally distrupted world [1].

**Sorumlu Yazar (Corresponding Author)*
*e-posta : erogluebru@hotmail.com*

In the future business models, educated and qualified workforce plays an important role in capital movements; therefore, low labor cost advantages may not well be seen as a key competitive advantage for emerging economies such as Turkey. Most cannot attract investments to the country. Countries now give importance to human capital investments in order to gain political and economic power. It has been argued that there is a close relationship between education and development in the future business models; consequently, countries should create further strategies for enhancing capabilities of their human capital [2].

Now, the fourth industrial revolution, built upon digital revolution, is taking place. None of the changes, technology, management, production etc. in the past have been as devastating as the changes in the fourth industrial revolution [3].

The term Industry 4.0, was first used in Germany and, is based on high technology strategies [4]. Meanwhile, possible effects of Industry 4.0 on employment vary among various occupations and sectors. According to the McKinsey report, the easiest occupations that entail repeatitive occupations physical tasks within predictable working environments such as operating machinery or preparing fast food will be objected to the full automation ; nevertheless, some of the distictive occupational categories related to big data also can be done better and faster by machines [1].

According to the study of Macurova, Ludvik and Žwakova (2017), organizations will be exposed to the risk of not finding the skilled workforce required by the changes coming with Industry 4.0. It has been claimed that due to the speed of change, the implementation of training strategies for new professions will be delayed, thereby the need for staff held required qualification will not be met [10]. While jobs requiring low-level skills are expected to decrease with Industry 4.0, it is expected to lead to increase in job supply in high-skilled labor force within the service sector, especially such as information technologies and scientific professions [11].

In the study of Kane et al, it was mentioned that some strategic changes in the education systems at national level are expected to accommodate new professions [12]. All the countries need to analyze their education systems and professional qualification programmes in the context of digitalization in production in order to determine their strengths and weaknesses [13], and should reorganize their education program to prepare their workforce for the future skills [14].

At a country-specific level, according to the professions' report prepared by the Ministry of Science and Industry in 2018, there is an intense youth unemployment in Turkey [15]. The reason for this unemployment is that the higher education is not geared towards the needs of industry level trends [2]. Current and future professionals should be career ready according to the skill sets needed for emerging in time with changes happening within various industries. As a matter of fact, in Turkey, YÖK (Higher Education Institution) announced the priority occupational groups for new professions and started to work on shaping the educational programs in this direction [17].

Identifying popular professions and creating future projections about professional jobs require the extraction of important information from data sources. In general, text mining techniques are used to extract hidden information in large data sets. For example, Bach et al. determined the knowledge and skills required by industry 4.0 organizations by examining job postings [18]. Meanwhile, Mauro et al. classified the abilities required by professions in their work [19]. In Frank et all.' s study, the barriers that inhibit scientists from measuring the effects of AI and automation on the future of work were discussed [20]. In the study of Dawson et all., job ads data and employment statistics were used to predict skills shortage and it was determined which skills are most important for occupations in shortage [21]. In the study of Boselli et al., If appropriately retrieved and analyzed the detailed and valuable information about the Web Labor Market dynamics and trends was obtained from the huge number of job vacancies available today on on-line job portals [22]. In the study of Papaoutsoglou et al., the digital sources of the software engineering labor market were analyzed, and the encountered issues were tried to be classified. Its aim was to connect different skill types, needs and goals of labor market with the utilization of digital sources and data analysis methods [23]. Based on our literature review, it was evident that in the existing literature, there are not enough studies about estimating the popular professions of the future. Therefore, the aim of this study is to predict current and future directions of professional job roles as a response to Industry 4.0 impact on digitally distrupted business enrironment in Turkey by making use of various Turkish resources, including documents belonging to institutions throughout the country for human resources employment. In short, it is tried to find answer to the question of "What is the popular occupation of current and future?". In this study, it was inspired from many studies related to the keywords in occupational groups in line with the work of Özköse (2017) [24].In this study, the documents obtained in the google search engine had search criteria of "IPA (Labor Market Research) Report", "Professions" and "Future Professions". Data sets passed through text mining processes are classified by machine learning algorithms of Knn [25], Naïve Base [26], Random Forest [27], Lassolars [28], [29], ElasticNet [30], [31], Stochastic Gradient Descent (SGD) [32] and Perceptron [33]. According to the measurement metrics applied to the learning models, estimation data with accuracy varying between ≅0.81 and ≅0.93 were generated. Furthermore, the calculations showed that in the analysis of the "IPA Report" and "Professions" documents, SGD and Perceptron algorithms, in the analysis of "Future Professions" documents, the Random Forest algorithm produced the best result; in the analysis of both, the

Elasticnet and Lassolars algorithms produced the worst result.

The ability to analyze changes in professions with the effect of disruptive technologies in business processes caused by Industry 4.0 adds value to the study in terms of literature. In our study, materials and methods will be mentioned in the second section, results and discussions will be presented in the third section, and the conclusions will be listed in the final section.

## 2. MATERIALS AND METHODS

### 2.1. Data Mining

Data mining is the process of discovering new meaningful correlations, patterns and trends by using pattern recognition technologies together with statistical and mathematical techniques, through the elimination of data stacks stored in storage media [34].

Processes specified by CRISP-DM (CRoss Industry Standard Process for Data Mining) are used in data mining applications [36]. These processes are as follows:

- Problem Definition: A problem definition is made from the project objectives determined according to the business requirements.
- Data Discovery: It includes the collection, definition and analysis of the data subject to analysis.
- Data Preparation: It is the process in which data is prepared for the modeling stage by using cleaning and transformation processes.
- Modeling: It is the process which it is tried to be found solutions in accordance with the problem definition by using mathematical models applied to the data.
- Evaluation: It is the process in which the obtained results are evaluated with various metrics.

Text mining is a data mining method in which the text is used for extracting important data for linguistic purposes. In text mining, it is tried to make inferences, which are previously unknown and useful for analysis, from unstructured data. With the aid of the information obtained, it is found that there are relationships, hypotheses and tendencies unseen clearly in the text [38], [39].

Firstly, text mining was used for R&D purposes, to extract key information from large volumes of data. Along with technological developments, it has been used to produce solutions to a research question with project-based approaches. Over time, with the growth of the volumetric size of the data subject to analysis, text mining applications have emerged where the end user experience is processed and the important information is extracted [40].

Most of text mining operations are as follows [41]:

- Feature Extraction tries to find facts and relations in text.
- Text-base navigation is used for extracting key informations.
- Search and Retrieval is used for searching internal.
- Categorization (Supervised Classification) is the process of classifying documents.
- Clustering (Unsupervised Classification) is to group documents on the basis of a criteria.
- Summarization is the operation that reduces the amount of text in a document while still keeping its key meaning.

There are some limitations in text mining arising from the complexity of natural language. One of them is ambuguity, which means one word may be used in different meaning in context or sentences may be interpreted as different meaning. Semantic analysis is required to obtain fine grained information, but it is expensive in terms of complexity. The other limitation is the multilingual text refining. Each language has different phonetic features which effect the information extraction, thereby, different natural language algorithm must be developed for each language [41].

In the study of Miner et al., it was presented that text mining was related with disciplines and methods of data mining, natural language processing, artificial intelligence and machine learning, statistics, computational languages, classification, clustering, etc. [42].

### *2.2. Machine Learning Algorithm*

There are two different learning models in machine learning algorithms: supervised and unsupervised. Supervised algorithms contain data related to the model to be taught to the system. The relationship between data input and output values are tried to be found by teaching this data to the system. Supervised learning algorithms can be of two different features: simple regression and classification. In classification, outputs are tried to be estimated in separate classes, while trying to produce predictions in a continuous trend in regression algorithms [43].

In unsupervised learning, there is only data. Just as there is no information about what the data is related to, there is no feedback on predictive values. In these algorithms, it is tried to make inferences about the relationships and properties of the data by making use of clustering methods [44].

Classification is one of the most common application areas of data mining. The main task in classification is to assign a known class label to data consisting of a set of variables. This classification process is done with a classifier model created with a learning algorithm applied to the training set. The performance of the established classifier model is evaluated by a separate test data set [45].

In the classification process, there are many methods and algorithms used to create models. Knn [25], Naïve Base [26], Random Forest [27], Lassolars [28], [29],

ElasticNet [30], [31], Stochastic Gradient Descent (SGD) [32] and Perceptron [33] algorithms, which were briefly explained as follows, were coded in Phyton by using scikit-learn library.

2.2.1. K-nearest neighbours (Knn) Algorithm

In the Knn algorithm, firstly, $n$ center points are determined. Data set elements are clustered in the group of their closest center point by calculating their distances to $n$ center points. After clusters of $n$ center points have been created, the clusters are compared with each other. As a result of this comparison, the displacement of the cluster elements and whether the change occurred in the clusters are checked. If there is a change in the clusters, the algorithm continues to run, otherwise the process stops. The distances of the $n$ points selected from the test data to the training data are calculated and then the distance values are ranked from small to large. The representation of the Knn algorithm in which the first $k$ points selected are assigned to a class according to the most frequently encountered class is given in Figure 1 [25].

Quick calculation time and simplicity in terms of complexity can be counted among its advantages.The high accuracy is the most important feature of the algorithm. In addition, the accuracy value depends on the quality of the data subject to analysis. Other downsides are high data usage, slow response time with the large volume of data [25].

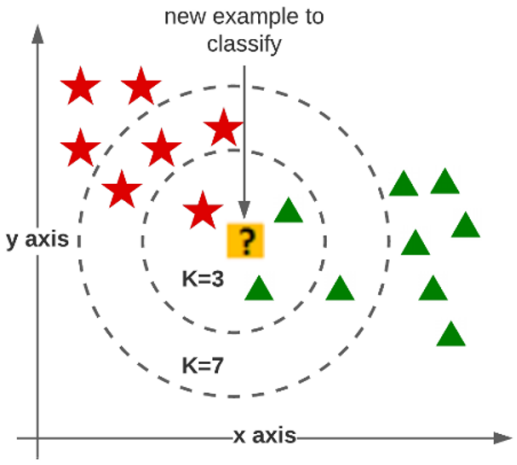


**Figure 1.** Knn Algorithm Classification.

2.2.2. Naive Bayes Algorithm

The Naive Bayes algorithm is a probability-based classifier based on the Naive Bayes theorem. Naive Bayes algorithm gives good results when the model consists of categorical variables [46]. Naive Base is based on the assumption that the variables based on the model are independent from each other. With this feature, it gives good results in very complex machine learning problems.

Bayes' theorem is a mathematical statement based on conditional probabilities. Conditional probability is defined as the probability of occurrence of these two events together if events $A$ and $B$ are interconnected. Provided that $P(A)$, $P(B)$ are the probability of the occurrence of event $A$ and $B$, $P\ (A \mid B)$ is the probability of event $A$ if event $B$ occurs, $P\ (B \mid A)$, the probability of occurrence of $B$ event if event $A$ occurs, is calculated by using the product rule of the probability theorem as in (1).

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad (1)$$

Naive Bayes algorithm is successfull in multiclass prediction problems and better suited for categorical input variables. Naive Bayes assumes all the features are independent. Although this assumption provides that it can perform better and requires less data, it has no use of real life [26].

2.2.3. Random Forest Algorithm

Random Forest algorithm is a tree-based decision making algorithm. The final decision tree is created by collecting the values from different subdecision trees created with the data set, as shown in Figure 2.

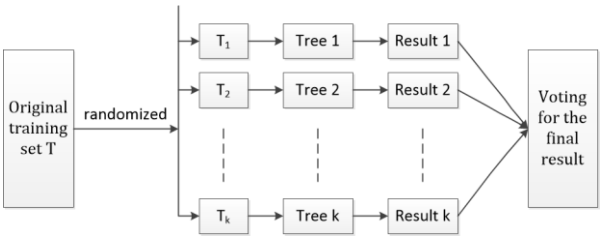


**Figure. 2.** Random Forest Algorithm [47]

A random forest consists of a collection of classifiers in the tree structure expressed as $\{h(x,\theta_k), k = 1, \ldots\}$ with the condition that $\theta_k$ is an independent random forest. Each subtree, which is a member of the random forest, has one vote in choosing the popular class for the $x$ input value. These votes are used in the formation of the final result [47] .

In Breiman's RF model, all subtrees form the classifier $h(x,\theta_k)$ for the input vector $x$, with the learning dataset and the random variable $\theta_k$ corresponding to the $k$‘ th tree [27]. After the classifiers run $k$ times, the classifier sequence $\{h_1(x), h_2(x), \ldots, h_k(x)\}$ is obtained. Using these classifier sequences, the classifier model consisting of the decision function below is reached:

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (2)$$

In this equation, $H(x)$ is the combination of classifier models, $h_i$ is a singular tree, $Y$ is the output value, $I(\cdot)$ is the pointer function.

Random Forest algorithm requires no normalization, thereby, data preperation is easy. The use of multiple subtrees give predictive power to the algorithm. Random forest algorithm is very sensitive to data. A small change

in data cause considerable amount of change in result [47].

### 2.2.4. Lasso ve ElasticNet Algorithms

Lasso and ElasticNet are based on linear regression algorithm. While error corrections use quadratic values in Lasso's algorithm, ridge regression is based on absolute values. ElasticNet algorithm uses two features at once [48]. The simple regression algorithm is represented by the expression of $y = \beta_0 + \beta_1 x_1 + \varepsilon$. As a matrix, it is expressed as follows:

$$H\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \dots & x_1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & x_n \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (3)$$

At this point, the coefficient is calculated as $\beta = (X'X)^{-1}X'Y$. In the known regression algorithm, incompatible results occur with regression curve in case high correlation between variables [49]. Therefore, in ridge regression, a variable is added to the expression to reduce error as follows (4):

$$\beta(ridge) = \arg min\|y - x\beta\|^2 + \lambda\|\beta\|^2 \quad (4)$$

The value of $\lambda \geq 0$ is defined as the correction or complexity coefficient in the above expression where the quadratic correction is made. Similarly, there is a correction in the Lasso algorithm. Differently, the correction is in the form of absolute value, as seen below:

$$\beta(lasso) = arg\, min\|y - x\beta\|^2 + \lambda\|\beta\| \quad (5)$$

Absolute value error correction is important to avoid high error values being doubled by square operation.

ElasticNet algorithm, which is overcoming the limitations of Lasso algorithms with the use of a penalty function, uses a mixture of two techniques in Ridge and Lasso algorithms and shown in Figure 3 and expressed as follows [30]:

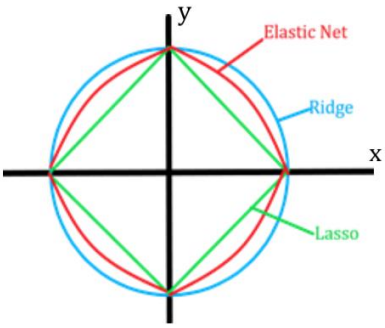


**Figure 3.** Lasso and ElasticNet Algorithm

$$\beta(elasticnet) = \arg min\|y - x\beta\|^2 + \lambda_1\|\beta\|^2 + \lambda_2\|\beta\| \quad (6)$$

Lasso and ElasticNet are regression algorithm which produce good results in continuos variables, not in categorical variables.

2.2.5. Stochastic Gradient Descent

The Stochastic Gradient Descent algorithm is a popular algorithm in machine learning and forms the basis for artificial neural networks [50]. Gradient descent algorithm concerns with the slope of the curves. By moving down along the slope, it is tried to find the $x$ value that constitutes the smallest $y$ value.

Gradient is the slope of a function. It is calculated by looking at how much the change in one variable affects the other. Gradient descent is mathematically a convex function whose output value is the partial derivative of input values.

In a simple learning model, let $z$ be a pair of data $(x, y)$ consisting of random input value of $x$ and output value of $y$. The loss function measures the cost of estimating $\hat{y}$ while $y$ is the actual value. There is a family of functions $f_w(x)$ defined by the weight vector w, and called $\mathcal{F}$. From this point on, the $f \in \mathcal{F}$ function that will make the loss function $Q(z, w) = \ell(f_w(x), y)$ minimum is tried to be found. The success of the training data set is measured by using the experimental risk ($E_n(f_w)$) expression shown below [51]:

$$E_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \quad (7)$$

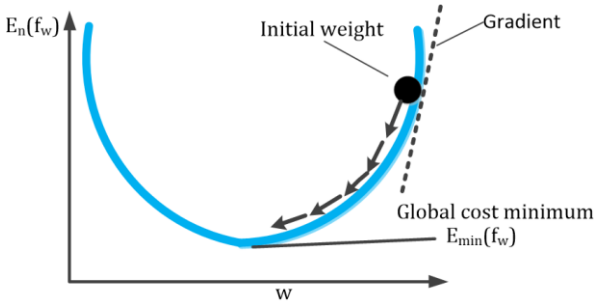


**Figure 4.** Stochastic Gradient Descent Algorithm

Gradient descent algorithm is presented as minimization of $E_n(f_w)$ loss function by using gradient descent and graphically represented in Figure 4 [52]. The weight values of $w$ are updated with the expression of (8) in order to make the loss function minimum in the gradient descent algorithm.

$$w_{t+1} = w_t - \gamma\frac{1}{n}\sum_{i=1}^{n} \nabla_w Q(z_i, w_t) \quad (8)$$

In the Stochastic Gradient Descent algorithm, the predicted value is used based on the randomly selected single $z_t$ sample instead of calculating all gradient values of the loss function $E_n(f_w)$ at each step. Thus, the weight update expression is simplified and shown below:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_i, w_t) \quad (9)$$

The stochastic process, $\{w\_t, t = 1, \dots\}$, depends on the randomly selected sample in algorithm steps.

Because of the use of a single training example and the processing one sample at a time, it is fast and easy to fit. Besides, achieving the convergence takes longer due to the noisy data steps.

2.2.6. Multi Layer Perceptron (MLP)

Perceptron is the mathematical function that models a neuron in the simplest way. Perceptron consists of one or more input values, processor and one output value. The input function of perceptron algorithm is shown in Figure 5. In a generalized perceptron algorithm, the total value is found by adding the products of *n* input values and weights. If this total value is greater than a threshold value, the function takes the value of 1, otherwise it takes the value of zero [53]. The power of generalization in Perceptron provides with classifying an unknown pattern with other known patterns. Perceptron algorithm used too much parameters, thereby, it is resulting in redundancy and inefficiency. The Perceptron algorithm is mathematically expressed in (10).

$$y = f(z)$$
$$z = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + b \qquad (10)$$
$$f(z) = \begin{cases} 0, & f(z) < 0 \\ 1, & f(z) \geq 0 \end{cases}$$

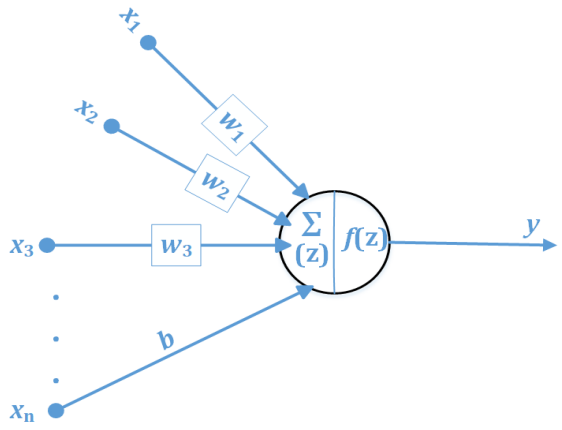


**Figure 5.** Perceptron Algorithm Input Function

### 2.3. Data Sources and Variables

Once the recent studies in the literature are examined, it is seen that search engines have turned into very useful web tools to create the data sets needed for data mining and text mining [54], [55], [56], [57]. IPA reports published by the Employment Agency are a valuable resource in terms of conducting research on professions and revealing the current situation. For this reason, the documents found with the "IPA report" and "Professions" search texts on the google search engine were selected as the subject of the research. Documents downloaded with the application, which was developed within the scope of the research, were analyzed by text mining techniques.

In the application of data preparation, documents was found according to the search criteria using google search libraries. The documents found were downloaded with http request. If the document format is pdf, the content was obtained by extracting text from the pdf file using the PDFMiner phyton library. For reading files in html format, Beatifulsoup method was used to parse the html documents and get the real text out of them. Beatifulsoup is phyton language library that extracts data from files in html format by using data scraping techniques. In short, Beautifulsoup helps to pull particular content from a webpage, remove the HTML markup, and save the information. The schematic representation of the data preparation application is given in Figure 6.

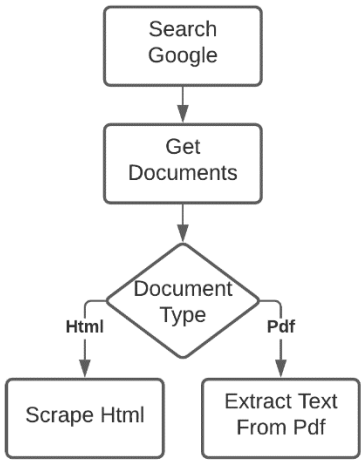


**Figure 6.** Data Preparation Application

The 151 documents were used in the scope of this study. While 76 of these documents were used for learning phase, 75 of them were used for estimation. The professions mentioned in the documents were classified by using the sector classification given in the Table 1. These list was taken from the ISCO (International Standard Classification of Occupations) [51].

**Table 1.** Profession Classification List.

| Value | Profession Class |
|---|---|
| 1 | Occupation related to the armed forces |
| 2 | Managers |
| 3 | Professional occupations |
| 4 | Technicians, technicians and assistant professional occupations |
| 5 | Employees working in office services |
| 6 | Service and sales staff |
| 7 | Skilled agriculture, forestry and aquaculture workers |
| 8 | Craftsmen and those working in related jobs |
| 9 | Plant and machine operators and installers |
| 10 | Those who work in jobs that do not require qualification |

The occupational names included in the frequency list were given to the model as a variable. In the study conducted with the "IPA report" and "Professions" documents, 2735 variables were used. The occupations found in documents were used as variables in the model.The variable value in the model input matrix was assigned as the number of occupations in the document

It is very unlikely to make inferences about future professions from IPA reports. For this reason, the documents reached with the search text "Future

Professions" in the google search engine have been worked on. Analysis on the future professions concentrated on jobs in the field of information technology. Of the 71 documents obtained within the scope of the study, 36 were used in the learning phase, while 35 were used for estimation. The professions of the future are classified according to the priority occupational groups announced by YÖK (Higher Education Institution) and modeled with machine learning algorithms. Classes used in classification are listed in Table 2 [17]. 117 variables were used in the model.

**Table 2.** Future Profession Classification List

| Value | Profeession Class |
|---|---|
| 0 | Network technologies (5G, Internet of Things) |
| 1 | Smart Energy Systems |
| 2 | Smart and Innovative Materials |
| 3 | Biomaterials and Tissue Engineering |
| 4 | Biomedical and Biomedical Technologies |
| 5 | Blockchain Technology |
| 6 | Algebra and Coding Theory |
| 7 | Industrial Engineering |
| 8 | Climate Change |
| 9 | Advanced Robotic Systems and Mechatronics |
| 10 | Advanced and Intelligent Manufacturing |
| 11 | Quantum Information and Quantum Machine Learning |
| 12 | Virtual and Augmented reality technologies |
| 13 | Cyber Security / Cryptology |
| 14 | System Engineering |
| 15 | Aquaculture and Fisheries Technology |
| 16 | Sustainable Agriculture (including Innovative and Good Agricultural Practices) |
| 17 | Sustainable and Intelligent Transportation |
| 18 | Remote Sensing and Geographic Information Systems |
| 19 | Data Science and Cloud Computing |
| 20 | Fuels (Fossil and Bio) and Combustion |
| 21 | Artificial Intelligence and Machine Learning |
| 22 | Software engineering |
| 23 | Innovative Food Processing Technologies and Food Biotechnology |
| 24 | IT Law |
| 25 | Digitalization in Education |
| 26 | Nursing |
| 27 | Molecular Biology and Genetics (Gene therapy and Genome Studies) |
| 28 | Rehabilitation Medicine and Assistive Technologies |
| 29 | Regenerative Medicine |
| 30 | Healthy Nutrition and Food Additives |
| 31 | Aging and Elderly Health |
| 32 | Human Robot Communication |
| 33 | Social media |

Data cleaning followed the process of creating a frequency matrix in text mining. The data cleanup process includes cleaning excess spaces, removing punctuation marks, conjunctions and prepositions, lowercase and stemmer algorithm. All processes except the stemmer algorithm were used in the study because it causes information loss.

### 2.4. Model Evaluation Metrics

The accuracy of the classifications of the proposed model is measured by some performance measurement metrics.

In a $n$-element data set, the error term is calculated by the expression of $e_j = A_j - P_j$, provided that $A_1, A_2, \dots, A_n$ are the actual values, $P_1, P_2, \dots, P_n$ are predicted values.

**Mean Absolute Error (MAE):** It is calculated by taking the average of the absolute value of the prediction errors and shown as below:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|e_j| \tag{11}$$

The absolute values of the difference eliminates the reducing effect of the negative values. Small error values indicate a high degree of agreement between the forecast and actual values. It is expected to produce more accurate results than the known error metrics, because, large error values are squared and their effect on the total error value is multiplied in quadratic calculation methods. For this reason, absolute error methods are used as a valid criterion in cases where data which are incompatible with the regression curve cannot be determined by data cleaning processes.

**Mean Absolute Percentage Error (MAPE):** The error is the mean of the percentages relative to the absolute values of the difference between the predicted and the true value. It is calculated by the following expression.

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\frac{|e_j|}{|A_j|} \tag{12}$$

The mean absolute percentage error (MAPE) is one of the most widely used measures of forecast accuracy, due to its advantages of scale-independency and interpretability. It works best if there are no extremes to the data and no zeros [58].

### 2.5. Confusion Matrix

Confusion matrix is the matrix shown in Table 3, which classifies the predicted values as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) in classification problems. In this classification method, the high values of predictive data in true positive (TP) and true negative (TN) classes are considered as good.

The total data count is calculated by summing the number of data of all classes in the confusion matrix ($TP + TN + FP + FN$). While the number of true ones is calculated as ($TP + FN$), the number of false ones is calculated as ($TN + FP$).

The accuracy and precision measurements of the model are performed with the metrics, which use error matrix classification [59]:

**Table 3.** Confusion Matrix

| | | Real Values | |
|---|---|---|---|
| | | False (0) | True (1) |
| Prediction | False (0) | TN | FN |
| | True (1) | FP | TP |

**Accuracy:** It is calculated by the ratio of correctly predicted data to the total number of data.

$$Accuracy = \frac{TP + TN}{TOTAL} \quad (13)$$

**Recall:** It shows how many of the correct classes in the forecast data were correctly predicted.

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

**Precision:** It is calculated as the ratio of the number of true positive data predicted correctly to the total number of positive predicted data.

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

**$f_1$ score:** $f_1$ score are used to combine two metric values and evaluate with a mixed value.

$$f_1 = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (16)$$

## 3. RESULTS AND DISCUSSIONS

### 3.1. Recommended Model

In the proposed method, the files obtained from the internet environment with the Data Preparation step shown in Figure 6 were subjected to frequency analysis. The contents of the document were divided into words and how many times each word occurs in the text was determined by frequency analysis. Frequency analysis was calculated for occupational names consisting of more than one word, using binary (bigram), triple (trigram), quadruple (fourgram) word groups with n-gram algorithms. Word groups were formed by consecutive words in the text. In the Data Cleaning step, unrelated words and alpha characters in the text were eliminated by the stopwords data set and alpha characters control specific to the Turkish language. The occupational names mentioned in the text were obtained by filtering with the ISCO occupation list. Machine learning data entries were created as a matrix in the Bag of Words Vector Creation step. Bag of Words is used for extracting key information from the text data. A bag of words is composed of keywords and occurrence of this keywords in the documents. Thus, text data were numberized and converted into the form that can be used by machine learning algorithms. In Learning and Prediction Phase, the machine learning algorithms described in section 2.2 were applied to the learning data set for training while the test data set for prediction. The predictions produced by machine learning algorithms were evaluated with the measurement metrics.

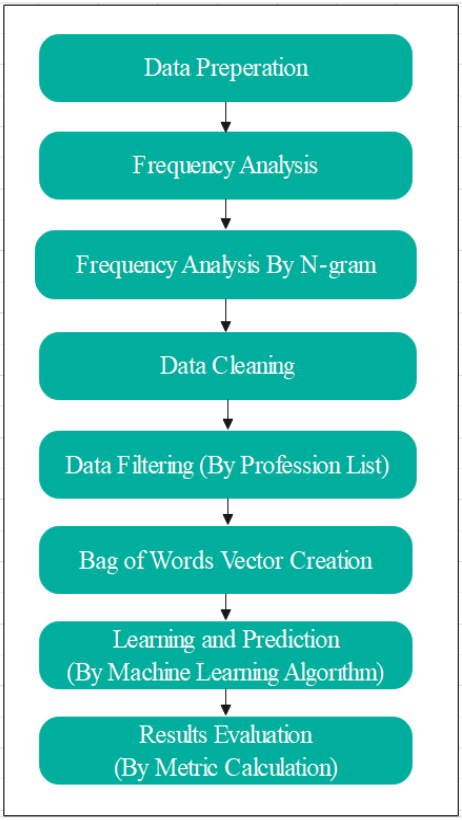


**Figure 6.** Text Mining Machine Learning Algorithm

Let the corpus of the proposed method be mathematically expressed by the pair $\mathcal{K} = (\mathcal{T}, \mathcal{D})$. In this model, while job names are expressed as $\mathcal{T} = (t_1, t_2, \dots, t_m)$, the documents are referred as $D = (d_1, d_2, \dots, d_n)$. In the light of these expressions, the frequency matrix of the model corpus is shown as $A = (a_{ij})$ for $A \in \mathbb{R}^{m \times n}$. The frequency of the job $t_i$ in the document of $D_j$ is calculated as how much it has passed in it, and shown as $a_{ij} = |\{p|t_{jp} = t_i\}|$. According to occupational classes, it is calculated as $a_{ij} = |\{p|t_{jp} = mesleksinif(t_i)\}|$.

Model output value is calculated by using occupational sector frequency analysis as follows: Provided that $B = (b_1, b_2, \dots, b_n)$ is one row of frequency matrix, occupational class value of $y$ is calculated as below:

$$y = \frac{\sum_{i=0}^{n} i \times b_i}{\sum_{i=0}^{n} b_i} \quad (17)$$

Machine learning algorithm input and output values were arranged as bag of words vector according to the occupational frequency values of the documents and were shown in Table 4. The words obtained from all documents by frequency analysis were labeled with a variable number in a matrix. If they were mentioned in the document, occurence value of them in the document

**Table 4.** Bag Of Words Model Input Matrix.

| Doc | accountant | cleaning staff | customer representative | Front accountant | chef | automation systems technician | y |
|---|---|---|---|---|---|---|---|
| Doc 1 | 30 | 28 | 19 | 12 | 17 | 1 | 6 |
| Doc 2 | 32 | 25 | 4 | 28 | 6 | 0 | 6 |

**Algorithm 1.** Recommended Model

**Input:** *N* documents, occupation_list, *M* ml_algorithm, *K* occupation_class
**for** *i* = 1,2, … , N **do**
*wt* ← tokenize( *docs(i)*)
*wt* ← eliminate_lambda_stopwords(wt)
*freq_lis*t ← frequency_analysis (wt)
*analysis(i)* ← ( *docs(i),* freq_list)
*total_frequency_list* ← calculate_total_frequency( *freq_list* )
*profession_list* ← filter_with_occupation_list ( *total_frequency_list, occupation_list*)
(*learn_analysis_list , test_analysis_list*) ← split_data_into_two_set ( *analysis* )
Splitting data into two set as odd, even, fifty, fifty
**for** *i* = 1,2, … , *learn_analysis_list.count* **do**
*frequency_list* ← *learn_analysis_list(1)*
**for** *j* = 1,2, … , *frequency_list.count* **do**
*index* ← index_of *( frequency_list(i)(0), occupation_list)*
*word_vector(index) = word_vector(index) + frequency_list(i)(1)*
*bag_vector(i)* ← *word_vector*
*learn_bag_vector* ← *bag_vector*
*test_bag_vector* ← generate_bag_vector (*test_analysis_list*)
Test bag vector was generated like learn_bag_vector calculation.
**for** *i* = 1,2, … , *learn_bag_vector.count* **do**
*bag* ← *learn_bag_vector(i)*
**for** *j* = 1,2, … , *bag.count* **do**
*profession_name* ← *profession_list(j)*
*profession_class* ← find_profession_class **(***profession_name, occupation_list)*
*profession_class_count(profession_class)* ←
*profession_class_count(profession_class) + bag(j)*
*toplam1 =0*
*toplam2 =0*
**for** *k* = 1,2, … , *profession_class_count.count* **do**
*toplam1* ← *toplam1 + (k+ 1) * profession_class_count (k)*
*toplam2* ← *toplam2 + profession_class_count(k)*
*y_train(i)* ← *toplam1 / toplam2*
*y_test = y_test_from_bag_vector(test_bag_vector)*
Test output value were calculated like *y_train*.
**for** *i* = 1,2, … , *ml_algorithm.count* **do**
learn ( *learn_bag_vector, y_train* )
*y_pred* ← predict ( *test_bag_vector, y_test* )
evaluate_metrics (*y_test, y_pred*)
confusion_matrix (*y_test, y_pred*)

if not, 0 were given to the algorithm as the variable value of *x*.

Model Learning algorithm results were subjected to consistency and accuracy analysis with measurement metrics.

The algorithm of represented model was shown in Algorithm 1. In this algorithm, firstly the documents obtained were tokenized into the words. Stopwords and lambda characters were eliminated from the tokenized words. Then, the frequency list of them was calculated and stored into analysis collections for each document. The professions encountered in all the documents were found by filtering the total frequency list with the occupation list. After that, the analysis data were splitted into two dataset as learning and test. Bag vectors, which are the inputs of the algorithm, were calculated as summing the frequency of the professions in each document and setting the index of the professions encountered in the profession list.

The output value of bag vector were calculated as dividing product sum of profession count and class index into sum of class index. This process was performed for both learn and test data sets. The machine learning algorithms were applied to the bag vectors in order of learning and predicting. Finally, the results obtained and the performance of the algorithms were evaluated by measurement metrics and confusion matrix.

Machine learning algorithms integrated in the recomended model has some limitation parameter given as follows:

In Knn algorithm, the parameter of number of neighbour, which is important for classification, was taken as 3 while the leaf size, the speed of the construction of the tree and query, was taken as 10. In naive base algorithm, smoothing variable is used as 1e-9. It is used for stability by adding to variances. In random forest algorithm, 10 trees were used in forest. In LassoLars algorithm, the coefficient of penalty term was taken as 0.01. In Elasticnet algorithm, the coefficient of penalty term was taken as 1. In SGD Algorithm, the stopping criterion was used as 1e-3 while the maximum iteration was taken as 1000. In Perceptron Algorithm, the stopping criterion was used as 1e-3 while the maximum iteration was taken as 1000.

In the proposed method, the documents obtained as search result of "IPA report" and "Professions" search text were downloaded via the application developed in phyton language. Frequency analysis, one of the text mining processes, was run on these documents, and then the frequency list were filtered by the professions specified in the occupation database. Machine learning algorithms were run by using 76 training and 75 predictive data out of 151 documents obtained from the internet. Machine learning algorithms trained with the learning data set were expected to produce estimation with the prediction data set.

The collective frequency analysis chart for the profession names was shown in Figure 7. In the frequency analysis graph, Among the 2735 professions mentioned in all documents, the 17 most common professions with the lowest frequency of 60 were listed. As the data are ordered in descending order, the graph is damped along the x-axis and approaches to 0.

According to the collective frequency analysis of the professions seen in Figure 7, the profession classes of "Professional occupations", "Technicians, technicians and assistant professional members", "Managers" and "Artisans and those who work in related jobs" are popular profession classes. The frequent inclusion of public professions in the analysis results shows that the appeal of public jobs has increased in the recent period of increasing unemployment.

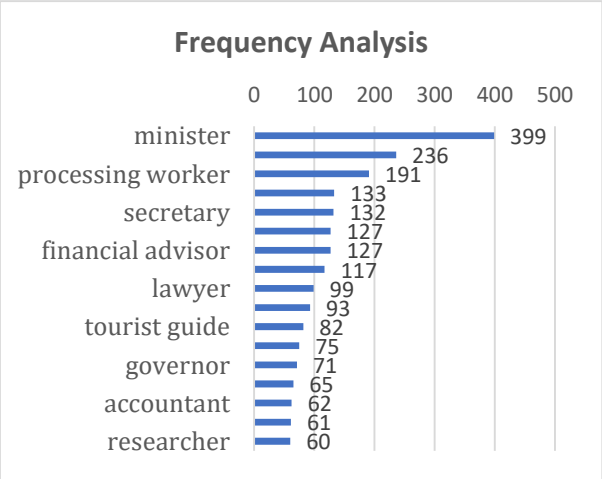


**Figure 7.** Frequency Analysis of Current Popular Occupations

According to the prediction produced by the machine learning algorithms, the profession classes of "Professional occupations", "Managers" and "Technicians and assistant professional members" have come to the fore as shown in Figure 8. As a result of the changes in business models, it is obvious that robots will not take over the professions that will coordinate the jobs and the professions that require high skill levels. The estimation results has also supported this fact.

Experimental measurement metrics were applied to all algorithms in the learning model and the comparative results were presented in Table 5. The best results were produced by SGD algorithm. This situation has been explained that SGD Algorithm is equipped with intelligence functions and produces successful results in categorical variables. The worst results were given by LassoLars and ElasticNet because LassoLars and ElasticNet are regression algorithms and the regression algorithms produce bad results in categorical variables. Perceptron is also based on intelligence function. So it produced good result like SGD algorithm. Additionally Knn algorithm gave good accuracy results with neighboor of 3. If the number of neighboor is increased, it is possible to obtain better accuracy value.

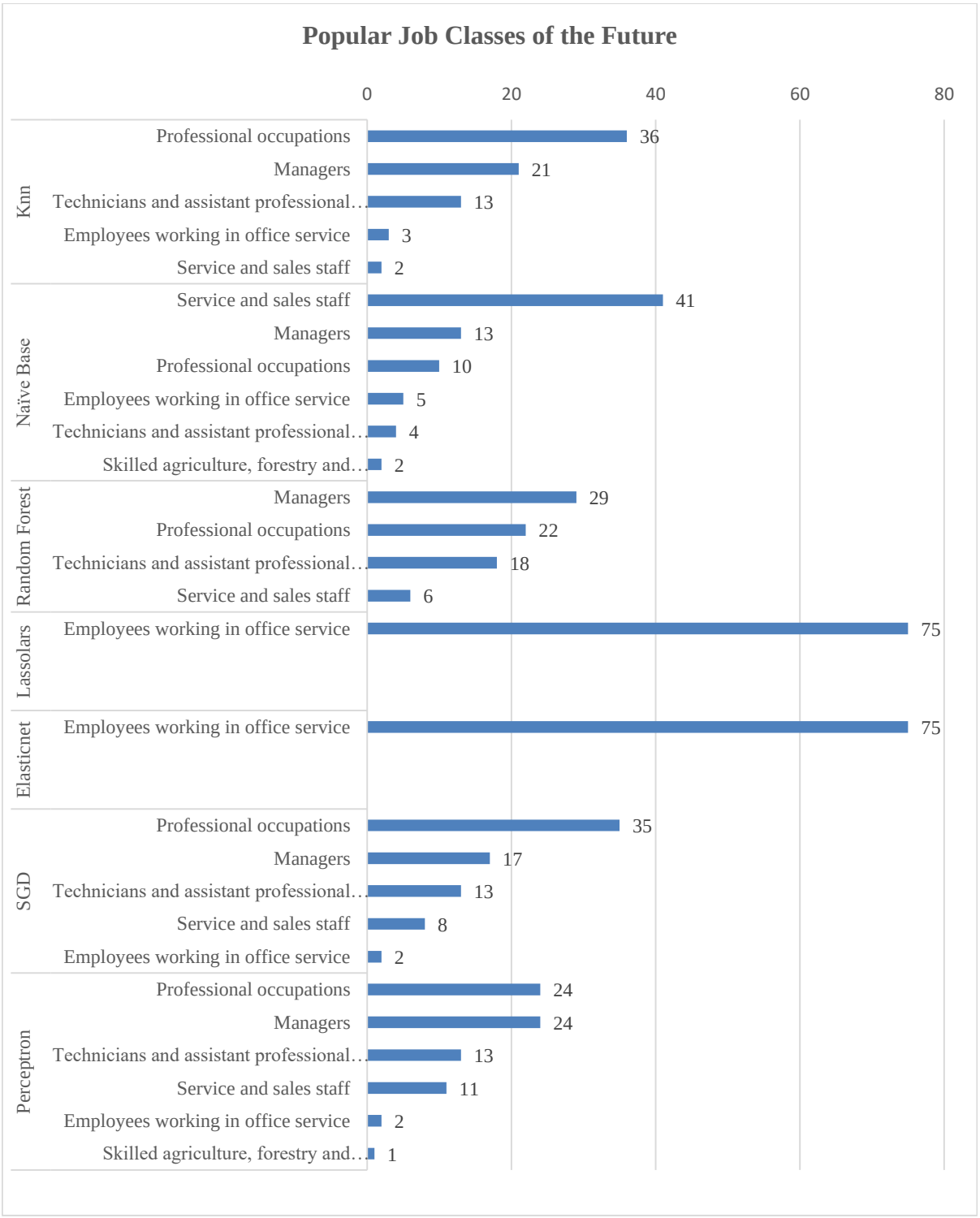


**Figure 8.** Current Popular Job Classes According to Estimation Results

That the number of neighboor value is 3 is optimal for the number of documents used in the analysis. The random forest algorithm, which usually produces high accuracy results as its characteristic feature, produced high accuracy results in the study. According to the number of documents used in the analysis, the number of subtrees was chosen as 10.

Better results can be obtained in terms of accuracy and measurement metrics by increasing the number of subtrees and documents used. Although Naïve Base algorithm creates good results in multiclass dataset, it couldn't produce best results. Because the limited number of document were subjected to the analysis. On the other hand, it produced better results than regression algorithms. That was also seen in the results that good results had been obtained according to Lassolars and Elasticnet algorithms.

The best performance was given by SGD algorithm in according to the mean absolute error metric which can produce correct results when data compatible with the classification cannot be determined properly. The worst performance was given by LassoLars and ElasticNet algorithm with ≅1.52. According to the mean absolute percentage error metric, the best result was given by SGD algorithm with ≅0.11, while the worst result was given by ElasticNet and LassoLars with ≅0.82.

**Table 5.** Comparative Metric Results

| Algorithm | MAE | MAPE(%) |
|---|---|---|
| Knn | 0,77 | 0,15 |
| Naïve Base | 1,33 | 0,57 |
| Rand-Forest | 0,78 | 0,16 |
| LassoLars | 1,52 | 0,82 |
| ElasticNet | 1,52 | 0,82 |
| SGD | 0,64 | 0,11 |
| Perceptron | 0,70 | 0,13 |

**Tablo 6.** Confusion Matrix Values

| Algoritma | TP | FP | TN | FN |
|---|---|---|---|---|
| Knn | 44 | 644 | 31 | 31 |
| Naïve Base | 16 | 616 | 59 | 59 |
| Rand-Forest | 43 | 643 | 32 | 32 |
| LassoLars | 3 | 603 | 72 | 72 |
| ElasticNet | 3 | 603 | 72 | 72 |
| SGD | 47 | 647 | 28 | 28 |
| Perceptron | 48 | 648 | 27 | 27 |

The confusion matrix results according to the algorithms were calculated by using the confusion matrix values in Table 6 and shown in Table 7.

The prediction results have the lowest confusion matrix accuracy value of ≅0.81. According to the accuracy value, the highest result was given by SGD and Perceptron algorithm with ≅0.93, while the lowest result was given by LassoLars and ElasticNet algorithm with ≅0.81. These results are compatible with the measurement metrics result seen in Table 5.

According to the recall, the best result was obtained from Perceptron algorithm with 0.64, while the lowest value was obtained from LassoLars and ElasticNet algorithm with the value of ≅0.04. The same results were obtained according to the precision metric.

According to $f_1$ score calculated by using recall and precision, the highest result was from Perceptron algorithm with a value of ≅0.64, while the lowest result was from LassoLars and ElasticNet algorithm with a value of 0.04.

**Tablo 7.** Comparative Confusion Matrix Result

| Algorithm | Acc. | Recall | Precision | f1 score |
|---|---|---|---|---|
| Knn | 0,92 | 0,59 | 0,59 | 0,59 |
| Naïve Base | 0,84 | 0,21 | 0,21 | 0,21 |
| Random Forest | 0,91 | 0,57 | 0,57 | 0,57 |
| LassoLars | 0,81 | 0,04 | 0,04 | 0,04 |
| ElasticNet | 0,81 | 0,04 | 0,04 | 0,04 |
| SGD | 0,93 | 0,63 | 0,63 | 0,63 |
| Perceptron | 0,93 | 0,64 | 0,64 | 0,64 |

According to the metric measurements presented, the profession classification results obtained with a high accuracy of ≅0.93 seem to be compatible with today's economic and occupational employment conditions, and shed light on the future. It has been inferred that there could be improvements in $f_1$ score by increasing the variety of the model data set and providing improvements in data cleaning processes.

At this stage of the study, the analysis processes were run with the documents obtained by searching with the search text "Future Professions". In the analysis process for future professions, the frequency analysis of the documents were created and filtered according to the occupation list created by adding various professions to the professions list having been announced by Özdemir and Kılınç [60].

Among the 117 professions mentioned in all documents, the 10 most common professions with the lowest frequency of 10 were listed in Figure 9. When the graphical representation is examined, the genetic profession (121 times) has created a serious awareness compared to other professions. Recently, many disease has been tried to to be treated, thanks to the science of genetics, that seems in line with this popularity. The popularity of information technology professions is an

indication that information technology processes have become an essential part of production processes with the Industry 4.0 revolution. Popular professions of the future are shown in Figure 9.

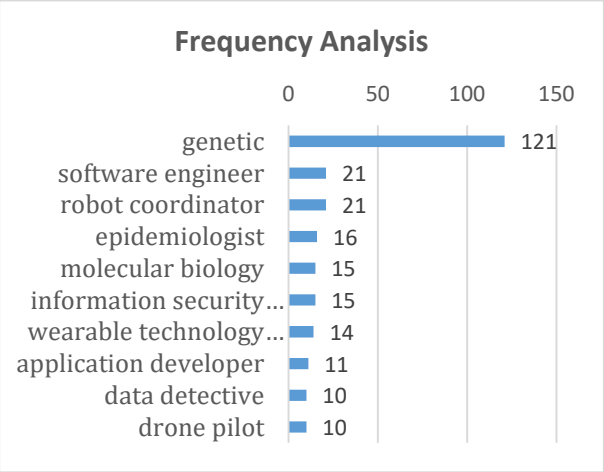


**Figure 9.** Frequency Analysis of Popular Occupations of Future

Technology-oriented occupational classes such as "Software Engineering" and "Digitalization in Education" stand out in the estimation results according to algorithms. "Nursing", "Molecular Biology and Genetics" are other important classes. Popular occupational classes according to the estimation results were given in Figure 10.

**Table 8.** Comparative Evaluation Metric Results

| Algorithm | MAE | MAPE(%) | Accuracy |
|---|---|---|---|
| Knn | 2,11 | 0,26 | 1 |
| Naïve Base | 2,29 | 0,35 | 1 |
| RandomForest | 1,88 | 0,24 | 1 |
| LassoLars | 2,33 | 0,35 | 1 |
| ElasticNet | 2,43 | 0,37 | 1 |
| SGD | 2,27 | 0,35 | 1 |
| Perceptron | 2,11 | 0,30 | 1 |

Analysis evaluation results for future professions were listed in Table 8. Algorithms gave results with high accuracy (≅1). According to the mean absolute percentage error, RandomForest algorithm gave the best result with ≅0.24, while algorithm gave the worst value with ≅0.37. According to the mean absolute error, Random-Forest gave the best result with a value of ≅1.88. The worst result was given by ElasticNet algorithm with ≅2.43. This situation is due to the fact that the model includes categorical variables.

The matching of the professions, which are the professions of the future in the studies conducted so far, with the predictions of YÖK's priority occupational areas [61] was presented in Table 9. These match shown in the table was found in the scope of the study as a result of the future profession analysis given in detail above. It was observed that compatible results were obtained according to the specified matches. The professions of the future used in the study were classified according to the priority occupational areas announced by YÖK. Priority occupational areas were matched with the occupations obtained as a result of the frequency analysis shown in Figure 9 for the future occupations. This match is an indication of the accuracy of the analysis results, despite a few data.

**Table 9.** YÖK Priority Occupational Areas and Profession Matches.

| Priority Occupational Areas | Occupations |
|---|---|
| Advanced and Intelligent Manufacturing | Industrial Data Scientists<br>Wearable Technology Designer |
| Sustainable and Intelligent Transportation | Smart City Expert |
| Data Science and Cloud Computing | Data Detective<br>Data Analysis Specialist<br>Personal Data Broker<br>Industrial Data Scientists<br>Cloud Computing Specialist<br>Software specialist<br>Software developer |
| Robot Technologies | Robot Coordinator<br>Robot Coordination Specialist |
| Cyber Security / Cryptology | Data Security Specialist<br>Information Security Analyst |
| Remote Sensing and Geographical Information Systems | Drone Pilot |
| Drug Studies | Molecular Biology<br>Bioengineering<br>Epidimiologist |
| Molecular Pharmacology and Pharmaceutical Research | Molecular Biology<br>Bioengineering<br>Epidimiologist |
| The Human Brain and Neuroscience | Genetic |

## 4. CONCLUSIONS

In this study, popular professions for today and the future were tried to be estimated by using text mining techniques. With Industry 4.0, it has been concluded that, the appeal of information technology-based jobs will increase, and the occupation classes of professional, managers and technicians will maintain their attractiveness in business models changing as a result of mechanization in production.

As a result of the analysis of today situation, the profession classes of "Professional occupations", "Managers" and "Technicians and assistant professional members" shows superiority to the other classes. The

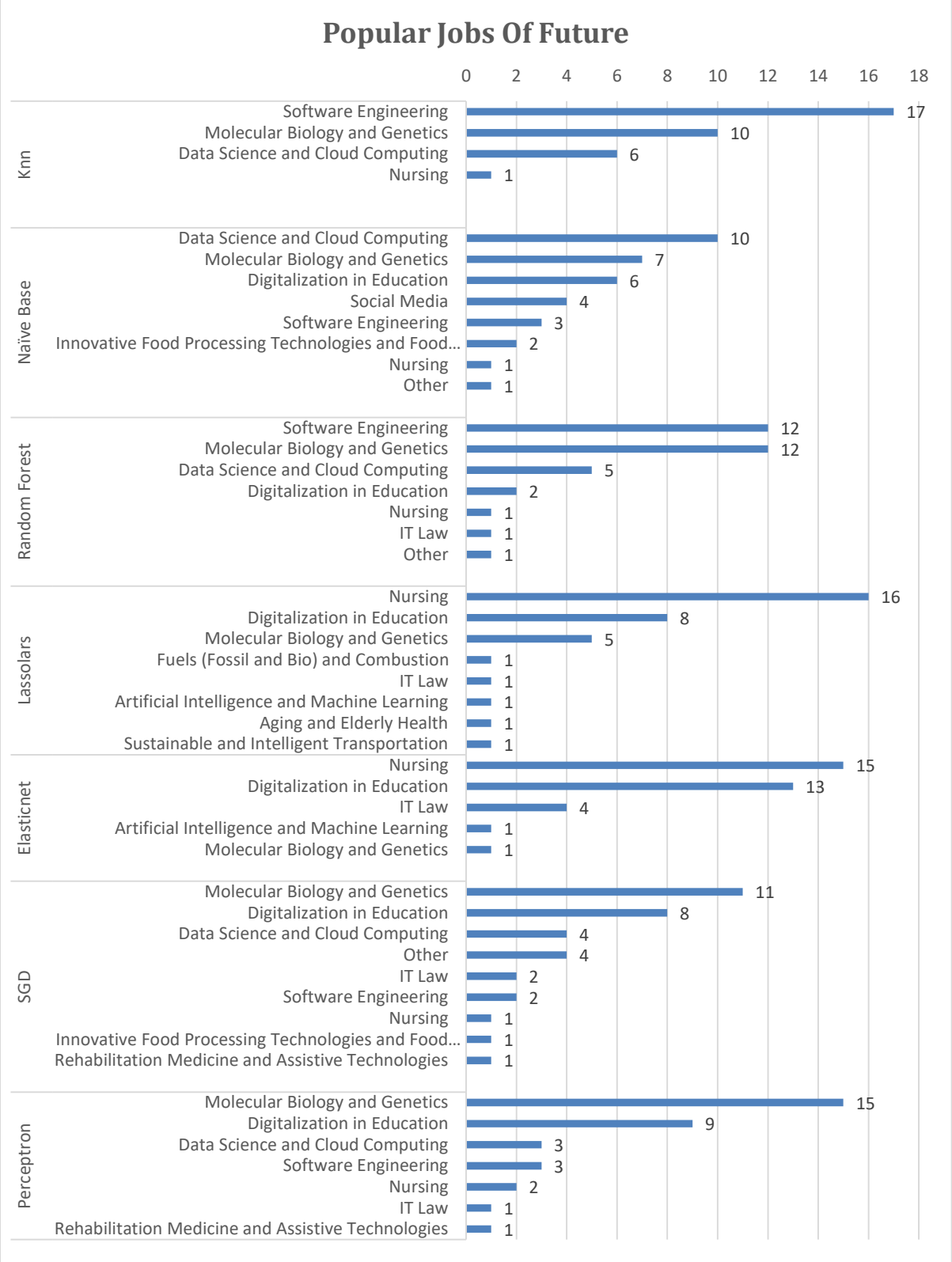


**Figure 10.** Popular Job Classes of Future According to Estimation Results

profession classes of "Professional occupations", "Managers" and "Technicians and assistant professional members" were seen frequently in the estimation results obtained from Knn, Naïve Base, Random Forest, SGD and Perceptron algorithms. It can be projected from these results that the changes in business models will not affect the professions that will coordinate the jobs and the professions that require high skill levels. Public jobs were highlighted in terms of popularity because of the economic imbalances in recent periods.

According to the result of the analysis of future professions, information technology professions showed great importance. This is explained as information technology professions have become an essential part of production processes with the Industry 4.0 revolution. In addition, it was clearly seen from the frequency graphics that the occupation of genetics had created serious awareness over other occupations. This could be easily explained since the genetics has great importance on treating incurable diseases.

According to the estimation results obtained from future profession analysis, information technology professions such as ”Software Engineering”, “Data Science and Cloud Computing” had dominated other job areas. In addition, the job areas of “Molecular Biology and Genetics” showed great attraction in parallel with the technological advancement in medicine. With the industry 4.0, it is expected that data science will become important role in all occupations. This was seen clearly from the estimation results that “Nursing” attracted attention rather than doctor.

In the study, it was tried to be matched the future professions, which was found as a result of the analysis conducted so far, with the priority occupational areas announced by YOK to verify the correctness of the model. As expected, it was observed that the compatible results were obtained.

The analysis of the future professions pointed out the existence of professions unclassified by the priority occupational groups of YÖK. it will be useful to add the occupational groups of "Management and Organization", "Smart City", "Ethics Expertise" to the priority occupational areas announced by YÖK.

According to the analysis, SGD and Perceptron algorithm showed superioty to the other algorithms. LassoLars and ElasticNet showed bad performance in according to the evaluation metrics because the model is composed of categorical variables and LassoLars and ElasticNet algorithms are regression algorithm. In the performance comparison according to the prediction results produced by the classification algorithms, predictions with an accuracy of ≅93% were produced. In case of increasing the variety and volume of data, it is expected that the differentiation of intelligence-based algorithms compared to other algorithms will become more evident. On the other hand, obtaining results compatible with the current economic conditions of Turkey shows that popular professions can be predicted by text mining techniques.

As a result, the contribution of this study to the literature can be summarized as reaching that today's popular professions, which are currently being tried to be determined by employer surveys, can be determined by text mining techniques and the future projections can be made with the pediction produced by the machine learning algorithms. Achieving results compatible with the priority occupational areas of YÖK offers a reinforcing effect to this study.

Some limitations were encountered in this study. Since there is no occupation list published by a standard authority for the future occupations, the occupation list used in the study were compiled from the works of futurists and the internet. The classification was made according to the priority occupational areas of YÖK within the scope of the study. In addition, there couldn't find enough documents for the analysis of future professions.

For future research, the study can be developed by identifying the skill required by industry 4.0 and revealing the relationships between these skills and professions. In the study, Turkey situation was tried to be analyzed from Turkish documents. The study can be extended to analyze the situation in the world by using English documents..

## DECLARATION OF ETHICAL STANDARDS

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

## AUTHORS' CONTRIBUTIONS

**Ebru KARAAHMETOĞLU:** Performed research, software development, development of experiments, analysis of results and writing of the paper.

**Süleyman ERSÖZ:** Performed research, development of experiments and analysis of results.

**Ahmet Kürşat TÜRKER:** Performed the processes of developing the experiments and analyzing the results.

**Volkan ATEŞ:** Performed the analysis of the results of the experiments and the review of the paper.

**Ali Fırat İNAL:** Performed the review of the article.

## CONFLICT OF INTEREST

There is no conflict of interest in this study.

## REFERENCES

[1] Manyika J., Chui M., Bughin J., Dobbs R., Bisson P., and Marrs A., “Disruptive technologies: Advances that will transform life, business, and the global economy,” ***McKinsey Global Institute***, (2013).

[2] Öztürk N., “İktisadi Kalkınmada Eğitimin Rolü,” ***Sosyoekonomi***, 1:27–44, DOI:10.17233/se.86714, (2005).

[3] Schwab K., "The Fourth Industrial Revolution", ***World Economic Forum,*** Geneva, Switzerland, (2016).

[4] Mosconi F., "The new European industrial policy: Global competitiveness and the manufacturing renaissance", London, (2015).

[5] Russmann M., "Industry 4.0: World Economic Forum", ***Bost. Consult. Gr.***, 1–20, (2015).

[6] Huimin M., "Strategic plan of 'Made in China 2025' and its implementation", ***Anal. Impacts Ind. 4.0 Mod. Bus. Environ.***, 19: 1–23, (2018).

[7] Kurt R., "Industry 4.0 in Terms of Industrial Relations and Its Impacts on Labour Life", ***Procedia Comput. Sci.***, 158: 590–601, (2019).

[8] Blinder A. S., "Education for the Third Industrial Revolution", Princeton University, Department of Economics, Center for Economic Policy Studies,Working Papers, (2008).

[9] Pamuk N. S. and Soysal M., "Yeni Sanayi Devrimi Endüstri 4.0 Üzerine Bir İnceleme", ***Verimlilik Dergisi***, 1:41–66, (2018).

[10] Macurova P., Ludvik L., and Žwakova M., "The driving factors, risks and barriers of the industry 4.0 concept," ***Journal of Applied Economic Sciences***, vol. 12(7): 2003-2011, 2017.

[11] Weber E., "Industry 4.0 – job-producer or employment-destroyer?", ***Institute for Employment Research***, (2016).

[12] Kane G. C., Palmer D., Phillips A. N., and Kiron D., "Is Your Business Ready for a Digital Future?", ***MIT Sloan Management Review***, 56(4):7–44, (2015).

[13] Kleinert C., Matthes B., and Jacob M., "IAB Forschungsbericht 5/2008", (2008).

[14] Özkan M., Al A., and Yavuz S., "Uluslararası Politik Ekonomi Açısından Dördüncü Sanayi-Endüstri Devrimi'nin Etkileri ve Türkiye", ***Siyasal Bilimler Dergisi,*** 1–30, (2018).

[15] Bilim ve Sanayi Bakanlığı, "Mesleklerin Geleceği Araştırma Raporu", (2018).

[16] Işık V., "Türkiye'de Genç İşsizliği ve Genç Nüfusta Atalet", ***HAK-İŞ Uluslararası Emek ve Toplum Dergisi***, 11:131–145, (2016).

[17] Yükseköğretim Kurulu Başkanlığı, "Geleceğin Meslekleri Çalışmaları Çalıştay Raporları", (2019).

[18] Pejic-Bach M., Bertoncel T., Meško M., and Krstić Ž., "Text mining of industry 4.0 job advertisements" ***International Journal of Information Management***, 50:416–431, (2020).

[19] De Mauro A., Greco M., Grimaldi M., and Ritala P., "Human resources for Big Data professions: A systematic classification of job roles and required skill sets", ***Information Processing & Management***, 54(5): 807–817, (2018).

[20] Frank M.R., Bessen J.E. , Brynjolfsson E., Cebrian M., Deming D.J., Feldman M., Groh M., Lobo J., Moro E., Wang D., Younk H. and Rahwana I., "Toward understanding the impact of artificial intelligence on labor,", ***PNAS***, 116(14):6531–6539, (2019).

[21] Dawson N., Rizoiu M. A., Johnston B. and Williams M. A., "Predicting Skill Shortages in Labor Markets: A Machine Learning Approach", ***2020 IEEE International Conference on Big Data***, 2:3052–3061, (2020).

[22] Boselli R., Cesarini M., Marrara S., Mercorio F., Pasi M.M.G. and Viviani M., "WoLMIS: a labor market intelligence system for classifying web job vacancies", ***Journal of Intelligent Information Systems***, 51:477–502, (2018).

[23] Papoutsoglou M., Ampatzoglou A., Mittas N., and Angelis L., "Extracting Knowledge from On-Line Sources for Software Engineering Labor Market: A Mapping Study", ***IEEE Access***, 7:157595-157613, (2019).

[24] Özköse H., "Yönetim Bilişim Sistemleri Alanının Türkiye ve Dünya'daki Bibliyometrik Analizi ve Haritası", ***Gazi University, Enformatic Institute***, (2017).

[25] Cover T. and Hart P., "Nearest Neighbor Pattern Classification", ***IEEE Transactions on Information Theory***, 13(1): 21–27, (1967).

[26] Berrar D., "Bayes' theorem and naive bayes classifier", ***Encyclopedia of Bioinformatics and Computational Biology ABC of Bioinformatics***, 1–3:403–412, (2018).

[27] Breiman L., "Random forests", Machine Learning, 45(1):5–32, (2001).

[28] Efron B., Hastie T.,Johnstone I. and Tibshirani R., "Least angle regression", ***Annals of Statistics***, 32(2): 407–499, (2004).

[29] ***statweb.stanford.edu/~tibs/lasso/simple.htm***, "A simple explanation of the Lasso and Least Angle Regression", (2015) .

[30] Zhang Z., Lai Z., Xu Y., Shao L., Wu J. and Xie S.G., "Discriminative Elastic-Net Regularized Linear Regression", ***IEEE Transaction on Image Processing***, 26(3):1466–1481, (2017).

[31] Zou H. and Hastie T., "Erratum: Regularization and variable selection via the elastic net", ***Journal of the Royal Statistical Society. Series B: Statistical Methodology***, 67(5):768, 2005.

[32] Ketkar N., "Stochastic Gradient Descent,", ***Deep Learning with Python***, 113–132, (2017).

[33] Rosenblatt F., "The perceptron: A probabilistic model for information storage and organization in the brain", ***Psychological Review***, 65(6): 386–408, (1958).

[34] Seifert J. W., "CRS Report for Congress Data Mining", ***Reading***, 1–16, (2004).

[35] ***www.baskent.edu.tr/~gmemis/courses/datamining/DM_1.pdf***, "Veri madenciliği 1", (2019) [36] Wirth R., ",***Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining***, 24959:29–39, (2000).

[37] Delen D. and Crossland M. D., "Seeding the survey and analysis of research literature with text mining", ***Expert Systems with Applications***, 34(3):1707–1720, (2008).

[38] Mecca G., Raunich S., and Pappalardo A., "A new algorithm for clustering search results," ***Data & Knowledge Engineering***, 62(3):504–522, (2007).

[39] Witten I. H., "Text mining: Practical handbook of internet computing", ***Chapman & Hall/CRC Press***, (2005).

[40] Iarrobino M., "The Evolution of Text Mining – Trends We're Seeing Across R&D Organizations", ***http://www.copyright.com/blog/trends-evolution-text-mining/***, 2021.

[41] Gupta V. and Lehal G. S., "A Survey of Text Mining Techniques and Applications", ***Journal of Emerging Technologies in Web Intelligence***, 1(1): 60–76, (2009).

[42] Miner G. D., Elder J., and Nisbet R. A., "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications", ***Academic Press*** , (2012).

[43] Lidy T. and Rauber A., "Classification and Clustering of Music for Novel Music Access Applications", ***Machine Learning Techniques for Multimedia, Springer***, 249:285, (2008).

[44] Kamikawa Y. and Kato T., "Development of liquid-crystalline folate derivatives: Effects of intermolecular hydrogen bonds at oligopeptide moieties", ***Polymer Preprints***, Japan, 55(2):2659–2660, (2006).

[45] Agaoglu M., "Predicting Instructor Performance Using Data Mining Techniques in Higher Education", ***IEEE Access***, 4:550, (2016).

[46] ***medium.com/@datalabtr/naïve-bayes-algoritması -ve-r-uygulaması-4d321869d371,*** "Naïve Bayes Algoritması ve R Uygulaması", (2019).

[47] Liu Y., Wang Y. and Zhang J., "New machine learning algorithm: Random forest", ***Lecture Notes in Compuer. Science***, 7473:246–252, 2012..

[48] Zou H. and Hastie T., "Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays", ***Journal of the Royal Statistical Society, Series B***, 67(1):301–320, (2003).

[49] Rençber Ö. F. and Bağcı H., "Sermaye Yeterliliğini Etkileyen Değişkenlerin Elastik Net Regresyon Yöntemi İle Belirlenmesi," OPUS Uluslararası Toplum Araştırmaları Dergisi, DOI: 10.26466/opus.561915, (2019).

[50] Shalev-Shwartz S. and Ben-David S., "Stochastic Gradient Descent", ***Understanding Machine Learning***, 150–166, (2014).

[51] Bottou L., "Stochastic gradient descent tricks", ***Lecture Notes in Computer***, 7700:421–436, (2012).

[52] Rumelhart D. E., Hinton G. E., and Williams R. J., "Learning internal representations by error propagation", ***Parallel Distributed Processing: Explorations in the Microstructure of Cognition***, MIT Press, 1:318–362, (1986).

[53] Chandra A. L., "Perceptron Learning Algorithm: A Graphical Explanation Of Why It Works.", ***https://towardsdatascience.com/perceptron-learning-algorithm-d5db0deab975***.

[54] Hu F. and Trivedi R. H., "Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content", ***International Journal of Hospitality Management***, 84, (2020).

[55] Vanhala M., Lu C., Peltonen J., Sundqvist S., Nummenmaa J. and Järvelin K., "The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research", ***Journal of Business Research***, 106:46–59, (2020).

[56] Hassani H., Beneki C., Unger S., Mazinani M. T. and Yeganeg M. R. i, "Text mining in big data analytics", ***Big Data and Cognitive Computing***, 4(1):1–34, (2020).

[57] Xie X., Fu Y., Jin H., Zhao Y. and Cao W., "A novel text mining approach for scholar information extraction from web content in Chinese", ***Future Generation Computer Systems***, 111:859–872, (2020).

[58] Glen S., "Mean Absolute Percentage Error (MAPE)." ***https://www.statisticshowto.com/mean-absolute-percentage-error-mape/***.

[59] Ohsaki M., Wang P., Matsuda K., Katagiri S., Watanabe H. and Ralescu A., "Confusion-matrix-based kernel logistic regression for imbalanced data classification", ***IEEE Transactions on Knowledge and Data Engineering***, 29(9):1806–1819, (2017).

[60] Özdemir D., Kılınç Ş., "Geleceğin Meslekleri Listesi", 2019.

[61] Yüksek Öğretim Kurumu, "Geleceğin Meslekleri Çalışmaları", ***Geleceğin Meslekleri, Mesleklerin Geleceği Çalıştayı***, (2019)

## Appendix

**Tablo** Profession Dictionary.

| English | Turkish |
|---|---|
| accountant | muhasebeci |
| application developer | uygulama geliştiricisi |
| auditor | denetçi |
| chartered accountant | yeminli muhasebeci |
| data detective | veri dedektifi |
| digital tailor | dijital terzi |
| drone pilot | drone pilotu |
| epidemiologist | epidemiyolog |
| financial advisor | finans danışmanı |
| general manager | genel müdür |
| genetic | genetik |
| governor | Vali |
| information security analyst | bilgi güvenliği analisti |
| lawyer | avukat |
| minister | bakan |
| molecular biology | moleküler biyoloji |
| nurse | hemşire |
| pilot | pilot |
| president | Başkan |
| processing worker | işletmen |
| researcher | araştırmacı |
| robot coordinator | robot koordinatörlüğü |
| secretary | Sekreter |
| servant | hizmetçi |
| software developer | yazılım geliştirici |
| software engineer | yazılım mühendisi |
| tourist guide | turist rehberi |
| wearable technology designer | giyilebilir teknoloji tasarımcılığı |
| web designer | web tasarımcısı |