PAPER Special Section on Intelligent Information Processing to Solve Social Issues

A Methodology on Converting 10-K Filings into a Machine Learning Dataset and Its Applications

Mustafa SAMI KACAR[†a)], **Semih YUMUSAK**[†], ***Members***, ***and*** **Halife KODAZ**[††], ***Nonmember***

SUMMARY Companies listed on the stock exchange are required to share their annual reports with the U.S. Securities and Exchange Commission (SEC) within the first three months following the fiscal year. These reports, namely 10-K Filings, are presented to public interest by the SEC through an Electronic Data Gathering, Analysis, and Retrieval database. 10-K Filings use standard file formats (xbrl, html, pdf) to publish the financial reports of the companies. Although the file formats propose a standard structure, the content and the meta-data of the financial reports (e.g. tag names) is not strictly bound to a pre-defined schema. This study proposes a data collection and data preprocessing method to semantify the financial reports and use the collected data for further analysis (i.e. machine learning). The analysis of eight different datasets, which were created during the study, are presented using the proposed data transformation methods. As a use case, based on the datasets, five different machine learning algorithms were utilized to predict the existence of the corresponding company in the S&P 500 index. According to the strong machine learning results, the dataset generation methodology is successful and the datasets are ready for further use.

key words: *10-K filings, XBRL, EDGAR, data pre-processing, machine learning*

1. Introduction

Especially in the last two decades, financial data are analyzed for various purposes using machine learning (ML) methods. It has been reported that the quality, quantity, and relevance of ML studies in the financial sector will increase daily [1]. One reason is the documentation obligations regarding all activities of companies operating on domestic stock exchanges and public institutions. In addition, there are restrictions on complying with predetermined laws and regulations while sharing these documents. The shared documents contain essential data about companies and the markets themselves; in particular, investors, shareholders, and auditors are highly interested in these data and the information to be derived from them. Significant contributions of supervisory and supporting institutions, such as generally accepted accounting principles and U. S. Securities and Exchange Commission (SEC), internal and external auditors, and other stakeholders, increase the potential of studies in this field. When companies share files, they have to share them with file codes determined according to their content. For example, 8-K files are shared when there are unexpected material events or a significant change that concerns all stakeholders, whereas 10-Q files are the reports shared in each financial quarter, which explain all developments about a company in the relevant period. Indeed, the establishment of the XBRL file format enabled a machine-understandable standard for data analysis. SEC has obliged companies to upload XBRL formats while uploading their reports to the Electronic Data Gathering, Analysis and Retrieval (EDGAR) database, and these files are publicly accessible. All developments facilitate data analysis in the financial field.

In this study, a web crawler was developed to obtain 10-K filings reports as text files shared with the EDGAR database for 2019. Data preprocessing methods are proposed to prepare the obtained files for analysis with ML algorithms. The datasets obtained by the proposed methods were analyzed by K-nearest neighbor (KNN), random forest (RF), decision tree (DT), Adaboost, and quadratic discriminant analysis (QDA) methods. Results with Precision, Recall, F-Score, ROC Curve, and Accuracy metrics are presented in Sect. 5.

2. Background

With its tabular data structure, financial data can be represented in many different formats (e.g XBRL, pdf, HTML, xlsx). The digitalized financial reports allow researchers to perform financial analysis among companies, sectors and the entire market. However, it is not very common to collect financial reports over the EDGAR database and use them in analysis. Mostly, databases like COMPUSTAT, a database storing statistical, financial and market information of global companies operating worldwide, are used. For instance, Chychyla and Kogan [2] compared 10-K data with COMPUSTAT. 30 different accounting items of 5000 different companies were compared and analysed. It was revealed that there are significant differences in terms of shared information among different companies. Data processing efficiency in financial reports was calculated by Rao and Guo [3] in their work. Therefore, they compared the EDGAR and COMPUSTAT databases. They concluded that the size, age and industry of the company in data sharing are not very effective on data processing efficiency. In a different study, Cunningham and Leidner [4] improved the quality of the information available at SEC, in order to provide valuable information for investors. In this context, some

Manuscript received January 21, 2022.
Manuscript revised August 4, 2022.
Manuscript publicized October 12, 2022.

[†]The authors are with Computer Eng. Dept., KTO Karatay Univ., Turkey.
[††]The author is with Computer Eng. Dept., Konya Techinical Univ., Turkey.
a) E-mail: samikacar@gmail.com

DOI: 10.1587/transinf.2022IIP0001

deficiencies which decrease the quality have been identified. In addition to quantitative analyzes used in accounting and finance, the effectiveness of textual analysis was investigated by Loughran and McDonald [5]. As a different approach, Hoitash and Hoitash [6] proposed an accounting reporting complexity criterion over XBRL tags. The criterion aimed to determine the inaccuracies, deficiencies in the application, and the parameters affecting the reliability of the data. Peterson et al. [7] conducted a study to identify the accounting consistency measures for the same firm with the change over time and the change for different firms by analyzing the textual similarity of the accounting policies footnotes described in 10-K files. As a result of that study, it was concluded that the accounting consistency and accurate analyst estimates triggered stronger stock returns. Henselmann et al. [8] presented a system design aimed at identifying suspicious companies with abnormal numbers in XBRL files. They have been found that in general, businesses show higher earnings in order to encourage investors. Chen et al. [9] presented a machine learning approach to predict the annual earnings using XBRL formatted 10-K filings shared by the companies. The effect of the necessity on the comparison of financial reports of sharing them in XBRL format by the SEC was investigated by Dhole et al. [10]. In that study, it has been concluded that the comparison of financial reports and firm specific taxonomy has become more challenging in recent years. On the other hand, Li [11] analysed the contents of the forward-looking statements (FLS) part in the Management Discussion and Analysis section (MD&A) of the 10-K (annual) and 10-Q (quarter) reports using the Naive Bayes algorithm. The research revealed that the content was generally positive, and although many factors change over time, the content had not been changed accordingly. Studies have also been carried out on tag structures in XBRL formatted reports. Plumbee and Plumbee [12] reviewed XBRL assurance in financial reporting via tags. They found that many shortcomings and mistakes were implemented in the current situation. Loukas et al. [13] have studied XBRL tags and proposed a database with sentences consisting of tags. For this, they provided semantic validation and association of tags with natural language processing and machine learning methods. Chen et al. [14] tried to predict future earnings using decision tree method. For this, they used XBRL formatted 10-K filings as in this study. Moreover, they worked on arranging tags and finding correlations through tags.

3. Financial Datasets

This section describes the data sets used within the study. Section 3.1 explains the meta data of the database including the targeted users, the storage platform. The schema of the data used in this study is explained in 10-K filings (see 3.2). Finally, in 3.3, XBRL data format, which has become an important concept for the analysis of financial information is established.

3.1 SEC's EDGAR Database

SEC is an unaffiliated government agency that is intended for protecting investors, retaining a fair and regular market, and catalyzing capital formation. The SEC monitors federal security laws to ensure equality and freedom of information. To this end, the SEC requires public firms, capital managers, investment professionals, and all market stakeholders to regularly share relevant information so that investors can access accurate, up-to-date, complete, and reliable information. The SEC enables firms and entrepreneurs to create new business opportunities, and keep up with relevant changes. SEC describes EDGAR Database as "EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by firms and others who are required by law to file forms with the U.S. Securities and Exchange Commission". EDGAR is mainly intended for an efficient, fair, and confidential security market for investors, firms, and all other participators with storing and presenting high volume financial data. Except some reports and special conditions, all publicly traded firms are required to load their filings on EDGAR since 1996. Thus, there is inestimable raw data on EDGAR, which are waiting to be analyzed for extracting significant information. To access data on EDGAR, even registration is not required. Anyone can reach all records on EDGAR at any time.

3.2 10-K Filings

10-K filings are annual reports in which publicly traded firms disclose business and financial condition information and audited financial statements for the most recently completed two or three fiscal years. Firms are required to disclose their 10-K forms in two or three months according to their revenues. 10-K forms provide a wide perspective of a firms' business, the perils it faces, and operational and financial outcomes for the fiscal year.

These reports consist of 4 sections and 15 parts (see Table 1). Business is located in the first section of the first part. In this section, firms are obliged to write down all operations that occur during the year according to their field of activity. Clause 1A of the first section contains risk factors; in clause 1B, there are caveats for situations that independent auditors noted to be resolved in previous years but not resolved. The second item includes important facilities, features and physical assets owned by firms. In the third item, there are legal transactions and courts in which firms are involved. In the fourth and last section, there are occupational health and safety violations or other legislation articles of firms. In the second part, the share values in the stock exchange in which firms are traded, consolidated financial data, management's comments on the status of firms, expectations, market risks in the field of activity, financial statements and other information about firms are presented. The third and fourth sections contain a commitment to the accuracy of shared infor-

Table 1 Indicates the outline of 10-K filings with sections and subsections.

1. Section	1. Business
	1A Risk Factors
	1B Unresolved Staff Comments
	2. Properties
	3. Legal Proceedings
	4. Mine Safety Disclosures
2. Section	5. Market
	6. Selected Financial Data
	7. Management's Discussion and Analysis of Financial Condition and Results of Operations
	7A Quantitative and Qualitative Disclosures About Market Risk
	8. Financial Statements and Supplementary Data
	9. Changes in and Disagreements With Accountants on Accounting and Financial Disclosure
	9A Controls and Procedures
	9B Other Information
3. Section	10. Directors, Executive Officers and Corporate Governance
	11. Executive Compensation
	12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
	13. Certain Relationships and Related Transactions and Director Independence
	14. Principal Accounting Fees and Services
4. Section	15. Exhibits, Financial Statement Schedules

mation, signatures of firm officials and accountant evaluations. In this study, firms were analyzed on their financial statement data. Hence, the focus is on the data in the second part, more specifically data in Item 8, of 10-K forms.

3.3 XBRL (eXtensible Business Reporting Language)

XBRL is a general standard for business reporting, owned and reformed by XBRL International for the public interest. XBRL is used in more than 50 countries as an authorized digital reporting standard, thus millions of XBRL documents are created yearly. XBRL is named as "barcodes for reporting" because it constitutively aims to increase the efficiency of all kinds of business information by enhancing the skills such as preparation, validation, publication, exchange, consumption, and analysis while preparing documents. XBRL is an implementation of Extensible Markup Language, which is a specification to organize and define data on the internet. XBRL mainly tags all data in it, in other words, all values in a document are labeled by a specific tag created according to determined standards. Thus, reports are convenient while analyzing and managing data with machines. For all these reasons, XBRL is used by regulators, governments, firms, accountants, data managers, analysts, and investors.

In this study, the year-end report 10-K filings of 10711 companies operating in the US stock markets were scanned in the EDGAR database. The reports obtained were subjected to data preprocessing steps and transformed into matrices suitable for analysis. The S&P 500 index was used as a class label for companies to classify with ML methods. Two different classes were determined according to existence of companies in that index. The companies that traded in this index were allocated data for one class and the other for the other. Then, five different classification algorithms were run on the final datasets; the results were presented using five metrics. Although the methods proposed in the study are based on 10-K filings submitted by the SEC, they can also be applied to other reports in XBRL format. Moreover, it can be used in XBRL format financial reports in different countries. The tag-value data presentation structure in XBRL format significantly increases the adaptability and analyzability with algorithms. In Sect. 4, data collection and preprocessing steps are explained. The analyses performed are explained in detail, and the results obtained at the end of the previous step are evaluated in Sect. 5. Finally, the last section contains the conclusion of the paper.

4. Methodology

The methodology is composed of two main processes. Firstly, Data Collection Process (Sect. 4.1), the process of obtaining the data used in the study is explained. Researchers and investors can access a company's annual report through the SEC. Although these data can be accessed on a company basis, there is no collective presentation of the reports. Moreover, there is no possibility of obtaining only financial shares of companies. The method proposed in the study provides the opportunity to obtain the financial information drawn from company data in an aggregated way. Secondly, data pre-processing techniques developed for organizing, cleansing, parsing and merging the obtained data are explained in ten subtitles (Sect. 4.2). Indeed, forms that were highly complex and ambiguous when shared were transformed into simple, understandable and scalable datasets.

4.1 Data Collection Process

Within the scope of the study, the 10-K filings of 10711 firms operating on the NASDAQ stock exchange as of 2020, which have been uploaded to the EDGAR database were crawled and downloaded. In this context, a web crawler was designed and developed. The workflow of the process is illustrated in Fig. 1 (The process steps are numbered from 1 to 13). It was created to both identify the links of firms' web

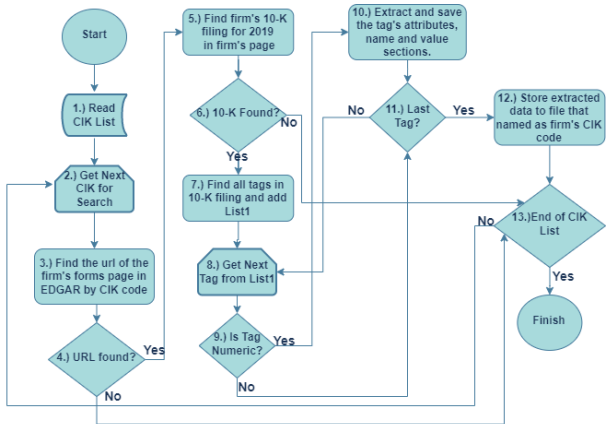


Fig. 1 Workflow of the data collection process.

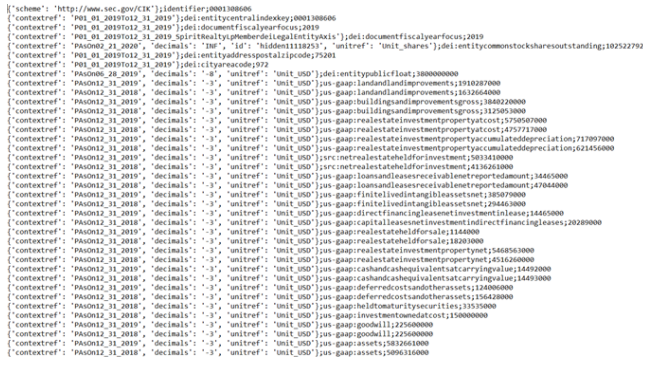


Fig. 2 An example of obtained and saved data before data prezprocessing steps.

Table 2 Period or date formats used in XBRLs of 10-K filings.

Period or Date Information
FD2019Q4YTD
D20190101-20191231
Duration_1_1_2019_To_12_31_2019
As_Of_12_31_2018_us-gaap_Statement
iee3fce9cc0c446609a421a12d16_I20200117
From2019-01-01to2019-12-31
From_January_to_December
Duration_01_January_2019_To_31_December_2019

pages on the SEC website, where all filings of firms are presented, and retrieve data from that website. First, read the CIK codes of the companies given specifically to each company from the previously prepared file (See Step 1). Loop with CIK codes for firm based search and extraction (See Step 2). Request the URL in the EDGAR database of all the forms each company has shared with the SEC (See Step 3). Check if the URL is found (See Step 4). If found, proceed to the next step. If not found, go to end of list check step (See Step 13). Among the company forms, find the 10-K filings in XBRL format shared by the company for 2019 (See Step 5). Check if the form was found (See Step 6). If found go to next step, if not found go to end of list check step (See Step 13). Find and list tags in 10-K filing (See Step 7). Loop through tag list (See Step 8). Check if each tag is numeric (See Step 9). If it's numeric, go to the next step. If not, go to tag list break check step (See Step 11). Extract and save the tag's attributes, name and value information (See Step 10). Check the tag list if the examined tag was the last or not (See Step 11). If the answer is yes, go to the next step. otherwise, go to (See Step 8). Write all the captured information to the file kept with the name of the company (See Step 12). Finally, check if it's the last CIK (See Step 13). If it is positive, finish it, if not, go to Step 2 and repeat the same steps for the next company.

4.2 Data Pre-Processing

Each of the files obtained in the data collection part was converted into comma-delimited file type (csv) formatted files. An example of a file that contains obtained data is shown in Fig. 2. Later, the following processes were applied to uncover information from the raw data. First, the yearly information from the attributes of all tags was obtained (see 4.2.1). After retrieving the years, lines without values for tags (see 4.2.2), without the same values for the same tags in different years (see 4.2.3), and with 0-1 values for related tags were removed (see 4.2.4). Tag repetition (see 4.2.5) and tag prefixes (see 4.2.6) were also removed. Later, tags that potentially have the same meaning were extracted (see 4.2.7). Then, tag usage percentage in files were detected (see 4.2.8). Next, documents, after the performed processes, were converted to data sets (see 4.2.9). Finally, class labels were created for samples according to the S&P500 index (see 4.2.10). In the following sections, all processes are described in detail.

4.2.1 Extract the Year from Attributes of Tags

Firms clarify the attributes of a tag with the source reference label (contextref) in their XBRL files. These labels include the year, decimal number, currency information of a tag. First, other features were deleted, except for the date information. Because, in this study, financial reports which have USD as currency at monetary sections were analyzed, and monetary information shared in XBRL files does not include long or short scales. Thus, only the date information is left in the contextref section. However, because not all firms share date information in the same way, obtained the date was difficult. Some of the forms to remark the date information in attributes are given in Table 2. As can be seen from the table, firms use very different forms for the date data. Although some are numeric, some are alphanumeric, and some are textual expressions, it was possible to retrieve date information in all forms in the table. However, some lines were deleted to avoid ambiguity due to the unclear time information, e.g., although the period of a tag is annual, given date range covers more than one year, or more than one period information is mentioned for a single tag. At the end of this section, structured files containing lines with date information, name, and value of a tag were obtained.

4.2.2 Remove Lines from Files without Values for Tags

At this stage, the lines obtained in the previous stage were examined. Lines in a file must have a value for each tag in the respective year. For example, there must be a numerical value indicating the status of a firm's assets corresponding to the "asset" tag. However, some lines contain incomplete information because of incomplete or incorrect data entries or because the tags of information that are not obligated to be shared are entered, but their values are left blank. Owing to lack of information, such lines were detected and removed from files.

4.2.3 Remove Lines with the Same Values for the Same Tags in Different Years

The lines that passed through the previous stages were compared at this stage on a date basis. Lines of a firm's file with the same value but different year information for the same tag name were deleted. Any shared item of a firm with the same value for different years may be questioned, but when the files are examined, it was observed that this might cause ambiguity and reduce the consistency of data. Considering that the aforementioned information might have been entered incompletely or incorrectly, the relevant lines were removed from the files. In addition, if there was the same tag belonging to different years in any file, first, the tag information related to the year it was examined will be obtained, so removing the tags with different date information with the same value did not cause any shrinkage in data.

4.2.4 Remove Lines That Have 0-1 Values for Related Tags

Lines containing 0 or 1 in the value information given for the relevant tags were removed from the files at this stage. When the lines containing 0 or 1 values and related tags in the files were examined, it was observed that some of them did not represent the financial information of the firms, and some might be related to incomplete or incorrect value entries. Thus, the lines that were irrelevant for this study, those entered incompletely or incorrectly, those that could cause complexity, and those that could negatively affect the results of the analysis were removed, and the files consisting of lines consisting of the year, tag name, and value represented by the tag were saved with the csv extension. In Fig. 3, a section from a file that has been saved through the steps mentioned above is shown.

4.2.5 Remove Repetitive Tags

As shown in Fig. 3, information regarding the previous or next year is also shared in 10-K filings. Generally, firms share the data of the last three years to define comparisons in the financial statements. Therefore, data can be included up to three years before the year examined. In addition, since

```
2018;us-gaap:accruedliabilitiescurrent;12276000
2019;us-gaap:accruedincometaxescurrent;1092000
2018;us-gaap:accruedincometaxescurrent;737000
2019;us-gaap:liabilitiescurrent;21280000
2018;us-gaap:liabilitiescurrent;22011000
2019;us-gaap:longtermdebtnoncurrent;74929000
2018;us-gaap:longtermdebtnoncurrent;76129000
2020;us-gaap:longtermdebtnoncurrent;69900000
2019;us-gaap:otherliabilitiesnoncurrent;158000
2018;us-gaap:otherliabilitiesnoncurrent;178000
2019;us-gaap:liabilities;99151000
2018;us-gaap:liabilities;98566000
2019;us-gaap:limitedpartnerscapitalaccount;37334000
2018;us-gaap:limitedpartnerscapitalaccount;34677000
2019;us-gaap:partnerscapital;53172000
2018;us-gaap:partnerscapital;50678000
```

Fig. 3 Shows the date;tag;value lines, after first data cleansing processes.

10-K reports are shared within the first three months of each fiscal year and some file updates are made on them afterward, they may also contain information for the next year following the relevant year. The tag name and value of that year were preferred, and others were removed from the files. However, when the files were examined, it was observed that some tags represented only an item in the past period, whereas some tags contained predicted information for the next year, such as future expectations. Therefore, if there is no value for the tag in the relevant year, the values for the previous or next year are taken according to their suitability. After this step, tag repeats were extracted, and independent tags were obtained in the files.

4.2.6 Remove Tag Prefixes

In XBRL file format, three different types of prefixes are used: 'us-gaap', ticker symbol (the firm's stock market code) and dei code. The 'us-gaap' is the initial prefix used in approximately 17.000 different tags, each of which has a special meaning, determined by the GAAP institution [15]. Firms share their own tags by defining their stock codes as prefixes as well. The word "dei" means data entry instructions. In the XBRL files, firms additionally share information such as year, city code, central index key with this prefix. In order to clean and simplify the features to be used, those prefixes were removed. By removing the prefixes, semantically same features were matched among different financial reports. Thus, tags with the same name except the prefix are accepted as the same tag. An example of file contents after data cleansing steps shown in Fig. 4.

4.2.7 Extract Tags That Potentially Have the Same Meaning

The correlation between different features is a common way of cleaning duplicated values from a data set. In this study, the similarity of values within different feature sets are investigated. In the data set, the values for the firms can share the income or expenses with different names. For instance,

Table 3 Tag pairs that were used for the same value in different documents and repetition number of that.

Tag-1	Tag-2	Number of repetitions
assets	liabilitiesandstockholdersequity	2880
commonstocksharesissued	commonstocksharesoutstanding	1295
adjustmentstoadditionalpaidincapitalsharebasedcompensation...	sharebasedcompensation	964
definedbenefitplanbenefitobligationbenefitspaid	definedbenefitplanplanassetsbenefitspaid	607
commonstocksharesissued	sharesoutstanding	492
commonstocksharesoutstanding	sharesoutstanding	488
paymentsforrepurchaseofcommonstock	treasurystockvalueacquiredcostmethod	432
proceedsfromstockoptionsexercised	stockissuedduringperiodvaluestockoptionsexercised	414
financeleaseinterestexpense	financeleaseinterestpaymentonliability	339
paymentsforrepurchaseofcommonstock	stockrepurchasedduringperiodvalue	288

```
2019;incometaxexpensebenefit;19076000
2019;profitloss;64656000
2019;netincomelossattributabletononcontrollinginterest;3094000
2019;netincomeloss;61562000
2019;comprehensiveincomenetoftaxattributabletononcontrollinginterest;3083000
2019;comprehensiveincomenetoftax;61269000
2019;weightedaveragenumberofsharesoutstandingbasic;20662750
2019;weightedaveragenumberofdilutedsharesoutstanding;20852361
2019;cashandcashequivalentsatcarryingvalue;27392000
2019;accountsreceivablenetcurrent;27961000
2019;inventorynet;67768000
2019;prepaidexpensecurrent;4530000
2019;assetscurrent;127651000
2019;propertyplantandequipmentnet;65756000
2019;goodwill;51404000
2019;intangibleassetsnetexcludinggoodwill;146061000
2019;deferredincometaxassetsnet;60407000
2019;otherassetsnoncurrent;35000
2019;assets;451314000
2019;accountspayablecurrent;21174000
2019;accruedliabilitiescurrent;49097000
2019;incometaxanddistributionspayablecurrent;1469000
2019;payablepursuanttotaxreceivableagreementcurrent;3592000
2019;liabilitiescurrent;75332000
2019;deferredincometaxliabilitiesnet;145000
2019;otherliabilitiesnoncurrent;1689000
```

Fig. 4 After cleansing steps, contents of files occured from date-tag-value of each item in 10-K filings.

indirect cost and non-production costs are being used for the same expenses. Additionally, some of the accounting items may balance each other as plus and minus, which results in the same value in the balance sheet. Moreover, sum of liabilities and stockholders' equity is equal to the total assets. Majority of the files contain 'assets' tag together with 'liabilitiesandstockholdersequity' tag containing the same values. Similarly, different named tags with the same value exist in the files. In order to identify those tags, First of all, different tags with the same value are identified. Then, the aggregated number of those items were counted. Tags with mostly the same value have been filtered. A sample of detected tags are given in Table 3.

4.2.8 Detect How Many Percent of All Files a Tag Is in

Common tags have been searched in the files subjected to the previous steps. No tag used in all 3.114 files was found. Then, it was researched again with various filters. For example, common tags were searched for files containing at least 100 or 200 tags or 500 files containing at most tags, but still the common tag number did not exceed five. Thus, in order to inference the usage frequency of the tags, it is calculated what percentage of all files each tag is used. Table 4 contains the most preferred 10 tags and their percentages.

Table 4 Top 10 most used tags by firms in 10-K filings.

Tag Name	Percentage
entitycommonstocksharesoutstanding	97.31
entitypublicfloat	95.08
liabilitiesandstockholdersequity	94.95
assets	94.82
stockholdersequity	87.76
cashandcashequivalentsatcarryingvalue	85.95
propertyplantandequipmentnet	84.66
accumulateddepreciationdepletionandamortization..	81.48
operatingleaserightofuseasset	80.54
operatingleaseliability	79.31

Table 5 Indicates for the limits,which shows the tags usage percentage in documents, how many tags were used in feature vectors.

Tags Used in Documents of at Least%	Number of Tags
30	167
40	119
50	83
60	43
70	22
80	8
90	3

4.2.9 Convert Documents to Data Sets

In the previous stages, the data that could be irrelevant for this study, those inaccurate or incomplete, and those that could cause confusion were cleared from the dataset. At this stage, the data that separated into documents were converted into predictive analysis datasets to be run in ML algorithms. In total, eight different datasets were produced. They were produced according to the tags whose usage frequency percentages in the documents were determined. First, the lower limit of the tag usage percentage was set. Since algorithms with lower values produced unbalanced results, we started with the tags used in at least 30% of the documents as the lower limit, and seven datasets with 10% increments up to 90% were created and named as "30DS", "40DS", "50DS", "60DS", "70DS", "80DS", "90DS". Accordingly, the tag percentages in the created files and the number of tags used as feature vectors in the datasets are shown in Table 5. Thus, each row represents the corresponding values of a firm for the specified tags, or 0 if the tag is not in the document, so each column also represents the value for that tag, if it has

value for the corresponding firm in the row, and 0 if it does not. The template of the datasets created is shown in Fig. 5.

The reason for entering 0 in places where there is no value in all dataset formation processes was that methods, such as column average, writing the most frequently repeating value, and writing the maximum or minimum value, negatively affected the performance of the algorithms. Only one of the tags, which were identified and grouped as mostly the same tags in the previous section, was kept in the final datasets to prevent data repetition, and the rest of the tags were removed. 25 rows that represent 25 firms' data and contain mostly missing data in the datasets created were also removed at this stage. Finally, when all data were examined, 125.000 independent tags were determined. A dataset of 125.000 columns, where detected tags were used as features, was also prepared for analysis and named "AllTags". There was an average of 60 tags in the files obtained after the cleaning processes. Therefore, in the last dataset obtained, approximately 60 values were meaningful, and the others were 0 in each row corresponding to 125.000 columns. As a result, after this stage, eight predictive analysis datasets were obtained with rows representing 3.089 firms and different number of columns.

4.2.10 Create Class Labels According to S&P 500 Index

The methods utilized to examine, clean, complete, and convert the 10-K Filings into a machine learning data set are explained in the previous sections. In order to measure how the created data set can be used for further data mining analysis, state-of-the-art classification algorithms were utilized. As the class label, the well known S&P 500 index was used to create a binary classification problem, identifying if the corresponding firm is in the index or not. The S&P 500 index is not just a list of companies in U.S. sorted by their market caps. It also shows a lot of information such as prominent American equities' and stock market overall [16]. Moreover, it affects many values such as price strength, volatility index [17]. Thus, the classification of companies according to the S&P 500 index from their annual reports may present new insides to researchers and investors.

As a result of the cleaning, filtering, and class label creation processes, 7 different data sets of 3.089 rows with different column numbers were obtained. 404 of the firms belong to companies in the S&P 500, while the remaining 2.685 do not.

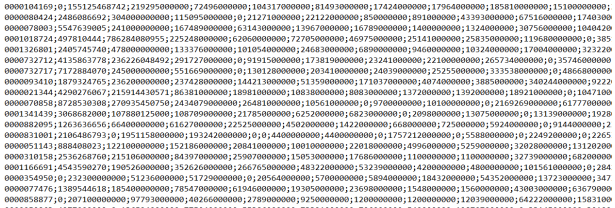


Fig. 5 Shows a section from one of the obtained data sets.

5. Results and Discussion

In this section, the results of the analyses performed in this study are given. In addition, the obtained results are interpreted.

5.1 Analyses

In the previous stages, the data were made ready for analysis by preprocessing techniques. At this stage, the data extracted from 10-K filings were analyzed in terms of the ability to represent the belonged firm, how useful the preprocessing techniques were, whether the most valuable companies in the market could be distinguished from the others with only 10-K filings information, and the difference between a large amount of data and an extraction on that data. For this, all eight datasets obtained in the previous stages were analyzed with five ML algorithms: KNN, DT, RF, Adaboost, and QDA [18]. In the algorithm selection, we aimed to select algorithms that use different mathematical methods while classifying the samples, and thus to determine the effectiveness of the obtained data in different methods and approaches. 10-fold cross-validation was used to resample the data, and before analyses, the datasets were shuffled [19]. The distribution of the data samples in a dataset according to the class labels, and accordingly, the division into 10 equal parts for each fold is given in Fig. 6.

Accuracy, Precision, Recall, F-Score given in formula 1–4, and ROC Curve metrics were used to evaluate the algorithms [20]. The results were calculated separately for each fold in terms of all metrics, and the final average values and confidence interval deviations were also extracted. Since the dataset was imbalanced according to the specified class tags, the macro average results of all metrics were extracted. All metrics can be extracted from the confusion matrix given in the Table 6. Accordingly, the number of true positives shows how many of the companies in the S&P 500 index

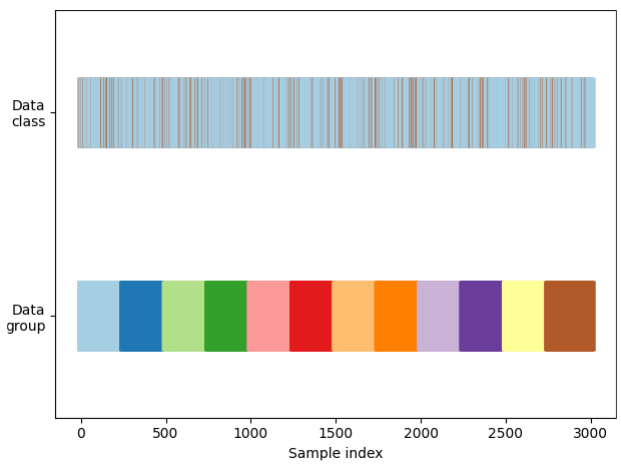


Fig. 6 Shows data distribution in data sets according to classes for 10 fold cross validation.

Table 7 Indicates the all performance results for eight different data sets in 5 different algorithms' analyzes corresponding to five different metrics.

Data Sets	Metrics	KNN	CI 95%	Decision Tree	CI 95%	Random Forest	CI 95%	Adaboost	CI 95%	QDA	CI 95%
30DS	**Accuracy**	0.9236	0.0073	**0.9317**	0.0087	0.9414	0.0075	**0.9433**	0.0064	0.899	0.0105
	Precision	**0.8589**	0.0244	**0.8457**	0.0247	0.88	0.0237	**0.8829**	0.0138	0.7759	0.0236
	Recall	0.7768	0.0211	**0.8516**	0.0246	**0.8511**	0.0288	**0.8591**	0.0193	**0.8792**	0.0216
	F-Score	0.8088	0.0204	**0.848**	0.0237	0.8626	0.0231	**0.8698**	0.0152	0.8121	0.0228
40DS	**Accuracy**	0.9223	0.0054	0.9249	0.01	0.9414	0.0107	0.9408	0.0052	0.9058	0.0115
	Precision	0.8549	0.0118	0.8387	0.0267	0.8867	0.0277	0.8803	0.0118	0.7872	0.0222
	Recall	0.7747	0.0159	0.8228	0.025	0.8483	0.0233	0.8515	0.0208	0.8639	0.0209
	F-Score	0.8067	0.0132	0.8297	0.0243	0.8646	0.0233	0.8635	0.0138	**0.8176**	0.0222
50DS	**Accuracy**	0.9246	0.0096	0.923	0.0101	**0.9433**	0.0046	0.9398	0.0069	**0.9107**	0.0121
	Precision	0.856	0.0254	0.8381	0.025	**0.8911**	0.0207	0.8787	0.0206	0.7985	0.026
	Recall	0.7888	0.0212	0.821	0.025	0.8474	0.0166	0.8521	0.0125	0.8433	0.0215
	F-Score	0.8164	0.0214	0.8268	0.0201	**0.8663**	0.0151	0.8633	0.0117	0.8171	0.0235
60DS	**Accuracy**	0.9243	0.0116	0.9126	0.0066	0.9408	0.0071	0.9401	0.0076	0.9081	0.0128
	Precision	0.8516	0.0247	0.8055	0.022	0.8822	0.0172	0.8754	0.0193	0.8054	0.0255
	Recall	0.7995	0.0166	0.8084	0.0187	0.848	0.0094	0.8558	0.0158	0.7797	0.0242
	F-Score	0.822	0.0189	0.8058	0.0185	0.8635	0.0116	0.8641	0.0147	0.7897	0.0227
70DS	**Accuracy**	0.9242	0.0084	0.9142	0.0073	0.9382	0.0062	0.9343	0.008	0.9071	0.0097
	Precision	0.8511	0.0146	0.812	0.0156	0.8716	0.0143	0.8602	0.015	**0.8227**	0.0287
	Recall	0.7975	0.0186	0.8103	0.0198	0.8453	0.0199	0.8442	0.0275	0.7268	0.0176
	F-Score	0.82	0.0156	0.8099	0.0153	0.8572	0.0165	0.8502	0.0201	0.7616	0.02
80DS	**Accuracy**	**0.9252**	0.0074	0.9126	0.0109	0.9388	0.0061	0.9298	0.0077	0.898	0.0093
	Precision	0.8515	0.0211	0.8098	0.0312	0.8799	0.0177	0.8552	0.0167	0.8281	0.0274
	Recall	0.8015	0.0229	0.8055	0.0227	0.836	0.0204	0.8234	0.0249	0.6547	0.0277
	F-Score	0.8217	0.0193	0.8062	0.026	0.8553	0.0179	0.8367	0.0197	0.6955	0.0339
90DS	**Accuracy**	0.9252	0.0057	0.9077	0.0067	0.9255	0.0042	0.9301	0.0061	0.8705	0.0125
	Precision	0.8461	0.0122	0.7967	0.0185	0.8485	0.0169	0.8559	0.0201	0.7049	0.0973
	Recall	**0.8056**	0.0157	0.7833	0.0162	0.8002	0.0184	0.8223	0.0116	0.5199	0.0093
	F-Score	**0.8234**	0.0134	0.7895	0.0168	0.821	0.0165	0.8372	0.0137	0.5058	0.018
AllTags	**Accuracy**	0.9132	0.0061	0.9149	0.0065	0.7941	0.006	0.8708	0.0075	0.1327	0.0089
	Precision	0.8491	0.0183	0.8117	0.0211	0.6087	0.0174	0.8807	0.0239	0.382	0.1343
	Recall	0.7265	0.0172	0.815	0.0207	0.6492	0.0148	0.5062	0.0197	0.4995	0.0037
	F-Score	0.7684	0.0174	0.8116	0.0169	0.6212	0.0155	0.4776	0.0176	0.1182	0.0069

Table 6 Shows the confusion matrix, which indicates the distribution of the actual classes with the predicted classes of the algorithm

	PREDICTED CLASS	
TRUE CLASS	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

correctly predicted the algorithm. The number of true negatives shows how many of the companies that are not actually in this index are correctly identified by the algorithm. The false positive and negative values represent, respectively, the number of companies that the algorithm predicts to be in this index although they do not actually have the S&P 500 index, and the number of companies that the algorithm predicts in this index even though they are actually in the S&P 500 index. In summary, the dataset obtained with the proposed data mining methods for evaluating the companies through 10-K filings was analyzed with five methods in terms of five metrics, and the results were obtained 7. The best results for the datasets are shown with underline, and the best results for the algorithms are shown in bold. In addition, ROC curves of the best and the worst 2 results are given in Fig. A· 1 at Appendix.

Accuracy:

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

Precision:

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

Recall:

$$Recall = \frac{TP}{TPFN} \tag{3}$$

F-Score:

$$F - Score = 2x\frac{PrecisionxRecall}{Precision + Recall} \tag{4}$$

5.2 Evaluation of Results

The dataset obtained from 10-K filings files downloaded from the EDGAR database provided by SEC, where companies disclose their information for the year completed in the first three months of each fiscal year, was analyzed. Eight datasets with different numbers of feature vectors were evaluated over five metrics with five ML methods. Commodity computer with 16-GB Ram, Intel i7 processor was used in the analysis. The average run times for all datasets used in analyses via algorithms with the 10-fold cross-validation method are given in Table 8. As expected, the run times were long when analyzing the dataset consisting of only 125.000 columns, whereas others completed the analysis in relatively short periods.

When the table was examined, successful results were

Table 8 Indicates runtimes of algorithms for each data set in seconds.

Data Sets	KNN	Decision Tree	Random Forest	Adaboost	QDA	Total
30	0.281234	0.218765	1.359348774	0.149994	0.149994	2.159335
40	0.176559	0.174996	1.447965407	1.393907	0.104681	3.298108
50	0.126557	0.198438	1.46185174	1.09846	0.059373	2.94468
60	0.082812	0.079683	0.812650275	0.612494	0.02812	1.61576
70	0.0375	0.048437	0.665624929	0.385936	0.010936	1.148433
80	0.02344	0.023437	0.450001836	0.220315	0.00625	0.723443
90	0.01719	0.0125	0.306246757	0.157811	0.006249	0.499997
AllTags	3.982077	21.84333	29.72249475	416.8544	156.2111	628.6134

seen in all datasets, except the ALLTags dataset. In addition, there was an insignificant difference between the ML methods used in the analysis. However, the dataset that has all independent tags as feature vectors had unsuccessful results for almost all algorithms. In the remaining seven datasets, where feature selection steps were used, the results were quite close to each other. When looking at the accuracy values of the algorithms in the ALLTags dataset, the difference between the best and the worst performance was 78%, whereas the difference was only 5% in the other seven datasets. Moreover, when looking at the fit and score times of the algorithms, Adaboost was the slowest. While the average total running time of the ALLTags dataset for each fold was 416 s for Adaboost, the highest average total time among all results in the other group was less than 2 s. According to these results, selecting some data shared by the firms increased the performance in classification problems, which was common for firms, whereas a major part of the identified tags was firm specific. RF was the most successful algorithm in the seven datasets for all metrics, followed by Adaboost. The most unsuccessful pair of algorithms for this dataset were KNN and QDA. In the ALLTags dataset, no algorithm, except DT, achieved successful results in all metrics. 95% was chosen as the confidence interval value. In this range, it is seen in Table 7 that deviation values were low in all metrics and algorithms. Thus, it was revealed that there were no significant differences between the values in 10 different folds. Looking at the seven datasets, where feature selection was applied, the most successful results were obtained in 30% and 50% datasets, but there was no significant difference in general. Because, as the percentages increase, although the number of tags representing the samples decreases, the number of missing information also decreases. Thus, although low percentage tags allow for generalization, high percentage tags make it easier to distinguish. It was concluded that using tags with a low percentage in the files on algorithms that achieve better results with more data and selecting the tags with a high percentage in algorithms that prefer strong distinguishing features among samples, albeit a little, will increase the classification accuracy. The most unsuccessful results in this group were also mostly in the 90% dataset. This is also expected because the 90% dataset had only three tag feature vectors. Nevertheless, these three tags could produce successful results in classification since they were included in at least 90% of the 10-K filings examined, and therefore their representation abilities were high. As a result, it was proven that, by processing the raw firm data downloaded from the EDGAR database with the methods described in this study, they could be transformed into datasets that could achieve successful results with ML methods; thus, publicly listed companies could be analyzed through the 10-K filings files they regularly share.

6. Conclusion

In this study, a web crawler designed to extract 10-K reports shared by public companies from the EDGAR database and data preprocessing methods analyzed by algorithms are presented. After the preprocessing steps, eight different datasets containing the reports of 3089 companies were obtained. The resulting datasets were analyzed using KNN, RF, DT, Adaboost, and QDA algorithms. The performance of the algorithms was measured with Accuracy, Precision, Recall F-Score, and ROC Curve metrics. Although many firms use their tags when sharing data in 10-K reports, successful results obtained by classifying firms with ML methods showed that common tags represent firms well. Highest scores were obtained by RF algorithm with 94% accuracy, 0.89 precision, 0.84 recall and 0.86 F-Score values on 50DS data set that was created with tags which were used in the half of all reports. Thus, it was shown that, even from a small part of data in the annual reports of firms, valuable information could be extracted with ML methods. Moreover, the methods proposed in this study will pave the way for researchers and investors to work on many issues such as specific tag groups (e.g tags related to 'Assets'), balance sheets, the effect of 10-K filings on stock markets, and information verification. Finally, since the biggest challenge is to establish a context between specific tags shared by different companies in the data preprocessing steps, it is concluded that examining tags with semantic studies, profiling of the companies with combining social media data and financial reports [21], and matching different tags will be a relevant field for future studies.

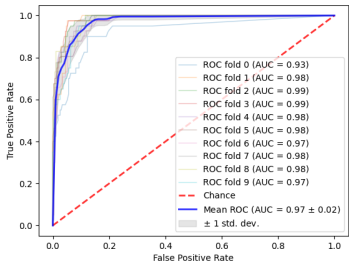
References

[1] D. Appelbaum, A. Kogan, and M.A. Vasarhelyi, "Big data and analytics in the modern audit engagement: Research needs," Auditing, vol.36, no.4, pp.1–27, 2017.

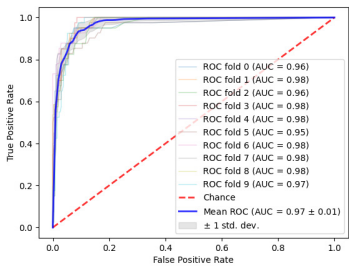
[2] R. Chychyla and A. Kogan, "Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in compustat and SEC 10-K filings," Journal of Information Systems, vol.29, no.1, pp.37–72, 2015.

[3] Y. Rao and K.H. Guo, "Does XBRL help improve data processing efficiency?," International Journal of Accounting and Information Management, vol.30, no.1, pp.47–60, 2022.
[4] L.M. Cunningham and J.J. Leidner, The SEC Filing Review Process: Insights from Accounting Research, 2019.
[5] T. Loughran and B. Mcdonald, "Textual Analysis in Accounting and Finance: A Survey," Journal of Accounting Research, vol.54, no.4, pp.1187–1230, 2016.
[6] R. Hoitash and U. Hoitash, "Measuring accounting reporting complexity with XBRL," Accounting Review, vol.93, no.1, pp.259–287, 2018.
[7] K. Peterson, R. Schmardebeck, and T.J. Wilks, "The earnings quality and information processing effects of accounting consistency," Accounting Review, vol.90, no.6, pp.2483–2514, 2015.
[8] K. Henselmann, D. Ditter, and E. Scherr, "Irregularities in accounting numbers and earnings management—a novel approach based on SEC XBRL filings," Journal of Emerging Technologies in Accounting, vol.12, no.1, pp.117–151, 2015.
[9] X. Chen, Y.H. Cho, Y. Dou, and B.I. Lev, "Fundamental Analysis of XBRL Data: A. Machine Learning Approach," SSRN Electronic Journal, 2022.
[10] S. Dhole, G.J. Lobo, S. Mishra, and A.M. Pal, "Effects of the SEC's XBRL mandate on financial reporting comparability," International Journal of Accounting Information Systems, vol.19, pp.29–44, 2015.
[11] F. Li, "The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach," Journal of Accounting Research, vol.48, no.5, pp.1049–1102, 2010.
[12] R.D. Plumlee and M.A. Plumlee, "Assurance on XBRL for financial reporting," Accounting Horizons, vol.22, no.3, pp.353–368, 2008.
[13] L. Loukas, M. Fergadiotis, I. Chalkidis, E. Spyropoulou, P. Malakasiotis, I. Androutsopoulos, and G. Paliouras, "FiNER: Financial Numeric Entity Recognition for XBRL Tagging," pp.4419–4431, 2022.
[14] X. Chen, Y.H. Cho, Y. Dou, and B. Lev, "Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data," Journal of Accounting Research, vol.60, no.2, pp.467–515, 2022.
[15] XBRL US, "2020 US GAAP financial and SEC reporting taxonomies," https://xbrl.us/xbrl-taxonomy/2020-us-gaap/, 2020.
[16] S&P Global, "Index Attributes," tech. rep., 2020.
[17] E.T.J. Chen and A. Clements, "S&P 500 implied volatility and monetary policy announcements," Finance Research Letters, vol.4, no.4, pp.227–232, 2007.
[18] C.C. Aggarwal, Data Classification Algorithms and Applications, Chapman and Hall/CRC, 2554.
[19] D. Berrar, "Cross-validation," Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, vol.1-3, no.April, pp.542–545, 2018.
[20] N. Lavesson and P. Davidsson, "Evaluating learning algorithms and classifiers," International Journal of Intelligent Information and Database Systems, vol.1, no.1, pp.37–52, 2007.
[21] K. Oztoprak, "Subscriber profiling for connection service providers by considering individuals and different timeframes," IEICE Trans. Commun., vol.E99-B, no.6, pp.1353–1361, 2016.

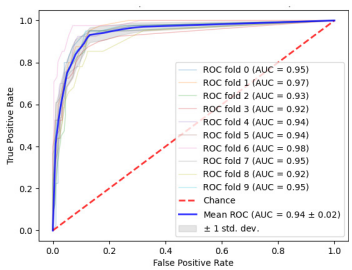
Appendix: ROC Curves of Test Results



(a)

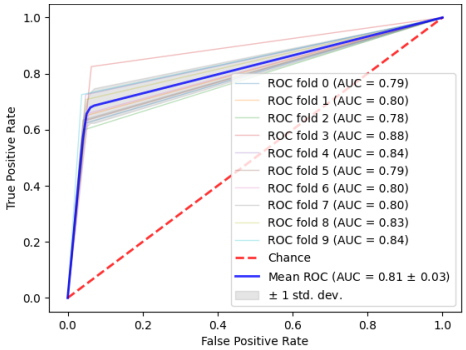


(b)

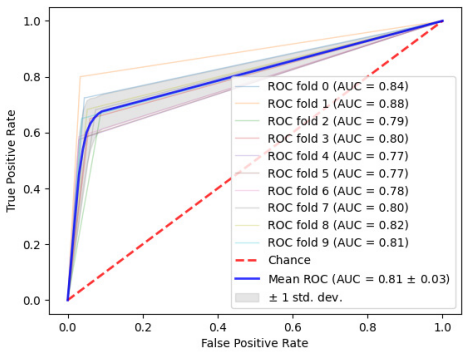


(c)

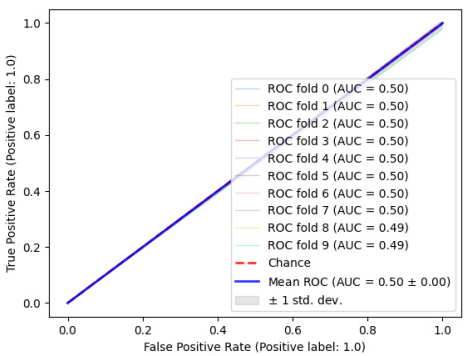
Fig. A· 1 ROC curves of best 3 results. (a) Random forest algorithm for 30DS dataset. (b) Random forest algorithm for 50DS dataset. (c) K nearest neighbor algorithm for 80DS dataset



(a)



(b)



(c)

Fig. A· 2 ROC curves of the worst 3 results. (a) Decision tree algorithm for 30DS dataset. (b) Decision tree algorithm for 90DS dataset. (c) QDA algorithm for AllTags dataset

Mustafa Sami Kacar earned the B.S. degree from Computer Engineering Department of Cankaya University and M.S. degree from Electrical and Computer Engineering program of KTO Karatay University in 2013 and 2017, respectively. He holds a Ph.D. in Computer Engineering at Konya Technical University. His research interests are Artificial Intelligence, Data Mining and Machine Learning.

Semih Yumusak holds a B.S. in Computer Engineering from Koc University (2005), an MBA from Istanbul Bilgi University (2008), and a PhD in Computer Engineering from Selcuk University (2017). With experience as a researcher at The Insight Centre for Data Analytics in Galway, Ireland (2015–2016) and AI4BD AG in Switzerland (2016–2021), he now serves as an Assistant Professor of Computer Engineering at KTO Karatay University. His areas of expertise include the Semantic Web, Linked Data, Web Mining, Data Mining, and Graph Algorithms.

Halife Kodaz graduated from Computer Engineering Department of Selcuk University with B.S. degree and M.S. degrees in 1999 and 2002, respectively. He received the Ph.D. degree in Electrical and Electronics Department from Selcuk University in 2008. He is a Professor at the Computer Engineering Department at Selcuk University. His research interests are artificial intelligence and machine learning.