Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 1 PAGE: 17-24 YEAR: 2023
DOI:10.17798/bitlisfen.1155725

Predictive Analysis Using Web Scraping for the Real Estate Market in Gaziantep

Ali Can ÜZÜMCÜ[1], Nazmiye ELİGÜZEL[2*]

[1] *Gaziantep University, Industrial Engineering, 27310 Gaziantep, Turkey*
[2] *Gaziantep Islam Science and Technology University, Industrial Engineering, 27010 Gaziantep, Turkey (ORCID: 0000-0002-8089-2285) (ORCID: 0000-0001-6354-8215)*

Keywords: Gaziantep, Machine learning, Real estate, Web scraping.

Abstract
For investors and people who want to own a property, real estate is a crucial industry. Real estate includes land and any enduring construction, whether natural or artificial, such as houses, residences, apartments, and commercial structures. In Turkey, it is common to believe that owning property makes you live comfortably. Therefore, house ownership is a common aspiration among Turkish families. However, a variety of factors, such as a country's economic structure, inflation, world events, politics, etc., have an impact on the real estate market. In addition, the location, neighborhood, size, and number of rooms of a house can all affect how much it costs to live there. Gaziantep city is considered for analysis in the proposed study. The goal of this study is to predict which neighborhood, given a prospective buyer's financial status and specific property attributes, someone can afford to live in. As a result, web scraping is used to collect real estate data from the website. Once the data has been gathered, forecasting the neighborhood of a house is done using machine learning algorithms including decision trees, random forest, and extra trees. The results demonstrate that all algorithms produce good results with a performance accuracy of over 80%. However, among these algorithms, decision tree classification offers the best performance.

1.Introduction

Gaziantep city is located in South-Eastern Anatolian Region in Turkey. The province's 6554 km2 territorial area represents about 1% of Turkey's total land area and there are about 2130432 people living there[2]. It has nine districts. Gaziantep is also one of the Turkey's most expensive cities. One of the most important investments in most people's portfolios is real estate. Real estate includes land and any enduring construction, whether natural or artificial, such as houses, apartments, commercial structures, and fences [1]. In 2021, there were 1491856 housing sales in Turkey, a 0.5 percent decline from the previous year. With 276223 homes and 18.5 percent of the market, Istanbul had the biggest percentage of house sales. Ankara, with 144104 house sales and a 9.7 percent market share, and Izmir, with 86722 house sales and a 5.8 percent market share, followed Istanbul in that order. The provinces with the lowest number of house sales were Hakkari (267), Ardahan (377), and Bayburt (871), in that order. Gaziantep, on the other hand, ranked 10th with 35610 houses [3]. Figure 1 provides information regarding 2021 housing sales data. House prices in Gaziantep have increased significantly, from just 145661TL/m^2

* Corresponding author: nazmiye.eliguzel@gibtu.edu.tr
Received: 04.08.2022, Accepted: 01.03.2023

[2] http://www.gaziantep.gov.tr/

[3]https://data.tuik.gov.tr

in 2014 to over 721839 TL/m^2 in 2022[4]. In Figure 2, the trend of change in house prices is given.

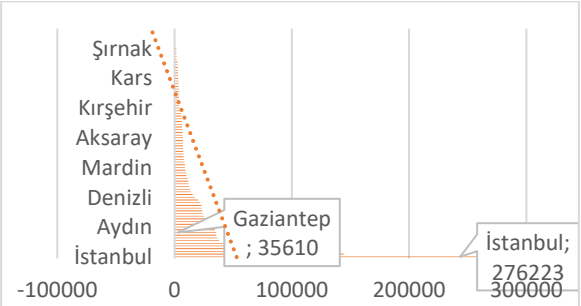

Figure 1. Housing sales in 2021 by provinces (TÜİK)

One of the biggest cities in Turkey, Gaziantep, has experienced significant housing inflation in recent years. Despite this, it has a high rank in terms of housing sales. Housing serves as a place to live as well as a significant investment that has an impact on life quality. The neighborhood of a house is very important to its stakeholders (investors, homeowners, developers, appraisers, and others) [2]. Therefore, it's critical for people to choose the neighborhood in which they can buy the house that has the features they want for their budget.

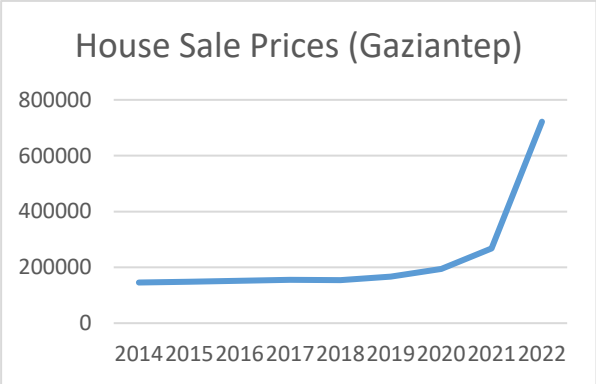

Figure 2. Change of house prices per square meter in Gaziantep by years

A rise in housing costs can be attributed to numerous things. The price of housing can alter due to a variety of factors, including a nation's economic structure, inflation, global events, politics, etc. These are the external variables. Other aspects of houses, such as the area, neighborhood, the size, and the number of rooms, can also have an impact on household pricing. The square meter of the house, the number of rooms, the district, real estate type, and the price are taken into account in the proposed study to forecast a house's neighborhood. Accurate predictions are required for the real estate and housing markets. We can notice a pattern that appears in the purchasing and selling of properties: for most people, owning a house is a lifelong ambition [3]. For various classification and regression problems, methods like the decision tree [4], extra trees [5], and random forest [6] are also frequently utilized. Machine learning algorithms were proposed by the authors as a reliable way to predict real estate property values. To increase prediction accuracy, they also examined feature importance and other data analysis techniques. In their research, they used linear regression, decision trees, XGBoost, extra trees, and random forests [7].

The objective of this study is to predict the neighborhood that a potential home buyer can afford to live in given her financial situation and certain property characteristics. Therefore, real estate data is taken from the website via web scraping. Machine learning methods like decision trees (DT), random forest (RF), and extra trees (ET) are used to forecast a house's neighborhood once the data has been collected.

The remainder of this paper is structured as follows: The literature review is presented in section 2, the research methodology is shown in section 3, and the findings and outcomes are presented in section 4. In section 5, the study is completed.

2. Literature Review

In order to assess real estate, there are some traditional approaches [2] in the literature such as cost approach [8] and income approach [9]. However, many machine learning-based methods have been utilized in recent years to forecast real estate values because they can identify functional relationships in past data.

The hedonic model, machine learning model, and geographically and temporally weighted regression (GTWR) were combined by Hu et al. [10]. They illustrated a novel method for segmenting the housing market. They then applied it to analyze the dynamics of the selling and renting submarkets in Shanghai, China, from 2018 to 2020. The method offered a useful and

[4] https://www.zingat.com/gaziantep-bolge-raporu

effective tool for segmenting the urban housing market.

Another study proposed by Xue and Yao [11] used a RF model to assess the significance and consequences of the physical environment, commute distance, housing costs, living area, and domestic socioeconomic and demographic factors on the housing relocation activity of members of the family with commuter employees or students. Their purpose was to offer baseline data for designers and real estate construction firms to use when planning and developing linked land, as well as to assist urban planners and directors in statistically examining the impact of causal factors on housing relocation activity.

Using data gathered from 24 randomly chosen Turkish cities, a Bayesian network driven by machine learning was created through a constrained structural learning method by Sevinç [12]. The information included various characteristics such as sales prices, details about each apartment's features, including its amount of bedrooms, the age of the building, whether a balcony is there, the net area, the heating system, mortgageability, the amount of bathrooms, seller type, and floor position.

Chou et al. [2] proposed a research that a brief overview of machine learning methods for forecasting home prices. The actual price registration system of Taiwan's Interior Ministry was used to gather information on housing transaction prices in Taipei City. Baseline and ensemble models were created using four popular artificial intelligence methods: Linear Regression, Artificial Neural Networks, Classification and Regression Trees, Support Vector Machines. Additionally, a hybrid model was created, and both baseline and ensemble methods were utilized to assess its prediction performance to that of the separate models.

Louati et al. [13] developed a collection of machine learning algorithms to perform analysis that could improve the accuracy of estimating land prices. The DT, RF, and linear regression are some of the machine learning methods used in their work. They gathered information from 5946 lands located in Riyadh, Saudi Arabia. Modern performance criteria, such as mean absolute error, median squared error, and mean squared error were utilized to assess the performance of the constructed models. According to the experiments, the RF-based model performs better than the other models.

Another research [14] that examined into the reasons Taiwanese home prices have risen steadily for the past ten years. Data relating to real estate gathered from publicly available websites was clustered employing a double-bottom map particle swarm optimization analytical technique. Population, rent and the money availability are three crucial elements that could impact real estate value trends, and they were highlighted in their assessment on the clustering results.

Instead of using different machine learning techniques to predict the transactions or list cost of real estate assets without distinguishing the building and land costs, Kim et al. [15] proposed a study that estimated cost of land utilizing a significant amount of land-utilization data information gathered from variety of building and land-related datasets. The RF and XGBoost algorithms were utilized to forecast 52,900 land costs in Seoul, South Korea, from January 2017 through December 2020. Additionally, the models underwent separate training for various land utilization and time periods. The overall findings showed that XGBoost produces a greater forecasting accuracy.

Some of the studies considered web scrapping technology in order to predict real estate [16]–[19]. Web scraping is a method for obtaining data from the Internet. The process of gathering data from websites and transferring them to a more convenient and more flexible form so that they can be examined and checked with ease is also referred to as web scraping, sometimes web crawling or data scraping, and sometimes data mining or text mining [16].

Web scraping and machine learning techniques are both utilized in the proposed study. While some studies focused solely on a few regions, others neglected crucial factors that affected real estate values, like the amount of rooms, neighborhood, and area. The studies' most typical drawback is that they estimate real estate prices in a limited region with few features. Even though the proposed study is focused on a particular area, it differs from prior studies in that it estimates neighborhood values rather than prices and deals with different real-estate features such as the square meter of real estate, the number of rooms, district, neighborhood, real estate type, and price. It is predicted where people can dwell based on a set of parameters and the amount of money they hold.

3. Material and Method

There are two parts to the planned study. Data is gathered in the first step using a web scraping technique. Utilizing machine learning techniques, the result is predicted in the second stage. The crucial steps in the use of web scraping and machine learning algorithms are presented in this section. First of all, web scraping is conducted to obtain data. The sequential procedures and methodologies for the web scrapping are shown in Figure 3.

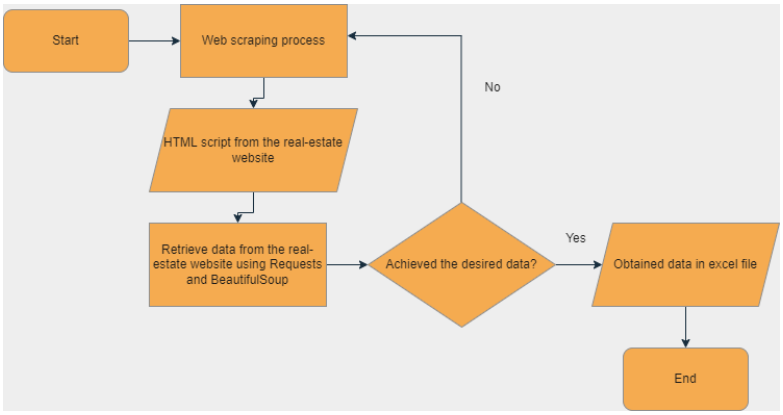

Figure 3. The structure of the proposed web scraping

Web scraping is a method of obtaining information from webpages using software. By utilizing low-level Hypertext Transfer Protocol (HTTP), this sort of software program simulates human browsing or investigation of the World Wide Web [20]. Python software is used for the web scraping. The steps of web scraping are given as follows:

- First of all, “EmlakBuluyoruz.Com” is chosen as real estate website[5].
- Search is conducted by considering Gaziantep city[6]. Houses for sale are considered in this study.
- In Python, Requests module is used to reach Hypertext Markup Language (HTML) contents of the website.
- The BeautifulSoup module is used to parse the contents of the HTML document and then discover elements using the "HTML class name" in the subsequent phase.
- After the mentioned processes, the data is obtained by utilizing the related class name.
- Finally, a excel file is created and data is saved for the next processes.

Through web scraping, a total of 548 pieces of data were collected on June 25, 2022. In Table 1, a sample of data is demonstrated. In the proposed study, to predict an house's neighborhood, the square meter of the house, the number of rooms, the district, real estate type, and the price are considered. Table 2 demonstrates the definition and data types of the variables. Scikit-learn supervised learning algorithms such as DT, RF, and ET are used to predict the neighborhood of a real estate.

[5] https://www.emlakbuluyoruz.com/

[6] https://www.emlakbuluyoruz.com/satilik/konut

Table 1. A sample data

The square meter, The number of rooms, District, Neighborhood, Real estate type, Price
170,3+1,Şehitkamil,Değirmiçem,Apartment,1275000
155,3+1,Şahinbey,Güneykent,Apartment,955000
195,3+1,Şehitkamil,İbrahimli,Apartment,2550000
250,4+1,Oğuzeli,Körkün,Cottage,3500000
120,2+1,Şahinbey,Kıbrıs,SingleHouse,400000
200,4+1,Şehitkamil,Onbeştemmuz,Apartment,2300000
175,3+1,Şehitkamil,İbrahimli,Apartment,1890000

In the proposed study, the target is to classify the target class, there are 75 classes available. neighborhood. Given that the neighborhood is

Table 2. Definition and data types of the variables

	Definition	Data Type
The square meter	Side x side = area of a square is the formula for calculating a square's surface area.	Integer
The number of rooms	Consists of rooms that may be occupied, such as bedrooms, kitchens, dining rooms, and living rooms.	String
District	A district is a particular kind of administrative division that the local government oversees in some countries.	String
Neighborhood	A neighborhood is a place where people reside and socialize.	String
Real estate type	Type of buildings	String
Price	Value of real estate	Integer

The collected data is processed via above mentioned machine learning techniques. These techniques are explained as follows:

3.1. Decision tree

DTs are one of the widely utilized techniques among classification algorithms. It is a non-parametric technique. DTs have nodes that form a rooted tree. That is, it is a directed tree without incoming edges at the root node. Each of the other nodes has just one incoming edge. Internal or test nodes are the name given to outgoing edges. Finally, the remaining nodes are known as leaves. [21]. Making a model that predicts the target value variable based on a variety of input factors is the aim of DT learning [22]. It is a graphical representation of decisions and their probable outcomes.

3.2 Random forest

One of the classifier algorithms used in it is composed of a number of DTs, each of which is created by putting an algorithm into practice. A majority vote is used to estimate the RF, rather than individual tree estimates [23]. Each tree in the ensemble is built individually utilizing a sample taken with substitution from the training set. As a result, RF is used to tackle mentioned problem due to its randomness [24].

3.3 Extra trees

ET, highly random trees based on the ensemble approach, lessen the weak generalization property and propensity for overfitting of conventional standalone DTs [25]. A meta-predictor that is fitted with several randomized DTs is utilized in this class. [26].

3.4. Evaluation

In the proposed study, the machine learning classification algorithms such as DT, RF, and ET are utilized. The Python programming language's Sklearn library is used. Classification accuracy (ACC) and F-measure are the metrics that are utilized for evaluation.

Evaluation metrics are computed by the given equations [27]:

ACC: Classifier accuracies are evaluated by using confusion matrices.

$$ACC = \frac{True\ Negatives + True\ Positives}{True\ Positives + False\ Positives + False\ Negatives + True\ Negatives} \quad (1)$$

F-measure: The other frequently used method for assessing the effectiveness of classification algorithms is the F-measure. It is a classification algorithm's harmonic mean of precision and recall. Improved predictive performance is indicated by higher F-measure values. The average F-measure for all one-versus-all classes is employed in this study as the macro-averaged F-measure.

$$\text{Macro} - \text{averaged F} - \text{measure} = \frac{1}{n}\sum_{i=1}^{n} \frac{2 * Precisioni * Recalli}{Preciisoni + Recalli} \quad (2)$$

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

4. Results and Discussions

The results of the algorithms are given in Table 3.

Table 3. Results of the applied machine learning algorithms

Algorithm	ACC	F-measure	Precision	Recall
DT	0.878	0.869	0.90	0.879
RF	0.848	0.845	0.879	0.848
ET	0.863	0.860	0.892	0.864

As can be shown from Table 3, all algorithms offer satisfactory outcomes of more than 84 percent when both ACC and F-measure are taken into account. However, DT provides better performance when compared with the other algorithms. ETs come after DT. The worst performance is provided by RF. The technologies of web scraping and machine learning are used to conduct this investigation. By considering DT algorithm, the neighborhood of a real estate can be predicted. Finally, the following are the implications of the proposed study:

- Web scraping technique is used to mine real estate data from a real estate website.
- Performances of various machine learning algorithms such as DT, RF, and ET are compared.
- While all algorithms perform well, DT produces the best results.
- An accurate prediction is given for determining a house's neighborhood.
- The proposed study can be beneficial for both people, investors, and government.

5. Conclusion

Recently, utilizing machine learning models with publicly available data has helped research on predicting real estate value. Unlike studies that make price estimations, the proposed study aims to predict the neighborhood of a house by using machine learning techniques such as DT, RF, and ET. We make an estimation of the Gaziantep real estate market. Based on a number of criteria and the amount of money they own, it is estimated where people can live. The data is retrieved from a real estate website via web scraping. The planned study includes an investigation into the city of Gaziantep. The real estate market is impacted by a number of variables, including a nation's economic structure, inflation, global events, politics, etc. These are the external factors. In addition, a house's size, number of rooms, neighborhood, and location can all determine how much it costs to live there as internal factors. In the proposed study, internal factors are considered to predict the neighborhood of a house. Based on the study, all algorithms provided good performance by estimating the neighborhood of a house.

However, DT provides the best results. These results can help decision-makers choose the neighborhood they want to live in or invest in. As we concentrated on Gaziantep for the regional study, data gathering was a constraint of our study. Conducting an analysis of Turkey is a potential direction for future research. In addition, data from other real estate websites can be gathered and analyzed for the future work.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Contributions of the Authors

The model and computational framework were developed by Ali Can ÜZÜMCÜ and Nazmiye ELİGÜZEL, who also carried out the experiments. The manuscript was written by Nazmiye ELİGÜZEL. Each author contributed ideas that helped refine the study, analysis, and article.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethic

References

[1] J. Ratcliffe, M. Stubbs;, and M. Keeping, *Urban Planning and Real Estate Development*. Routledge, 2021.

[2] J. S. Chou, D. B. Fleshman, and D. N. Truong, *Comparison of machine learning models to provide preliminary forecasts of real estate prices*, no. 0123456789. Springer Netherlands, 2022.

[3] A. S. Ravikumar, *Real Estate Price Prediction Using Machine Learning*. 2017.

[4] M. F. Lee, G. S. Chen, S. P. Lin, and W. J. Wang, "A Data Mining Study on House Price in Central Regions of Taiwan Using Education Categorical Data, Environmental Indicators, and House Features Data," *Sustain.*, vol. 14, no. 11, 2022, doi: 10.3390/su14116433.

[5] C. H. Cheng and M. C. Tsai, "An Intelligent Homogeneous Model Based on an Enhanced Weighted Kernel Self-Organizing Map for Forecasting House Prices," *Land*, vol. 11, no. 8, 2022, doi: 10.3390/land11081138.

[6] Y. Ahn and C. K. Uejio, "Modeling Air Conditioning Ownership and Availability," *SSRN Electron. J.*, vol. 46, no. October, p. 101322, 2022, doi: 10.2139/ssrn.4211073.

[7] M. Cekic, K. N. Korkmaz, H. Mukus, A. A. Hameed, A. Jamil, and F. Soleimani, "Artificial Intelligence Approach for Modeling House Price Prediction," pp. 1–5, 2022, doi: 10.1109/icmi55296.2022.9873784.

[8] D. Kulikauskas, "The user cost of housing in the Baltic states," *J. Eur. Real Estate Res.*, vol. 10, no. 1, pp. 17–34, 2017, doi: 10.1108/JERER-11-2015-0042.

[9] G. J. Rangel, J. W. J. Ng, T. T. Murugasu, and W. C. Poon, "Measuring Malaysian housing affordability: the lifetime income approach," *Int. J. Hous. Mark. Anal.*, vol. 12, no. 5, pp. 966–984, 2019, doi: 10.1108/IJHMA-02-2019-0023.

[10] L. Hu, S. He, and S. Su, "A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets," *Comput. Environ. Urban Syst.*, vol. 94, no. January, p. 101775, 2022, doi: 10.1016/j.compenvurbsys.2022.101775.

[11] F. Xue and E. Yao, "Adopting a random forest approach to model household residential relocation behavior," *Cities*, vol. 125, no. May 2021, p. 103625, 2022, doi: 10.1016/j.cities.2022.103625.

[12] V. Sevinç, "Determining the Flat Sales Prices by Flat Characteristics Using Bayesian Network Models," *Comput. Econ.*, vol. 59, no. 2, pp. 549–577, 2022, doi: 10.1007/s10614-021-10099-5.

[13] A. Louati, R. Lahyani, A. Aldaej, A. Aldumaykhi, and S. Otai, "Price forecasting for real estate using machine learning: A case study on Riyadh city," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 6, pp. 1–16, 2022, doi: 10.1002/cpe.6748.

[14] C. H. Yang, B. Lee, and Y. Da Lin, "Effect of Money Supply, Population, and Rent on Real Estate: A Clustering Analysis in Taiwan," *Mathematics*, vol. 10, no. 7, pp. 1–17, 2022, doi: 10.3390/math10071155.

[15] J. Kim, J. Won, H. Kim, and J. Heo, "Machine-learning-based prediction of land prices in Seoul, South Korea," *Sustain.*, vol. 13, no. 23, pp. 1–14, 2021, doi: 10.3390/su132313088.

[16] T. G. D. Souza, F. D. R. Fonseca, V. D. O. Fernandes, and J. C. Pedrassoli, "Exploratory spatial analysis of housing prices obtained from web scraping technique," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 43, no. B4-2021, pp. 135–140, 2021, doi: 10.5194/isprs-archives-XLIII-B4-2021-135-2021.

[17] H. Salem and M. Mazzara, "ML-based Telegram bot for real estate price prediction," *J. Phys. Conf. Ser.*, vol. 1694, no. 1, 2020, doi: 10.1088/1742-6596/1694/1/012010.

[18] A. Grybauskas, V. Pilinkienė, and A. Stundžienė, "Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00476-0.

[19] H. Ahmed, T. A. Jilani, W. Haider, S. N. Hasany, M. A. Abbasi, and A. Masroor, "Producing standard rules for smart real estate property buying decisions based on web scraping technology and machine learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 498–505, 2020, doi: 10.14569/ijacsa.2020.0110363.

[20] V. S. Katti and S. H. N, "Patents and Publications Web Scraping," *IJCSN Int. J. Comput. Sci. Netw.*, vol. 5, no. 2, pp. 2277–5420, 2016, [Online]. Available: www.IJCSN.org.

[21] L. Rokach and O. Maimon, "DECISION TREES," in *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*, no. January, 2005, pp. 165–192.

[22] M. A. Jun and J. C. P. Cheng, "Selection of target LEED credits based on project information and climatic factors using data mining techniques," *Adv. Eng. Informatics*, vol. 32, pp. 224–236, 2017, doi: 10.1016/j.aei.2017.03.004.

[23] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. New York: Cambridge University, 2013.

[24] G. Khanvilkar and D. Vora, "Product Recommendation using Sentiment Analysis of Reviews : A Random Forest Approach," *Int. J. Eng. Adv. Technol.*, no. January, 2019.

[25] S. Galelli and A. Castelletti, "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 7, pp. 2669–2684, 2013, doi: 10.5194/hess-17-2669-2013.

[26] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems : A comparison of support vector regression , random forest , extra trees and regression trees," *J. Clean. Prod.*, vol. 203, pp. 810–821, 2018, doi: 10.1016/j.jclepro.2018.08.207.

[27] A. Onan, "Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling," *Comput. Math. Methods Med.*, vol. 2018, 2018, doi: 10.1155/2018/2497471.