Development of a Python-Based Classification Web Interface for Independent Datasets

İpek Balikci Cicek, İlhami Sel, Fatma Hilal Yagin and Cemil Colak

Abstract—**Classification; biomedical, bioinformatics, medicine, engineering etc. It is a fundamental approach that is frequently used in many research areas, such as especially in the field of health; it has become common to classify diseases with machine learning methods using risk factors of these diseases and to determine the effect levels of these risk factors on the related disease. There are both commercial and free software tools that researchers can analyze their data with classification methods. The aim of this study is to develop a user-friendly web-based software for classification analysis. Python sklearn and Dash libraries were used during the development of the software. Among the classification algorithms in the developed software; Logistic regression, Decision trees, Support vector Machines, Random Forest, LightGBM, Gaussian Naive Bayes, AdaBoost and XGBoost methods are available. In order to show how the software works, a classification model was created with the Random forest algorithm using the cervical cancer data set. Different metric values were evaluated for the models. Obtained from a random forest classification model;accuracy, sensitivity, specificity, negative predictive value, matthews correlation coefficient, and F1 score values obtained from the model were 94.44%, 100%, 93.33%, 100%, 83.67%, and 94.44 respectively. It is thought that the classification software developed in this study will provide great convenience to clinicians and researchers in the field of medicine, in terms of applying predictive classification algorithms for the disease without any software knowledge.**

Index Terms— **Classification, machine learning, web based software.**

I. Introduction

THE NUMBER of data produced in parallel with the developing technology is increasing day by day. It is of great importance to store, manage and make useful the enormous amount of data produced. Therefore, it is of great importance to be able to use techniques that can process large amounts of data. Data mining is the process of discovering patterns and trends hidden in large data sets [1].

Data mining is a technique that attempts to identify previously unknown hidden relationships among data in databases [2]. Data mining is a multidisciplinary field that bridges many technical fields such as database technology, statistics, artificial intelligence, machine learning, pattern identification and data visualization [1].

Models used in data mining are examined under two main headings: predictive and descriptive. In predictive models, it is aimed to develop a model based on the data with known results and to estimate the result values for datasets with unknown results by using this established model. In descriptive models, patterns in existing data that can be used to guide decision making are defined [3].

It is possible to examine data mining models under three main headings, namely classification and regression, clustering and association rules, according to their functions. Classification and regression models are predictive models, clustering and association rules models are descriptive models [4]. It is of great importance that the patterns in the heap datasets are extracted by data mining/machine learning techniques and used as prediction and decision support components. In this context, one of the frequently applied data mining/machine learning topics is classification [5].

Classification is a very common process in scientific studies because of the benefits it provides in solving problems. Especially in the field of medicine, the classification of diseases and the development of treatment methods according to this classification are among the most prominent examples. In addition to medicine, the functionality of classification can be seen in other branches of science. Classification is an estimation process that assigns the observations that make up the data set to previously determined classes within the framework of certain rules [6]. Each observation in the data set has a feature and these features are divided into classes. Creates a model with the observations determined to belong to which class. The success of the model is measured with the observations that are not included in the training set. Quite different algorithms are used in classification methods. Different results have been obtained due to the fact that there are many algorithms. Each algorithm works with different parameters and has more than one version. The studied algorithms are for different purposes, the data source used is different, the algorithms support different data types, and the

İPEK BALIKCI CICEK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (ipek.balikci@inonu.edu.tr) https://orcid.org/0000-0002-3805-9214

İLHAMI SEL, Inonu University Computer Engineering Department, Engineering Faculty,44280 Malatya, Turkey, (ilhamisel23@gmail.com) https://orcid.org/0000-0003-0222-7017

FATMA HILAL YAĞIN, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (hilal.yagin@inonu.edu.tr) https://orcid.org/0000-0002-9848-7958

CEMIL COLAK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr) https://orcid.org/0000-0001-5406-098X

Manuscript received June 9, 2021; accepted Jan 26, 2022.
DOI: 10.17694/bajece.949935

preprocessing on the data depends on the practitioner. Knowing in which areas and in which types of variables these algorithms give more effective results increases the success of the methods. For this reason, it is important to apply classification methods by comparing them [7].

In this research, we aimed to develop a new user-friendly web-based software developed with the Python Dash library that will allow the comparison of Logistic regression, Decision trees, Support vector Machines, Random Forest, LightGBM, Gaussian Naive Bayes, AdaBoost and XGBoost data classification methods.

II. Material and methods

A. Dataset

The UCI data repository's open access dataset "Cervical Cancer Behavior Risk Data Set" was used to demonstrate how the software works. The dataset includes 72 cervical cancer samples with 18 predictive variables and one outcome variable. 50 (69.45%) of the samples tested negative for cervical cancer, while 22 (30.55%) tested positive for cervical cancer [8].

B. Methods

One of the most widely used data mining methods, which is used to classify large data sets and reveal important data classes, or to predict future data trends, is classification models [9].

Classification is used to reveal hidden patterns in databases. It is used to estimate the class of the data whose class has not been determined by using the existing classed data, or to determine whether the previously classified data is classified correctly and if there is a misclassification, it is used to assign the data to the correct group [4].

In this web-based software, there are algorithms such as Logistic regression, Decision trees, Support vector Machines, Random Forest, LightGBM, Gaussian Naive Bayes, AdaBoost and XGBoost, which are classification methods. The classification algorithms included in the software are described below. In addition, an application was made on the cervical cancer dataset in order to evaluate the outputs of the software.

B.1. Logistic Regression Analysis

The main purpose of logistic regression analysis is to model in order to define the relationship between the dependent variable and the independent variable without being subject to a certain precondition, when the dependent variable can be categorical and the independent variables can be both categorical and continuous. In other words, it is an analysis method that investigates the cause and effect relationship between the dependent variable and the independent variables. The relationship between the variables need not be linear. It can also be an exponential or binomial distribution relationship [10].

In the medical applications of logistic regression models, independent variables are risk variables or variables that determine whether a disease will occur or not. In short, logistic regression is a regression method that helps to assign and classify the expected value of the dependent variable according to the independent variables [11].

B.2. Decision Trees

Decision trees in data mining are the most widely used technique among classification models because they are cheap to set up, easy to interpret, easily integrated with database systems, and have good reliability. Decision tree, as the name suggests, is a predictive technique in a tree view. It is the most popular classification technique that can create easy-to-understand rules and integrates easily with information technology processes with its tree structure [12].

Decision trees create tree-based classification models. They classify records into groups or make an estimation of the target (dependent) variable value, which depends on the values of the independent variables [13].

To create a classification tree, there is a feature that best determines the examples in the learning set. With this feature, the so-called branch and leaves of the tree are separated and a new sample set is created. A new defining attribute is found from the instances on this parsed branch and new branches are created. If all instances in each sub dataset, that is, on the branch, belong to the same class, there are no other attributes to parse the instances, and there are no other instances with the value in the remaining attributes, the branching process ends. Otherwise, there is a respecifying attribute to parse the sub dataset [14].

B.3. Support Vector Machines

Support vector Machines is a machine learning model developed by Vapnik-Chervonenkis, used in clustering and regression problems, especially in classification [15].

The SVM method has been used frequently in recent years, especially in data mining, for classification problems in data sets where the patterns between variables are unknown. This method was originally thought of as a linear classifier for solving two-class problems, then generalized to the solution of nonlinearly separable or multi-class classification problems, and started to be widely used in solving these problems [16].

Support vector machine models is an algorithm that has become popular recently. The main purpose of the SVM model is to determine the hyperplane that will best separate the classes of the target variable from each other. In other words, it is to maximize the distance between support vectors belonging to different classes. Support vector machines use an iterative training algorithm used to minimize the error function to create an optimal hyperplane [15].

B.4. Random Forest

The random forest (RF) classifier is made up of a number of tree classifiers, each of which is constructed using a random vector sampled separately from the input vector, and each tree casts a unit vote for the most popular class in order to classify an input vector. A decision tree's design necessitated the selection of a feature selection measure as well as a pruning procedure. There are several techniques for choosing features for decision tree induction, and most of them assign a quality measure to the feature directly. The Information Gain Ratio criterion and the Gini Index are the most commonly utilized feature selection measures in decision tree induction. The Gini

Index is used by the random forest classifier as a feature selection measure, which measures the impurity of a feature in relation to the classes [17].

B.5. Gaussian Naive Bayes

Gaussian nave Bayes (GNB) classification is a supervised learning approach that employs Bayes' theorem as a framework for categorizing observations into one of a pre-defined set of classes based on predictor variables' information. GNB classifiers estimate the conditional probability that an observation belongs to a certain class based on the values of the predictor variables, assuming that the predictor variables are class-conditionally independent, and hence (naively) ignore predictor variable covariance. GNB classifiers beat other, more sophisticated classifiers in classification tasks, even when assumptions aren't met [18].

B.6. XGBoost

Using gradient-boosted decision trees, XGBoost was primarily developed for speed and performance. It represents a method for machine boosting, or applying boosting to machines, pioneered by Tianqi Chen and adopted by a large number of developers. It's a part of the Distributed Machine Learning Community's toolkit (DMLC). For tree boosting methods, XGBoost (eXtreme Gradient Boosting) aids in maximizing memory and hardware resources. It has the advantages of improving the algorithm and modifying the model, and it can also be used in computing settings. Gradient Boosting, Regularized Boosting, and Stochastic Boosting are the three major gradient boosting techniques that XGBoost can perform. It also distinguishes itself from other libraries by allowing the addition and adjustment of regularization parameters. The approach is very efficient in lowering computation time and making the best use of memory resources [19].

B.7. LightGBM

Microsoft's LightGBM is a free and open source Gradient boosting algorithm. The parallel voting decision tree approach, which uses the histogram-based method to speed up the training process, minimize memory consumption, and combine advanced network connectivity to maximize parallel learning, employs the histogram-based method. In each cycle, divide the training data into different machines and make a local voting choice to select the top-k attributes and a global voting choice to receive the top-2k attributes. To locate the leaf with the highest splitter gain, LightGBM employs a leaf-by-leaf approach [20].

B.8. AdaBoost

The AdaBoost algorithm creates strong classifiers from weak ones. The AdaBoost algorithm's weak classifiers are members of the ensemble classifier. By adaptively modifying the weights in each cycle, AdaBoost develops a committee of member weak classifiers. The weights of the training samples that a current weak classifier misclassified are increased, whereas the weights of the training samples that a current weak classifier successfully classified are dropped [21].

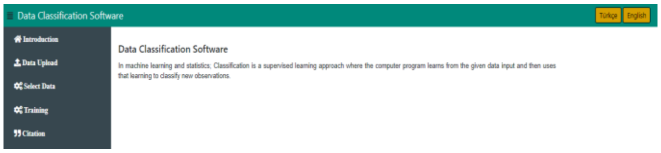
C. Model Validation and Performance Evaluation

For model validation, the data set was splitted into training (75%) and testing (25%) datasets. In the evaluation of classification results for all classification methods available in the software; Performance criteria of accuracy, sensitivity, specificity, negative predictive value, false positive rate, false negative rate, matthews correlation coefficient, positive likelihood ratio, negative likelihood ratio and F1 score are given. Obtained from a random forest classification model; accuracy is 94.44%, sensitivity 100%, specificity 93.33%, and F1 score value of 94.44%. The random forest model created according to the relevant performance criteria successfully classifies cervical cancer.

III. RESULTS

D. Data Classification Software

The data classification software user interface was created using Python Dash and Html codes. The data set file can be loaded from the data loading menu, and the predictive and predicted (class variable) variables can be selected in the data selection menu. The next menu is the training menu, which includes comprehensive classification algorithms. The main menu of the software is as in Figure I.

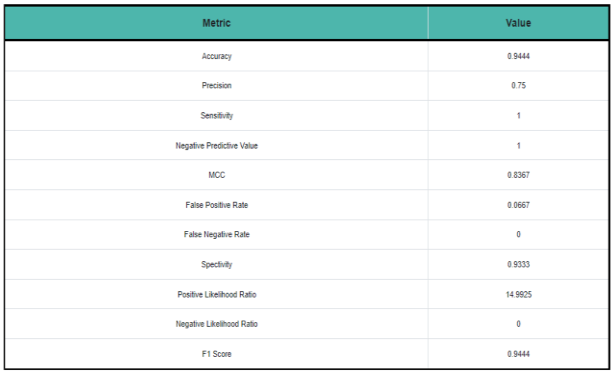


Data Classification Software

In machine learning and statistics; Classification is a supervised learning approach where the computer program learns from the given data input and then uses that learning to classify new observations.

Figure I. The main menu of the software

The values of the performance criteria for the generated Random forest model are shown in Figure II. Obtained from a random forest classification model; Accuracy is 94.44%, Sensitivity 100%, Specificity 93.33%, and F1 score value of 94.44%. The random forest model created according to the relevant performance criteria successfully classifies cervical cancer.



Metric	Value
Accuracy	0.9444
Precision	0.75
Sensitivity	1
Negative Predictive Value	1
MCC	0.8367
False Positive Rate	0.0667
False Negative Rate	0
Spectivity	0.9333
Positive Likelihood Ratio	14.9925
Negative Likelihood Ratio	0
F1 Score	0.9444

Figure II. Results for Performance Metrics for Random Forest Model Generated with Data Classification Software

IV. Conclusions

Classification analysis is one of the basic machine learning methods and is used by a large scientific community. There are many analysis tools used to guide researchers in this type of analysis. There are both commercial and free software tools that users can analyze their data with classification methods. In general, easy-to-use and well-designed interfaces are offered by commercial software packages [22].

Generally, well and comprehensively designed interfaces are offered by commercial software packages. One of these packages, Stata, is command-based and dependent on its commercial environment, the operating system. One of the free software packages that provides advanced possibilities for data classification analysis is the R environment. However, the R environment, like Stata, is OS dependent and command based. There are also free, open-source tools such as Weka that provide advanced analysis techniques. However, the Weka environment is an analytics tool that works as a desktop application. The fact that Weka is a desktop application can be difficult and time-consuming, especially for users (physicians, etc.) who do not have the Weka program installed on their computer. In addition, performing analysis with the Weka interface can be complex for most physicians [23-25].

Stata, R environment, Python, Rapidminer and WEKA programs can be both time consuming and difficult for researchers when data analysis needs to be evaluated quickly as they have to install these programs on their computers. Also, for most researchers, performing their analysis with such software can become more complex. On the other hand, the web-based software developed by this study is free, user-friendly and can perform data classification analysis from any device with internet access without writing any code, and provides comprehensive performance criteria and outputs of the classification results.

The software developed in this study, on the other hand, offers researchers a new user-friendly web-based software where they can easily perform data classification analysis and the analysis results can be easily understood. As a result, it is expected that this web-based software will allow them to compare comprehensive classification methods in disease prediction, particularly when compared to other analysis tools for physicians and healthcare professionals.

References

[1] S. Özekes, "Veri Madenciliği Modelleri ve Uygulama Alanları," 2003.

[2] S. Y. B. Dalı, "Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (Crm) Bir Uygulama."

[3] N. Zhong and L. Zhou, Methodologies For Knowledge Discovery And Data Mining: Third Pacific-Asia Conference, PAKDD'99, Beijing, China, April 26-28, 1999, Proceedings: Springer, 2003.

[4] H. Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi, İşletme Fakültesi Dergisi, C," ed: XXIX, 2000.

[5] G. AkgülL, A. A. Çelik, Z. E. Aydın, and Z. K. Öztürk, "Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı," Bilişim Teknolojileri Dergisi, vol. 13, pp. 255-268, 2020.

[6] B. Gülmez, "Yapay Sinir Ağlarının Yeni Metasezgisel Algoritmalar ile Eğitimi ve Veri Madenciliğinde Sınıflandırma Alanında Kullanımı," Erciyes Üniversitesi, Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı.

[7] Y. E. Kuyucu, "Lojistik Regresyon Analizi (LRA), Yapay Sinir Ağları (YSA) ve Sınıflandırma ve Regresyon Ağaçları (C&RT) Yöntemlerinin Karşılaştırılması ve Tıp Alanında Bir Uygulama," Gaziosmanpaşa Üniversitesi, Sağlık Bilimleri Enstitüsü, 2012.

[8] R. Machmud and A. Wijaya, "Behavior Determinant Based Cervical Cancer Early Detection With Machine Learning Algorithm," Advanced Science Letters, vol. 22, pp. 3120-3123, 2016.

[9] F. Köktürk, H. Ankaralı, and V. Sümbüloglu, "Veri Madenciliği Yöntemlerine Genel Bakis/Overview to Data Mining Methods," Türkiye Klinikleri Biyoistatistik, vol. 1, p. 20, 2009.

[10] N. Bayram, "Multinominal Lojistik Regresyon Analizinin İstihdamdaki Işgücüne Uygulanması," İstanbul Üniversitesi İktisat Fakültesi Mecmuası, vol. 54, p. 61, 2004.

[11] H. Bircan, "Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama," Kocaeli Üniversitesi Sosyal Bilimler Dergisi, pp. 185-208, 2004.

[12] G. Ulusoy, "Karar Ağacı Analizi ile AB Genişleme Kriterlerinin Değerlendirilmesi," 2013.

[13] G. Silahtaroğlu, "Veri Madenciliği," Papatya Yayınları, İstanbul, 2008.

[14] E. Akçetin And U. Çelik, "İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması," Internet Uygulamaları ve Yönetimi Dergisi, vol. 5, pp. 43-56, 2014.

[15] V. Vapnik, The Nature Of Statistical Learning Theory: Springer science & business media, 2013.

[16] Ö. Y. Akşehirli, H. Ankaralı, D. Aydın, and Ö. Saraçlı, "Tıbbi Tahminde Alternatif Bir Yaklaşım: Destek Vektör Makineleri," Türkiye Klinikleri Journal of Biostatistics, vol. 5, 2013.

[17] M. Pal, "Random Forest Classifier For Remote Sensing Classification," International journal of remote sensing, vol. 26, pp. 217-222, 2005.

[18] J. C. Griffis, J. B. Allendorfer, and J. P. Szaflarski, "Voxel-Based Gaussian Naïve Bayes Classification Of Ischemic Stroke Lesions In Individual T1-Weighted MRI Scans," Journal Of Neuroscience Methods, vol. 257, pp. 97-108, 2016.

[19] S. S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost," Information, vol. 9, p. 149, 2018.

[20] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An Effective Mirna Classification Method In Breast Cancer Patients," in Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 7-11.

[21] T.-K. An and M.-H. Kim, "A New Diverse Adaboost Classifier," in 2010 International conference on artificial intelligence and computational intelligence, 2010, pp. 359-363.

[22] İ. Perçin, F. H. Yağın, A. K. Arslan, and C. ÇOLAK, "An Interactive Web Tool for Classification Problems Based on Machine Learning Algorithms Using Java Programming Language: Data Classification Software," in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019, pp. 1-7.

[23] T. C. Sharma and M. Jain, "WEKA Approach For Comparative Study of Classification Algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, pp. 1925-1931, 2013.

[24] L. StataCorp, "Stata Data Analysis and Statistical Software," Special Edition Release, vol. 10, p. 733, 2007.

[25] R. RStudio Team, "RStudio: Integrated Development for R," RStudio, Inc., Boston, MA URL http://www. rstudio. com, vol. 42, p. 14, 2015.

Biographies

İPEK BALIKÇI ÇİÇEK obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

İLHAMİ SEL obtained his BSc. degree in Computer Education from Fırat University in 2007 and Computer Engineering from Inonu university in 2018. He received MSc. degree in Electronic and Computer Education from Fırat University in 2013. He currently continues his Ph.D. studies in Computer Engineering at the Inonu University. He has been working as a computer teacher in the Ministry of Education since 2007. His research interests are Natural Language Processing, data mining, machine learning, deep learning.

FATMA HİLAL YAĞIN obtained her BSc. degree in Statistics from Gazi University in 2017. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2020. She currently continues Ph.D. education in biostatistics and medical informatics from the Inonu University. In 2019, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, and image processing

CEMİL ÇOLAK obtained his BSc. degree in Statistics from Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in the general image processing, artificial intelligence, data mining, and analysis.