Journal of Basrah Researches (Sciences) **48**(2), 112 (2022)
DOI: https://doi.org/10.56714/bjrs.48.2.11

Big Data Framework Classification for Public E-Governance Using Machine Learning Techniques

Mohammed H. Altamimi*, Maalim A. Aljabery, Imad S. Alshawi

Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq.

ARTICLE INFO

Received 27 August 2022
Accepted 22 September 2022
Published 30 December 2022

Keywords:

Big data, Classification, Data mining E-Government, Machine learning, Prediction.

Citation: M.H. Altamimi, et al., J. Basrah Res. (Sci.) **48**(2), 112 (2022).
DOI:https://doi.org/10.56714/bjrs.48.2.11

ABSTRACT

Using Machine Learning (ML) in many fields has shown remarkable results, especially in government data analysis, classification, and prediction. This technology has been applied to the National ID data (Electronic Civil Registry) (ECR). It is used in analyzing this data and creating an e-government project to join the National ID with three government departments (Military, Social Welfare, and Statistics_ Planning). The proposed system works in two parts: Online and Offline at the same time; based on five (ML) algorithms: Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), and Naive Bayes (NB). The system offline part applies the stages of pre-processing and classification to the ECR and then predicts what government departments need in the online part. The system chooses the best classification algorithm, which shows perfect results for each government department when online communication is made between the department and the national ID. According to the simulation results of the proposed system, the accuracy of the classifications is around 100%, 99%, and 100% for the military department by the SVM classifier, the social welfare department by the RF classifier, and the statistics-planning department by the SVM classifier, respectively.

1. Introduction

Data Mining (DM) refers to extracting or mining knowledge from large amounts of data. It is also known as Knowledge-Discovery in Databases (KDD) or Knowledge Discovery and Data Mining. It is the process of automatically searching large volumes of data for patterns like association rules. DM applies many older computational techniques from statistics, information retrieval, ML, and pattern recognition [1]. ML is a technology that can handle Big Data classification for statistical or even more complex purposes such as decision making. The use of advanced technologies of ML fits perfectly with the scope of the new generation of government e-government [2,3]. The government is working to improve its performance through the use of modern technology resources such as the mobile, internet, etc., It is known as e-government. The government strives to enhance the political and social climate, as well as to effect fundamental change in how functions are carried out. These e-services provide better delivery of government services to citizens.

*Corresponding author email : itpg.mohammed.haron@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://bjrs.uobasrah.edu.iq

In addition, it improves interactions with industry and business, enabling citizens' access to information and more efficient government management. The resulting benefits can be mentioned as increased transparency, less corruption, decreased time and effort, revenue growth, and/or cost reductions, and greater convenience for citizens [4,5]. A classification is a form of data analysis that extracts models describing important data classes. Such models are called classifiers which predict categorical (discrete, unordered) class labels [6,7]. Many classification methods have been proposed by researchers in ML, like the Decision Tree (DT) classifier [8,9] and Neural Network (NN) classifier [10,11]. In this paper, the researchers propose to create an e-government project to join the National ID with three government departments (Military, Social Welfare, and Statistics_ Planning). by applying ML to five classifiers (SVM, DT, KNN, RF, and NB), using the ECR data, and choosing the algorithm with the highest accuracy for all government departments. the proposed system works in two parts: Online and Offline at the same time; test classifiers on the offline side and predict with the high-accuracy classifier on the online side. ECR data were previously paper records manually 100%. Paper records hinder the process of benefiting from their data, limited access, lack of clarity, inability to access remote files, and the cost of storing. To increase the need for the use of this data by government departments, it has been converted into electronic data by the National ID Law No. 3 of 2016. The comprehensive data was generated which contains a lot of information government departments need in their work. Where it becomes necessary to take advantage of this data

The rest of this study is structured in four sections. Within section two, the related works information is provided since it is necessary for having a look at similar works in the similar fields of the present study, while section three presents the researchers' proposed method which underlies the present research. Besides, section four includes the result and discussion, and finally, section five displays the conclusion and future work.

2. Related Work

Several researchers analyzed and classified government data using different DM techniques and ML algorithms. Charalampos Alexopoulos et al. [2] declare that their study contributes to this research topic by offering a thorough examination of government usage of ML. Rajagopalan M.R al. [4] presented a model that illustrates how big data can result in the transformation of the government by increased efficiency and effectiveness in the e-governance service with citizen engagement in decision-making. Ebenezer Agbozoet al. [12] provides an overview of the emergence of big data analytics of public sector e-services in governance. Kwok Tai Chui al. [13] provided a review of the use of machine learning algorithms, optimization algorithms, and applications in smart healthcare. Ayman Mir et al. [14] explains that built a classification model, to classify health data and the prediction of diabetes disease by a set of algorithms such as Naive Bayes, Random Forest, Support Vector Machine, and Simple CART algorithm. Maalim A. Aljabery et al. [15] analyzed the data of the National Health System using data mining and ML techniques and discovered what kind of hearing aids the patient needed. Mucahid Mustafa et al. [16] confirm that their study is utilized to evaluate the likelihood of getting breast cancer by using health data. World Bank [17] make recommendations for moving toward a more inclusive, trusted, and service delivery-oriented NID system. Pooja Thakar et al. [18] introduce a comprehensive survey, Journey (2002-2014) towards the exploration of educational data and its future scope. Mohammad Sultan et al. [19] present a comprehensive survey of the methods and techniques of data partitioning and sampling concerning big data processing and analysis. Fadi Salo et al. [20] apply a criterion-based approach to select 95 relevant articles from 2007 to 2017. The researchers identify 19 separate DM techniques used for intrusion detection, and the analysis includes rich information for future research based on the strengths and weaknesses of these techniques. Nawaf Alsrehin et al. [21] focus on traffic management approaches based on DM and ML techniques to detect and predict traffic. Mr. Sudhir et al. [22] presented a study of various DM classification techniques like Decision Tree, KNearest Neighbor, Support Vector Machines, Naive Bayesian Classifiers, and Neural Networks. Abdullah H et al. [23] show that their

research is a comparative evaluation of a variety of free DM and Knowledge Discovery tools and software packages The results reveal that the type of dataset utilized and how the classification algorithms are implemented inside the toolkits impact the performance of the tools for the classification job. Ivan Garcia et al. [24] proposed using big data analytics techniques, such as Decision Trees for detecting nodes that are likely to fail, and so avoid them when routing traffic. This can improve the survivability and performance of networks. Muhammet Sinan et al. [25] mention that the Bank marketing data set in UCI Machine Learning Data Set was used by creating models with the same classification algorithms in different DM programs. Saba Abdul W. Saddam et al. [26] propose a secure framework for mining cloud data in a privacy-preserving manner. A secure KNN classifier is used. Mohammed Z. Al-Faiz et al. [27] work to achieve different motions of the prosthetic arm by better classification with multiple factors using K nearest neighbor.

3. Proposed Method

Due to the presence of comprehensive data that contains a lot of information in ECR, the departments need to facilitate their work and then serve the citizens. The researchers propose to create e-government from three government departments (Military, Social Welfare, and Statistics-Planning) by applying five ML algorithms (SVM, DT, KNN, RF, and NB). The proposal consists of the following phases.

3.1. Data Collection Phase

The database was created with a structure similar the real data of the national identity data ECR in a large percentage with some simple changes to certain fields that do not have a significant impact on the nature of the data in order to get out of its secrecy and at the same time to be in conformity with what is required in our research. due to the fact that the real data was not obtained for its confidentiality and the privacy of citizens' information. The collected data consists of 10,000 records, as shown in Fig 1, each record with content for (46) attributes. It contains important citizens' data such as ID number, gender, name, family, mother's name, date of birth, health status, number of children, country of residence...etc. They were entered manually by Microsoft Excel with a missing percentage of 20.6%.

	ID number	gender	first name	second name	third name	fourth name	family	Mother's name	Mother's father	mother's grandfather	...	nationalism	Birth certificate No	Card Nationality No	Card Living No
0	201696670961	Male	Rashid	Faith	Muhammad	Zahir	Al-Hatti	retag	ali	mortada	...	Kurdish	124921	796486	26414
1	198158372498	Male	Abbas	Naji	Adam	Ghafel	Al-Lami	rwan	tahsin	abd	...	Yezidis	435924	800971	93970
2	201719788953	Male	Jassim	Tawfiq	Hassan	Suhim	Al-Ali	krama	ali	alkarim	...	?	181693	314360	49633
3	198044925325	Male	Jamil	Hammoud	Attia	Khwain	Al-Maliki	ahlam	hbib	kamil	...	Kurdish	352265	576842	44183
4	195063896422	Male	Star	Jabbar	Sugar	Obaid	Al-Maliki	isra	rahim	tahsin	...	Kurdish	297168	236707	83409
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9994	196542764483	female	hyam	abdalabas	kamil	rahim	Al-Bahadli	aya	kmal	gfar	...	?	868274	853401	23440
9995	199173020654	female	hind	abd	abass	alawi	kmal	sarmed	hadi	allatif	...	Yezidis	905420	198127	21340
9996	200810754260	female	duha	allatif	eihab	tahsin	?	ruqi	abdallatif	nsaif	...	Arabic	430723	347434	73620
9997	193028959752	female	iyman	eihab	nabil	riyad	?	fuda	hadi	raoof	...	?	319408	932646	23653
9998	200557048552	female	nanjs	abdalhusain	mortada	faris	?	zanib	abdal	nabil	...	?	818078	856774	10297

9999 rows × 46 columns

Fig. 1. Database of ECR.

3.2. Data Pre-Processing Phase

In this phase, the inconsistent, incomplete, and missing data are processed such as deleting the duplicate records that contain missing data by DM tools to get more accurate data. This accurate data helps in creating an accurate database and gives good results when training ML algorithms on it. After this step, data containing 500 records and 46 attributes were obtained with a missing rate of 1.4 percent, as shown in Fig.2.

	ID_number	gender	first_name	second_name	third_name	fourth_name	family	Mother's_name	Mother's_father	mother's_grandfather	...
0	201679624052	male	Haider	Rashid	Badan	Hassan	Al-Fakhri	om albnin	alkarim	eihab	...
1	198160894108	?	Amer	Najm	Abdullah	Mohammed	Al	hind	hadi	riyad	...
2	201789713024	male	Ali	Nasser	Musa	Muhammad	Al-Bouhiyah	hnadi	abdalabas	gfar	...
3	198069986689	?	Najat	Nima	Saleh	Aboud	Al	thani	zbali	hadi	...
4	201775048808	male	Nasser	Gary	Obaid	as	Eid	rfil	raid	abdalabas	...
...	...	...	...	...	...	...	...	...	...	...	...
494	200247470037	male	Jassim	Hassan	Muhammad	Ghaleb	Al-Maliki	ahlam	abdal	ahmaid	...
495	200172148932	male	Jassim	Hassoun	Latif	Saadoun	Al-Mayi	retag	husiain	raoof	...
496	195696348196	male	Jassim	Salem	Sender	of	Shakban	narjis	alkarim	abdalabas	...
497	200791571592	male	Jassim	Sherida	Jassim	Mohammed	Al-Mohammed	fuda	abdalabas	gmal	...
498	197458335261	?	Jassim	Sharif	Abdul	Hassan	Al-Husainat	rwan	hbib	talib	...

499 rows × 48 columns

Fig. 2. Data Warehouse of ECR.

Perform normalizing, which boosts the time processing of ML algorithms. The researchers achieve the targets of this process by increasing the processing time of the proposed model. Therefore, this part enhances the data set, which becomes more ready to use in the later parts [28], as shown in Fig.3.

_name	Mother's_father	mother's_grandfather	...	Living_governorate	Living_country	Passport_No	Telephone_No	male	healthily	death	free_work	Living_iraq
albnin	alkarim	eihab	...	Kuwait	Kuwait	A50062035	7805617216	1	1	1	0	0
hind	hadi	riyad	...	Maysan	iraq	A71495003	7806671726	0	1	1	0	1
hnadi	abdalabas	gfar	...	Basrah	iraq	A30107795	7803677076	1	1	1	1	1
thani	zbali	hadi	...	Stockholm	Sweden	A62779580	7801093325	0	0	0	1	0
rfil	raid	abdalabas	...	Babylon	iraq	A76733252	7809946500	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
ahlam	abdal	ahmaid	...	Baghdad	iraq	A34188742	7805506023	1	0	1	1	1
retag	husiain	raoof	...	Karbala	iraq	A88728022	7805182510	1	0	1	1	1
narjis	alkarim	abdalabas	...	Dhi Qar	iraq	A87955318	7802912666	1	1	1	1	1
fuda	abdalabas	gmal	...	Dhi Qar	iraq	A45831606	7806996245	1	1	1	1	1
rwan	hbib	talib	...	Basrah	iraq	A51431646	7804625292	0	1	0	0	1

Fig. 3. Normalizing.

In addition, it is important to extract significant features for each government department. In this step, it is an essential part to build a rating model. These features are extracted based on the needs of each government department. Moreover, these features differ from each other, for example, health status is one of the features used by the military to classify the soldier, whether he is armed or unarmed, while the state of health is not a feature of social welfare. The current phase works to correct and reduce the size and dimensions of the data so that increasing the performance of the classifier[6].

3.3. Train-Test Splitting and Classification Phase

This is the basic phase of the proposed system where data is divided and a classification model is built. The data is divided into two groups for training and testing, with 70% training data and 30% test data. It is a significant step that plays an influential role in preparing the data for classification. This division is so important in training ML algorithms to reduce errors and increase accuracy, as shown in Fig. 4.

Fig. 4. Train-Test Splitting.

The goal of the present paper is to build a model to classify and predict the needs of three government departments (Military, Social Welfare, and Statistics-Planning) The proposed system works on two sides (online and offline) and is based on five ML algorithms: SVM, DT, KNN, RF, and NB. After the train-test splitting stage, the data set is ready for the later stage. The department's feature set is entered into ML algorithms in an offline way, then the algorithm with the highest accuracy is selected and determined to work online in the future for this government department, as shown in Fig. 5.

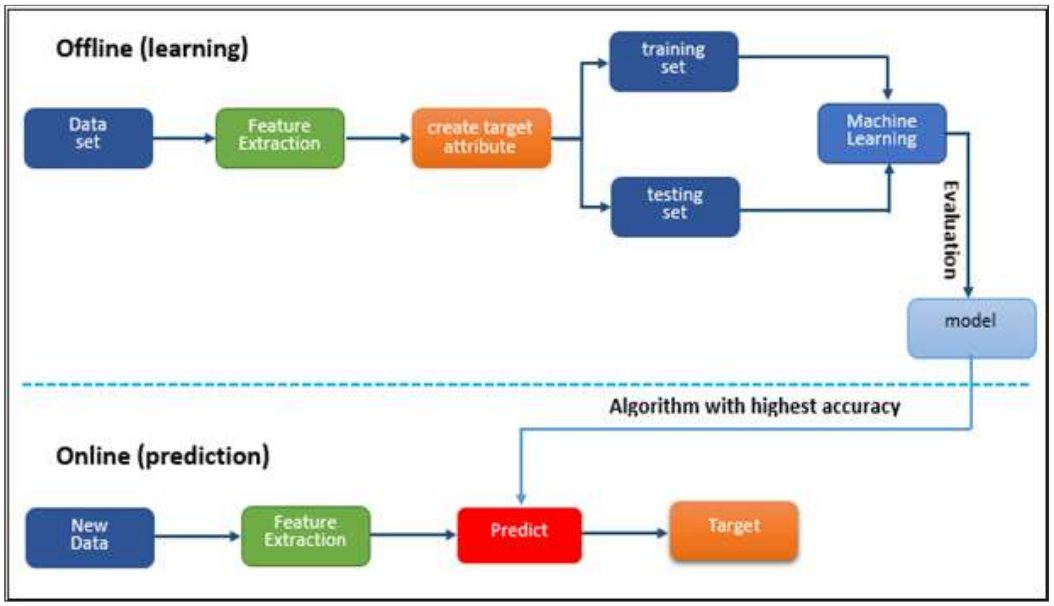

Fig. 5. Classification Model.

4. Results and Discussions

Experimental results were determined to explain the benefit and performance of the work. According to the fast-paced, and evolving lifestyle in recent years, the need to use e-government is increasing. Our work focuses on two parts: the first explains the role and usefulness of machine learning algorithms in the proposed system and reflects the e-government system in the second part. The proposed e-government system is characterized by data processing on the offline side, to maintain the confidentiality and privacy of data. And share the information obtained with the three government departments (military, social welfare, and statistical planning) on the online side. The proposed system was built using the programming language Python and using the Orange Data Mining program to compare the results between the five algorithms SVM, DT, KNN, RF, and NB. The proposed system runs on a hp computer with an Intel(R) Core (TM) i5-2410M CPU @ 2.30GHz (4 CPUs), ~2.3GHz, 8 GB RAM, and a Windows 64-bit operating system. the proposed system shows the following results.

4.1. **Military**

The SVM algorithm shows higher accuracy (100%) compared to other algorithms in the government Military department, as shown in the figures. Where Fig. 6 shows a comparison between the algorithms and Fig. 7 shows the targets confusion matrix for the military and Fig. 8 shows the distribution of the targets for the military.

Model	AUC	CA	F1	Precision	Recall
kNN	0.999	0.995	0.995	0.995	0.995
Tree	0.976	0.957	0.960	0.967	0.957
SVM	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	0.997	0.997	0.997	0.997
Naive Bayes	1.000	0.992	0.993	0.994	0.992

Fig. 6. Evaluation Results.

		Predicted			
		armed soldier	not a soldier	unarmed soldier	∑
Actual	armed soldier	100.0 %	0.0 %	0.0 %	105
	not a soldier	0.0 %	100.0 %	0.0 %	630
	unarmed soldier	0.0 %	0.0 %	100.0 %	15
	∑	105	630	15	750

Fig. 7. Confusion Matrix of SVM.

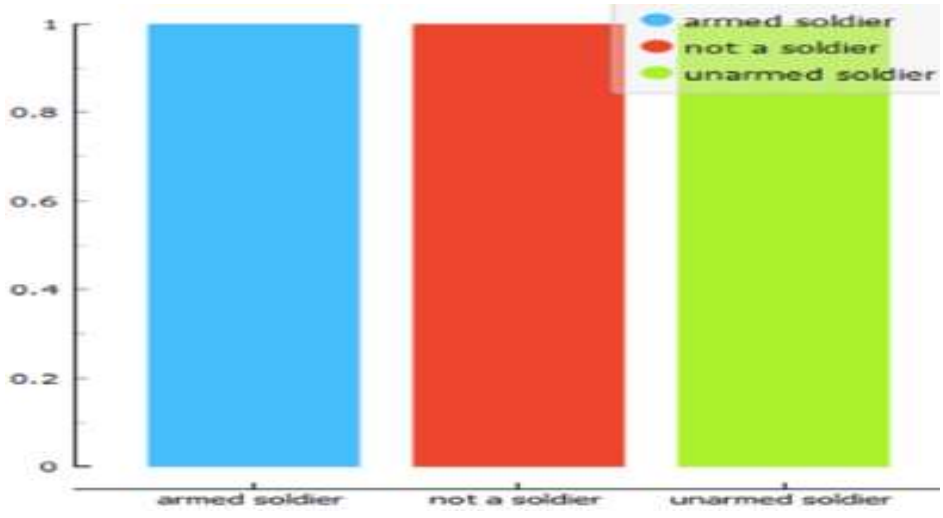

Fig. 8. Distribution of SVM.

4.2. Social Welfare

The random forest algorithm shows higher accuracy (99%) compared to other algorithms in the government Social welfare department, as shown in the figures. Where Fig. 9 shows a comparison between the algorithms and Fig. 10 shows the targets confusion matrix for social welfare and Fig. 11 shows the distribution of the targets for social welfare.

.

Model	AUC	CA	F1	Precision	Recall
kNN	0.999	0.975	0.976	0.978	0.975
Tree	0.951	0.887	0.833	0.786	0.887
SVM	1.000	0.967	0.959	0.954	0.967
Random Forest	1.000	0.993	0.993	0.994	0.993
Naive Bayes	0.997	0.933	0.944	0.964	0.933

Fig. 9. Evaluation Results.

		Predicted						
		Social insurance_1	Social insurance_2	Social insurance_3	Social insurance_4	Social insurance_5	not Social insurance	∑
Actual	Social insurance_1	80.0 %	20.0 %	0.0 %	0.0 %	0.0 %	0.0 %	5
	Social insurance_2	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	10
	Social insurance_3	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	20
	Social insurance_4	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	20
	Social insurance_5	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	30
	not Social insurance	0.0 %	0.0 %	0.0 %	0.0 %	0.6 %	99.4 %	665
	∑	4	11	20	20	34	661	750

Fig. 10. Confusion Matrix of Random Forest.

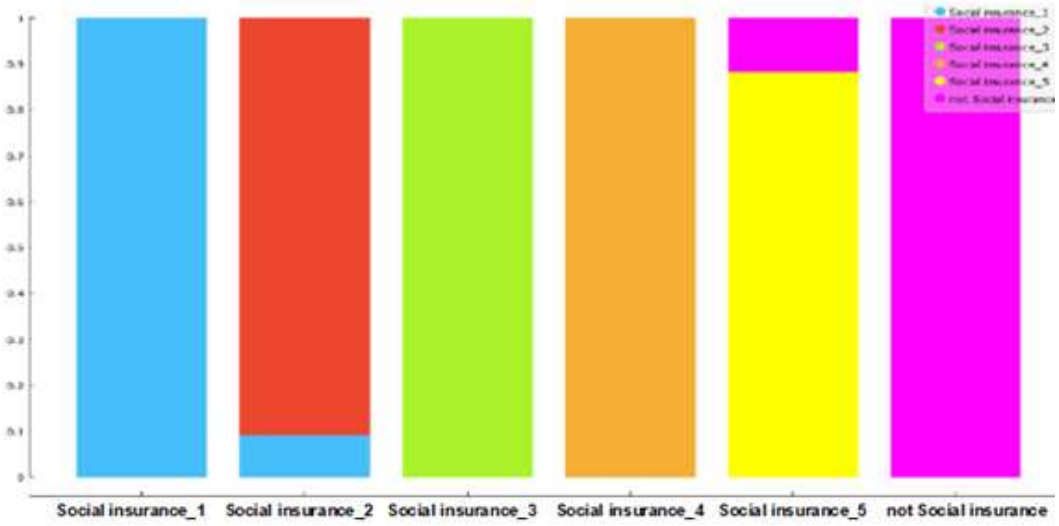

Fig. 11. Distribution of Random Forest.

4.3. Statistics_ planning

The SVM algorithm shows higher accuracy (100%) compared to other algorithms in the government Statistics_ planning department, as shown in the figures. Where Fig. 12 shows a comparison between the algorithms and Fig. 13 shows the targets confusion matrix for Statistics_ planning and Fig. 14 shows the distribution of the target for statistics_ planning.

Model	AUC	CA	F1	Precision	Recall
kNN	0.994	0.975	0.974	0.977	0.975
Tree	0.958	0.735	0.673	0.692	0.735
SVM	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	0.996	0.996	0.996	0.996
Naive Bayes	1.000	0.977	0.977	0.980	0.977

Fig. 12. Evaluation Results.

		Predicted											
		baby boy	baby girl	man	no	old man	old woman	teenage boy	teenage girl	woman	young man	young woman	∑
Actual	baby boy	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	50
	baby girl	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	30
	man	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	55
	no	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	90
	old man	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	60
	old woman	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	60
	teenage boy	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	40
	teenage girl	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	20
	woman	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	40
	young man	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	185
	young woman	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	120
	∑	50	30	55	90	60	60	40	20	40	185	120	750

Fig. 13. Confusion Matrix of SVM.

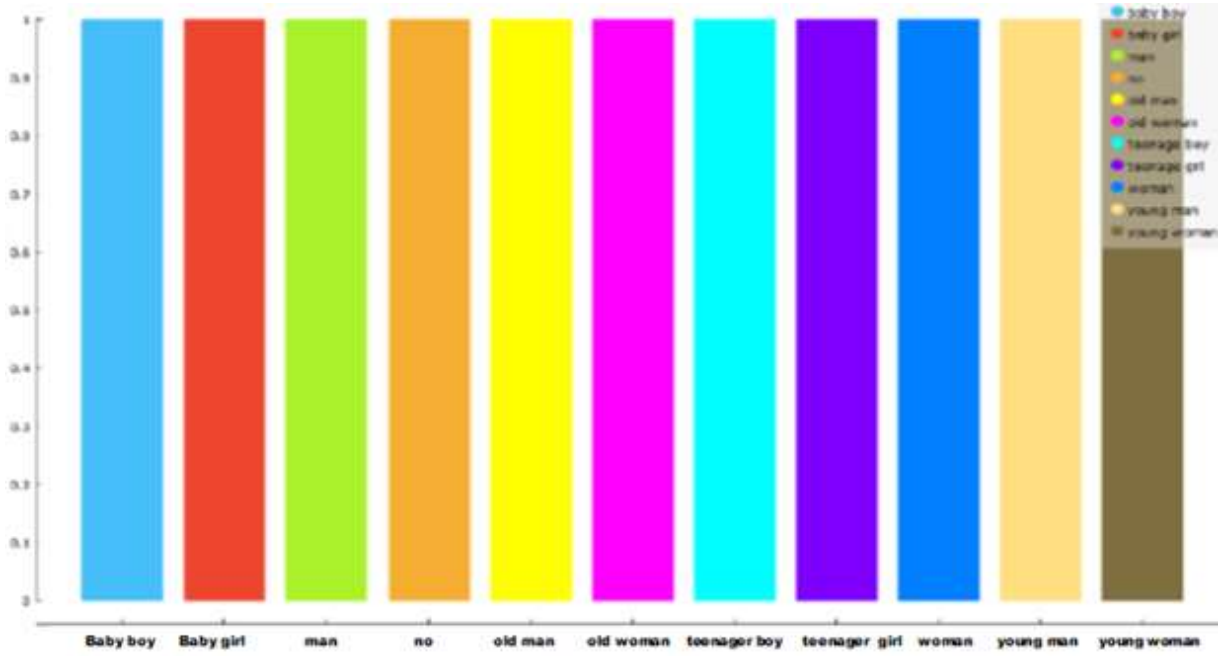

Fig. 14. Distribution of SVM.

5. Conclusion and Future Works

Data mining is the process of discovering hidden knowledge in the data. [29] Including ECR data. Classification is the DM technique that allocates a category label to a set of unclassified data. The main objective of this paper is to create an e-government project by sharing ECR data with a selected group of state departments (online) way after predicting the categories and needs of these state departments in the (offline) way using a set of Machine learning algorithms and choosing the most accurate algorithm for later online use. We used multiple algorithms to expand the scope of comparison in order to choose the best among them and also for the diversity of results due to the diversity of data that we adopted in our research. In the other hand, several algorithms examined to determine the algorithm which give the more accurately prediction, especially that the effort in the diversity of algorithms was for choosing the best among them to depends in the future work. Furthermore, it is important to mention that the work should be performed online, and offline at the same time to maintain the confidentiality and privacy of such data and create an integrated e-government project that includes all the joints and ministries of the Iraqi state.

References

[1] A.M. Hirudkar, M.S.S. Sherekar, International Journal of Computer Science and Applications **6**(2), 232 (2013).
[2] C. Alexopoulos, V. Diamantopoulou, Z. Lachana, Y. Charalabidis, A. Androutsopoulou, M.A. Loutsaris, ACM Journal, ICEGOV '19: Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance **1481**, 354 (2019).
[3] D. K. Altmemi, I. S. Alshawi, Journal of Positive School Psychology **6**(5), 1898 (2022).
[4] M.R. Rajagopalan, S. Vellaipandiyan, International Conference on ICT and Knowledge Engineering, (2013).
[5] M.D. Aljubaily, I.S. Alshawi International Journal of Electrical and Computer Engineering, **12**(2), 1776 (2022).
[6] J. Han, M. Kamber, J. Pei, Elsevier Science, (2011).
[7] I.S. Alshawi, Z.A. Abbood, A.A. Alhijaj, Telkomnika (Telecommunication Computer Electron. Control) 20(1), 212 (2022).
[8] N. Indumathi, R. Ramalakshmi, V. Ajith, International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 811 (2021).
[9] H.H. Al-badrei, I.S. Alshawi, Advances in Mechanics **9**(3), 1467 (2021).
[10] I. S. Alshawi, M. H. K. Jabbar, R. Z. Khan, International Journal of Management & Information Technology 6, 794 (2013).
[11] Z.A. Abbood, I.S. Alshawi, A.A. Alhijaj, F.P. Vidal, Telkomnika (Telecommunication Computer Electron. Control) **18**(5), 2439 (2020).
[12] E. Agbozo, K. Spassov, Proceedings of the 11th international conference on theory and practice of electronic governance 662 (2018).
[13] K.T. Chui, W. Alhalabi, S.S.H. Pang, P.O. de Pablos, R.W. Liu, M. Zhao, Sustainability **9**(12), 2309 (2017).
[14] A. Mir, S.N. Dhage, 2018 fourth international conference on computing communication control and automation (ICCUBEA), 1 (2018).
[15] M.A. Aljabery, S. Kurnaz, Journal of Information Science and Engineering **36**(2), 205 (2020).
[16] M.M. Saritas, A. Yasar, International Journal of Intelligent Systems and Applications in Engineering **7**(2), 88 (2019).
[17] World Bank, "Toward More People-Centered Service Delivery: Opportunities for the National ID System in Lesotho. Washington", World Bank Group, (2022).
[18] P. Thakar, A. Mehta, Manisha, International Journal of Computer Applications **110**(15), 60 (2015).

[19] M.S. Mahmud, J. Z. Huang, S. Salloum, T.Z. Emara, K. Sadatdiynov, Big Data Mining and Analytics **3**(2), 85 (2020).
[20] F. Salo, M. Injadat, A.B. Nassif, A. Shami, A. Essex, IEEE Access **6**, 56046 (2018).
[21] N.O. Alsrehin, A.F. Klaib, A. Magableh, IEEE Access **7**, 49830 (2019).
[22] S.M. Gorade, A. Deo , P. Purohit, International Research Journal of Engineering and Technology (IRJET) **4**(1), 3112(2017).
[23] A.H. Wahbeh, Q.A. Al-Radaideh, M.N. Al-Kabi, E.M. Al-Shawakfa, International Journal of Advanced Computer Science and Applications **8**(2), 18 (2011).
[24] I. Garcia-Magarino, G. Gray, R. Lacuesta, J. Lloret, IEEE Access **6**, 27958 (2018).
[25] M. S. Basarslan and I. D. Argun, Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT) 1 (2018).
[26] S.A.W. Saddam, J. Basrah Res. (Sci.) **43**(2), 44 (2017).
[27] M.Z. Al-Faiz, A.A. Ali, A.H. Miry, International Conference on Energy, Power and Control (EPC-IQ), **159** (2010).
[28] N. A. Noori, A.A. Yassin, Iraqi Journal for Electrical & Electronic Engineering **17**(2), 120 (2021).
[29] S. Kurnaz, M.A.H. Aljabery, ICEMIS '18: Proceedings of the Fourth International Conference on Engineering and MIS, **57**, 1 (2018).

Journal of Basrah Researches (Sciences) **48**(2), 112 (2022)
DOI: https://doi.org/10.56714/bjrs.48.2.11

تصنيف البيانات الكبيرة للحكومة الالكترونية باستخدام تقنيات التعلم الالي

محمد هارون التميمي*، معالم عبد علي الجابري، عماد شعلان الشاوي

قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة البصرة، البصرة، العراق.

معلومات البحث

الاستلام 27 آب 2022
القبول 22 ايلول 2022
النشر 30 كانون الاول 2022

الكلمات المفتاحية

البيانات الضخمة، التصنيف، التنقيب عن البيانات ، الحكومة الإلكترونية، التعلم الآلي، التنبؤ،|.

Citation: M.H. Altamimi, et al., J. Basrah Res. (Sci) **48**(2), 112 (2022).
DOI:https://doi.org/10.56714/bjrs.48.2.11

الملخص

أظهر استخدام التعلم الآلي (ML) في العديد من المجالات نتائج ملحوظة ، لا سيما في تحليل البيانات الحكومية وتصنيفها والتنبؤ بها. تم تطبيق هذه التقنية على بيانات البطاقة الوطنية (السجل المدني الإلكتروني) (ECR). يتم استخدامه في تحليل هذه البيانات وإنشاء مشروع حكومي إلكتروني لربط دائرة البطاقة الوطنية بثلاث دوائر حكومية وهي (التجنيد، والرعاية الاجتماعية ، والاحصاء والتخطيط). يعمل النظام المقترح في جزأين: Online و Offline في نفس الوقت ؛ استنادًا إلى خمس خوارزميات (ML): Support Vector Machine (SVM) و Decision Tree (DT) و K-Nearest Neighbour (KNN) و Random Forest (RF) و Naive Bayes (NB). يطبق الجزء Offline في النظام مراحل المعالجة المسبقة والتصنيف على ECR ثم يتنبأ بما تحتاجه الدوائر الحكومية من بيانات البطاقة الوطنية. يختار النظام أفضل خوارزمية تصنيف ، والتي تظهر نتائج مثالية لكل دائرة حكومية عند إجراء الاتصال Online بين الدوائر الثلاثة والبطاقة الوطنية. وفقًا لنتائج محاكاة النظام المقترح ، تبلغ دقة التصنيفات حوالي 100٪ و 99٪ و 100٪ لدائرة التجنيد بواسطة مصنف SVM ، ودائرة الرعاية الاجتماعية بواسطة مصنف RF ودائرة الاحصاء والتخطيط . بواسطة مصنف SVM ، على التوالي . يساعد النظام المقترح الدوائر الحكومية للاستفادة من اكبر قاعدة بيانات في الدولة العراقية وهي بيانات البطاقة الوطنية وذلك عبر مشاركة هذه البيانات مع تلك الدوائر كل حسب حاجته مع المحافظة على سرية وخصوصية هذه البيانات . في بحثنا استخدمنا بيانات حقيقية بنسبة كبيرة مع بعض التغييرات البسيطة في مجالات معينة ليس لها تأثير كبير على طبيعة البيانات من أجل الخروج من سريتها وفي نفس الوقت لتكون في التوافق مع ما هو مطلوب في بحثنا. واستخدمنا خوارزميات متعددة لتوسيع نطاق المقارنة من أجل اختيار الأفضل بينها وكذلك لتنوع النتائج بسبب تنوع البيانات التي اعتمدناها في بحثنا. من ناحية أخرى ، تم اختبار عدة خوارزميات لتحديد الخوارزمية التي تعطي تنبؤًا أكثر دقة ، خاصة أن الجهد المبذول في تنوع الخوارزميات كان لاختيار الأفضل من بينها ليعتمد في العمل المستقبلي.

*Corresponding author email : itpg.mohammed.haron@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://bjrs.uobasrah.edu.iq