Iraqi Journal of Intelligent Computing and Informatics (IJICI)
Vol. 1, 1. 1, December 2022, pp. 41~55
ISSN: 2791-2868, DOI: 10.52940/ijici.v1i2.11

Big Data in E-government: Classification and Prediction using Machine Learning Algorithms

Mohammed H. Altamimi, Maalim A. Aljabery, and Imad S. Alshawi
Department of Computer Science, College of Computer Science and Information Technology, University of Basrah
Basrah, IRAQ

Article Info

Article history:

Received August 15, 2022
Revised September 10, 2022
Accepted September 25, 2022

Keywords:

Big data
Classification
Data mining
E-government
Machine learning
Prediction

ABSTRACT

Many countries have used big data to develop their institutions, such as Estonia in policing, India in health care, and the development of agriculture in the United Kingdom, etc. Data is very important as it is no longer oil that is the most valuable resource in the world, but data. This research examines ways to develop the Iraqi state institutions by using the big data of one of its institutions (electronic civil registry) (ECR) in the national identity using the mining and analysis of this data. The pre-processing and analysis of this data are carried out depending on the needs of each institution and then using Machine Learning (ML) techniques. Its use has shown remarkable results in many areas, especially in data analysis, classification and forecasting. We applied five ML algorithms that are Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), and Naive Bayes (NB) carried out by python language and the Orange Data Mining (DM) tool. Choosing the workbook with the highest accuracy to work with to predict the needs of three government departments (military, social welfare, and statistical planning). According to the simulation results of the proposed system, the accuracy of the classifications was 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.

Corresponding Author:

Mohammed H. Altamimi
Department of Computer Science, College of Computer Science and Information Technology, University of Basrah Basrah, IRAQ
Email: itpg.mohammed.haron@uobasrah.edu.iq

1. INTRODUCTION

refers DM to extracting or mining knowledge from large amounts of data. It is also known as Knowledge-Discovery in Databases or Knowledge Discovery and DM. 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. 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]. 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 DT classifier [8], [9] and Neural Network classifier [10], [11]. In this paper, the researchers propose to create e-government from three government departments (Military, Social Welfare, and Statistics_ planning) by applying ML for five classifiers (SVM, DT, KNN, RF, and NB), using the ECR data, and choosing the algorithm with the highest accuracy for all government departments. 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. 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 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. Ayman Mir et al. [12] explain that built a classification model using WEKA tools, to classify medical 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, whereas experimental results show that the Support Vector Machine has the highest accuracy. Mucahid Mustafa et al. [13] confirm that their study is utilized to evaluate the likelihood of getting breast cancer by using anthropometric data and standard blood analysis factors. The classification performance of Artificial Neural Networks (ANN) and Naïve Bayes classifiers was calculated and compared on data with 9 inputs and one output. Pooja Thakar et al. [14] introduce a comprehensive survey, Journey (2002-2014) towards the exploration of educational data and its future scope. Mohammad Sultan et al. [15] present a comprehensive survey of the methods and techniques of data partitioning and sampling concerning big data processing and analysis. Fadi Salo et al. [16] 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. [17] focus on traffic management approaches based on DM and ML techniques to detect and predict traffic. Mr. Sudhir et al. [18] 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. [19] 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. [20] 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. [21] 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. [22] propose a secure framework for mining cloud data in a privacy-preserving manner. Secure KNN classifier is used. Mohammed Z. Al-Faiz et al. [23] work to achieve different motions of the prosthetic arm by better classification with multiple factors using K nearest neighbor. Maalim A. Aljabery et al. [24] choose the type of hearing aids that patients need by DM techniques.

3. PROPOSED METHOD

In this section, we discuss the main components of the proposed system of ECR data processing and classification by ML algorithms. we using an hp computer running Windows 10 and a core i5 processor, using python language and the Orange DM tool, and runs repetition 5 times. Five algorithms were randomly selected (SVM, DT, KNN, RF, and NB). For the purpose of establishing an e-government from three government departments (military, social welfare, and statistical planning). The proposed system consists of five main phase, as shown in Figure 1.

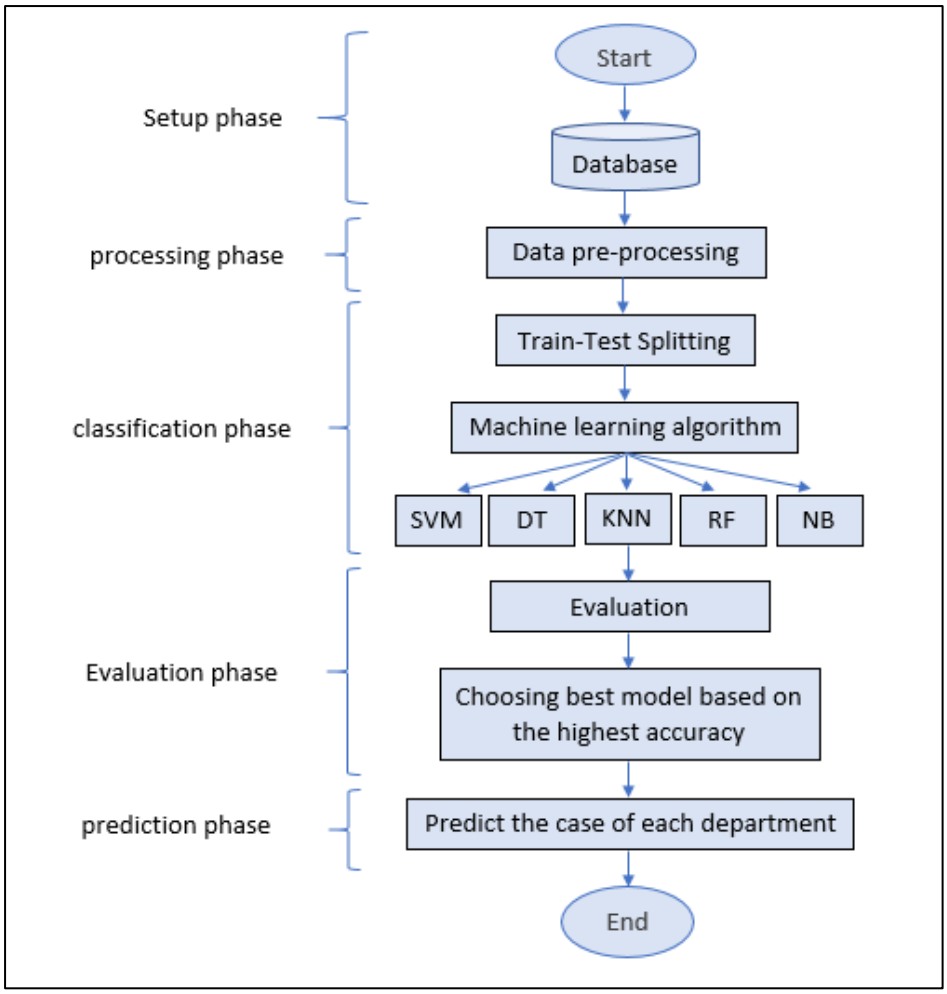


Figure 1. The proposed system phase

3.1. Setup Phase

In this phase, the database through which the proposed system is working is initialized, which is ECR data.

.

3.1.1. Data Collection

The database was created with a structure like the real data of the national identity data ECR 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 Figure 2, each record 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	Falih	Muhammad	Zahir	Al-Halfi	retag	ali	mortada	...	Kurdish	124921	796486	26414
1	198158372488	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-Malki	isra	rahim	tahsin	...	Kurdish	297168	236707	83409
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9994	198542764483	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	?	ruqii	abdallatif	nsaif	...	Arabic	430723	347434	73620
9997	193028959752	female	iyman	eihab	nabil	riyad	?	fuda	hadi	raoof	...	?	319408	932646	23653
9998	200557048552	female	narjis	abdalhusiain	mortada	faris	?	zanib	abdal	nabil	...	?	818078	856774	10297

9999 rows × 46 columns

Figure 2. Database of ECR

3.2. Data Pre-Processing Phase

The primary function of this phase, is enhance data that existed in the data set.

3.2.1. Inconsistent Data Processing

In this step, the inconsistent, incomplete, and missing data are processed such as deleting the duplicate records and that contain missing data by DM tools to get more accurate data. This 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 Figure 3.

	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
0	193629945390	female	Haider	Rashid	Badan	Hassan	Al-Fakhri	om albnin	alkarim	eihab	...	Kurdish	172975	816685
1	198759100110	?	Amer	Najm	Abdullah	Mohammed	Al	hind	hadi	riyad	...	Arabic	746107	429923
2	197662783958	female	Ali	Nasser	Musa	Muhammad	Al-Bouhiyah	hnadi	abdalabas	gfar	...	Arabic	541656	398759
3	195568327873	?	Najat	Nima	Saleh	Aboud	Al	thani	zbali	hadi	...	Arabic	963692	647037
4	198523472469	female	Nasser	Gary	Obaid	as	Eid	rfil	raid	abdalabas	...	Turkmen	746128	872769
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
494	196663008621	Male	Jassim	Hassan	Muhammad	Ghaleb	Al-Maliki	ahlam	abdal	ahmaid	...	Turkmen	464157	247025
495	194237395863	Male	Jassim	Hassoun	Latif	Saadoun	Al-Mayi	retag	husiain	raoof	...	Arabic	505817	833007
496	197366671877	Male	Jassim	Salem	Sender	of	Shakban	narjis	alkarim	abdalabas	...	Arabic	348438	439992
497	198586942824	Male	Jassim	Sherida	Jassim	Mohammed	Al-Mohammed	fuda	abdalabas	gmal	...	Arabic	627289	924274
498	201469274079	?	Jassim	Sharif	Abdul	Hassan	Al-Husainat	rwan	hbib	talib	...	Arabic	217238	294615

499 rows × 46 columns

Figure 3: Data Warehouse of ECR

3.2.2. Normalization

In this step, perform normalizing, it means transforming the data, namely converting the source data in to another format that allows processing data effectively. 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 [25], as shown in Figure 4. Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbor classification and clustering [6].

Mother's_father	mother's_grandfather	...	Living_country	Passport_No	Telephone_No	male	healthily	death	free_work	Living_iraq	age_18_28	Living_basrah
alkarim	eihab	...	Kuwait	A50062035	7805617216	1	1	1	0	0	0	0
hadi	riyad	...	iraq	A71495003	7806671726	0	1	1	0	1	0	0
abdalabas	gfar	...	iraq	A30107795	7803677076	1	1	1	1	1	0	1
zbali	hadi	...	Sweden	A62779580	7801093325	0	0	0	1	0	0	0
raid	abdalabas	...	iraq	A76733252	7809946500	1	1	1	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
abdal	ahmaid	...	iraq	A34188742	7805506023	1	0	1	1	1	1	0
husiain	raoof	...	iraq	A88728022	7805182510	1	0	1	1	1	1	0
alkarim	abdalabas	...	iraq	A87955318	7802912666	1	1	1	1	1	0	0
abdalabas	gmal	...	iraq	A45831606	7806996245	1	1	1	1	1	0	0
hbib	talib	...	iraq	A51431646	7804625292	0	1	0	0	1	0	1

Figure 4: Normalizing.

3.2.3. Extract features and target building

In this step, to enhance the quality of the data, and to build an accurate classifier, the important features of each government department must be extracted. As each government department has the information it needs from the big data set.

3.3. Classification Phase

The core phase of the proposed system includes building a classify and prediction model the needs of three government departments (Military, Social Welfare, and Statistics-Planning) for big data ECR on two sides (online and offline) This phase includes steps:

3.3.1. Train-test splitting

In this step 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 [26]. 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 Figure 5.

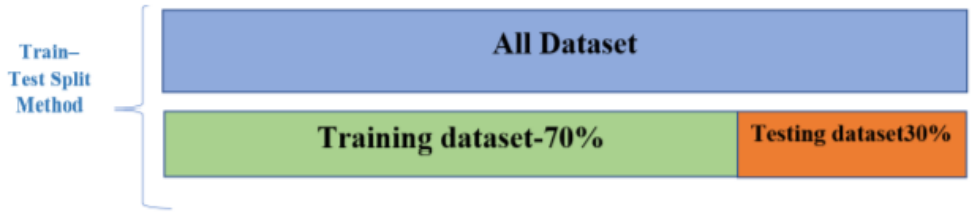


Figure 5: Train-Test Splitting.

The training set feeds the model to train the ML algorithms to learn and by extracted features and use them in the classification process. When the training stage is completed, will validate to Proposed System Via test set. This validation dataset involves altering or neglecting variables and tuning model settings until the process reaches a suitable accuracy level.

3.3.2. Building a classifier using machine learning algorithms

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 [2].

The primary task of the proposed system is to build a classifier that meets the needs of the three government departments (Military, Social Welfare, and Statistics-Planning) from ECR data. This is done by testing a set of ML algorithms. Five ML algorithms were selected randomly: SVM, DT, KNN, RF, and NB

Each algorithm is applied to the data set generated from the previous steps using Python language. The proposed system works on two sides (online and offline) in order to maintain the confidentiality and privacy of information citizens and not to disclose them, except within the limits of the desired targets. Then choose the algorithm with the highest accuracy for each government department through the offline side. The chosen algorithm is being worked on in the future through the online side to share the results with the relevant government department, as shown in Figure 6.

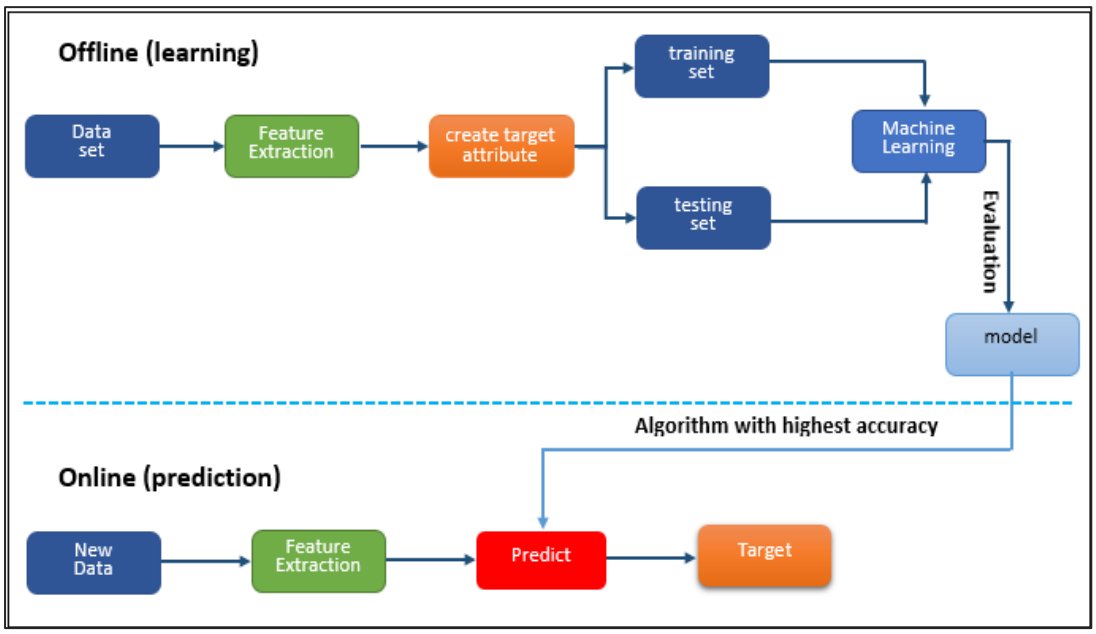


Figure 6: Classification Model

3.4. Evaluation Phase

In this phase, the five previously selected algorithms are evaluated randomly: SVM, DT, KNN, RF, and NB. that were previously applied in the classification phase. this is done by calculating the Confusion Matrix, Accuracy, Recall, Precision and F-Score for each algorithm. The benefit of this phase is to compare the algorithms applied to the features of each of the three government departments. And then choose the algorithm with the highest accuracy to work with in the future to contact the government department online.

3.4.1. Choosing best model based on the highest accuracy

In this step, choosing the best model that gave the highest accuracy among the other models. To install it and work with it in the future in the online side. After the model is well trained on a sample of ECR data and gives good results, this model is able to classify and predict the government department's need of the ECR data entered into it. So that each government department has its own model that works on predicting its needs. The benefit of this step is to build a model with good capabilities in predicting the needs of the government department.

Orange DM program was used to compare the results between the five algorithms SVM, DT, KNN, RF, and NB. The results were similar to the Python program in a pictorial form, through which it is possible to know the results with ease. [27].

3.5. Prediction Phase

In this phase, and after we have chosen the best model for working with the concerned government department, where the prediction is as follows:

1- Military department

The proposed system predicts the citizens covered by the military within the age group and conditions (features), chosen by the Military Department, the result is Armed soldier or unarmed soldier or not a soldier.

2- Social Welfare department

The proposed system predicts the citizens covered by the Social welfare within the age group and conditions (features), chosen by the Social Welfare Department, the result is eligibility for social insurance, first or second or third or fourth or fifth degree.

3- Statistics-planning department

The proposed system predicts the citizens covered by the Statistics and planning within the age group and conditions (features), chosen by the Statistics and planning department, the result is baby girl, baby boy, teenage girl, teenage boy, young woman, young man, woman, man, old woman, old man. and The Following Figure 7 depicts the workflow in detail for our proposed system.

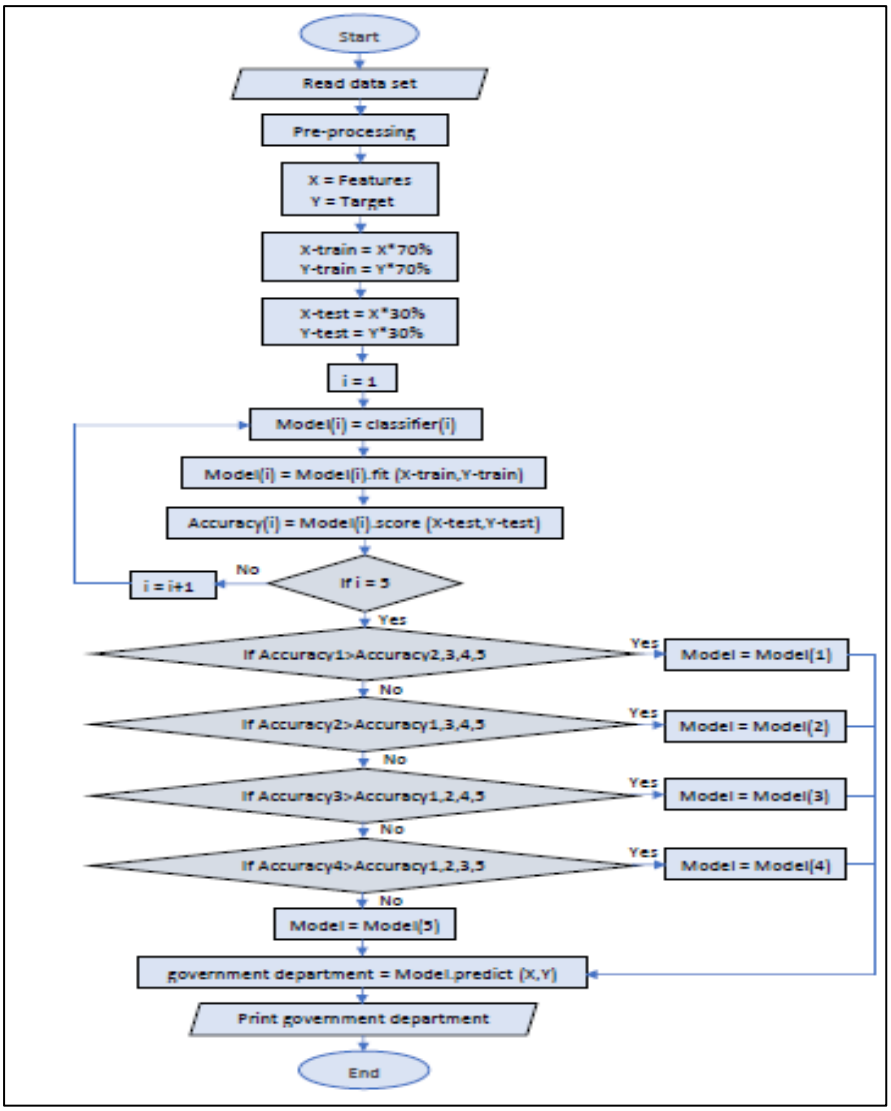


Figure 7: workflow in the proposed system.

3.6. Prediction Phase

After training the model for each government department of the three departments (Military, Social Welfare, and Statistics-Planning), which obtained the highest accuracy. Work is being done to create websites for each government department that will display the results that were predicted via the online side. Data is initialized and SVM, DT, KNN, RF, and NB algorithms are trained and the model with the highest accuracy is selected in the offline side in the National ID Department in order to maintain the confidentiality and privacy of citizens' data. The results of each department are being prepared through the department's website, in the online side. This data is updated over time and as a result of the ECR data refresh. The government department cannot modify the original data, but it can use the data that appears on its website, such as copying, printing and other uses.

4. RESULTS AND DISCUSSION

Based on the above test of five ML algorithms (SVM, DT, KNN, RF, and NB) on the ECR data for classification and prediction of the needs of three government departments (Military, Social Welfare, and Statistics-planning), shows the proposed system the following results.

4.1. Military

Offline way, the SVM algorithm shows higher accuracy (100%) Compared to other algorithms in the government Military department, as shown in Figures 8, 9, and 10:

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

Figure 8: 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

Figure 9: Confusion Matrix of SVM

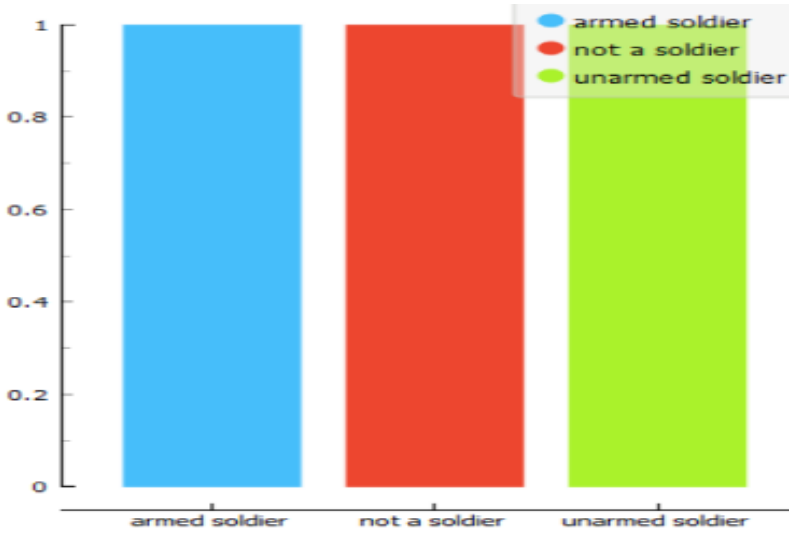


Figure 10: Distribution of SVM

After selecting the SVM algorithm that showed the highest accuracy, the system predicts the needs of the military department, and then shares the results via the department's website in an online way, as shown in Figures 11 and 12:

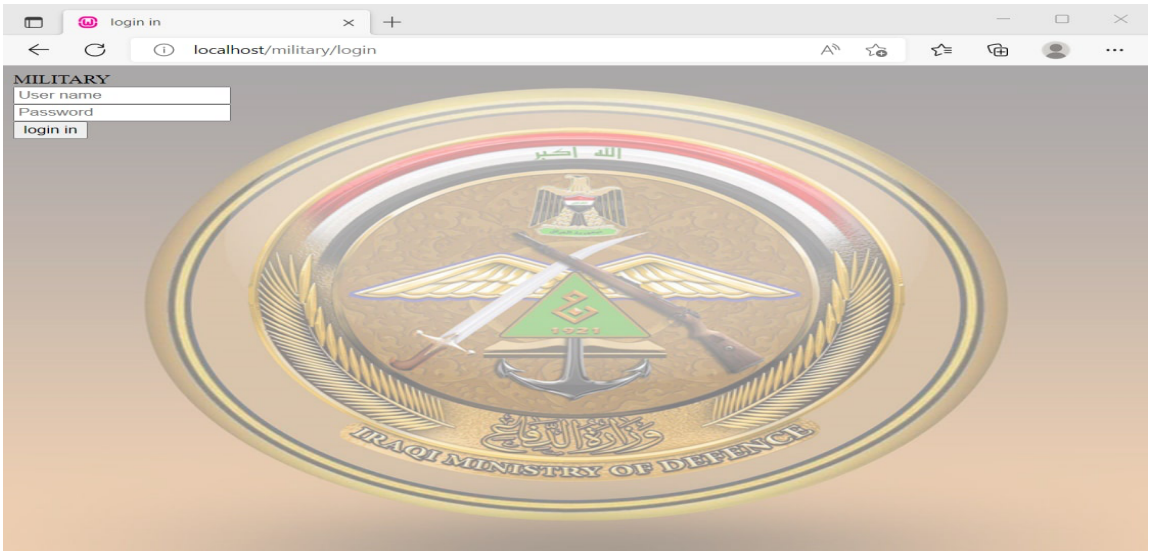


Figure 11: Website login to the military department

	Name	Mother_name	Birth_date	Living_governorate	Class_military
1	Mohammed Kazem Abdullah Abdul alhlichi	duha abd raoof	2001	Basrah	armed soldier
2	Ibrahim Salem Taher Mohamed altimimi	sara nashit hbib	2001	Basrah	armed soldier
3	Ibrahim Abdul-Zahra Abdul-Hassan Butti Al-Bahadli	ruidi kamil nashit	2001	Babylon	armed soldier
4	Ibrahim Mohamed Sayhoud Yahya developer	qmair gfar nsaif	2001	Basrah	armed soldier
5	Ahmed Jassim Dagher rahim alfriji	thani nsaif kmal	2002	Basrah	armed soldier
6	Ahmed Jassim Mohammed Mahmoud alharoon	hager abdalkarim abdalabas	2002	Maysan	armed soldier
7	Ahmed Jabbar Saeed Jbara Al-Asadi	hyat abdal rahim	1999	Maysan	armed soldier
8	Ahmed Jabbar Awaja Hamidi Al-Hamidi	kukab nabil kamil	1998	Basrah	armed soldier
9	Ahmed Jabbar Qassem Mohammed Al	ruqii abdalhusiain hadi	1994	Dhi Qar	armed soldier
10	Ahmed Chard Dhahi Othman Al-Abadi	ruqii nashit abass	2002	Basrah	armed soldier
11	Ahmed Abdel-Razzaq Abdel-Sada Ibrahim alsaiid	qmair zbali abdal	2002	Karbala	armed soldier
12	Ahmed Alwan, dressed by Hamil	kukab abd nashit	2002	Baghdad	armed soldier
13	Asaad Saddam Adai Abbas al-Maliki	sara raoof abdalabas	2000	Basrah	armed soldier

Figure 12: Data website of the military department

4.2. Social welfare

Offline way, the random forest algorithm shows higher accuracy (99%) Compared to other algorithms in the government Social welfare department, as shown in Figures 13, 14, and 15:

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

Figure 13: 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

Figure 14: Confusion Matrix of Random Forest

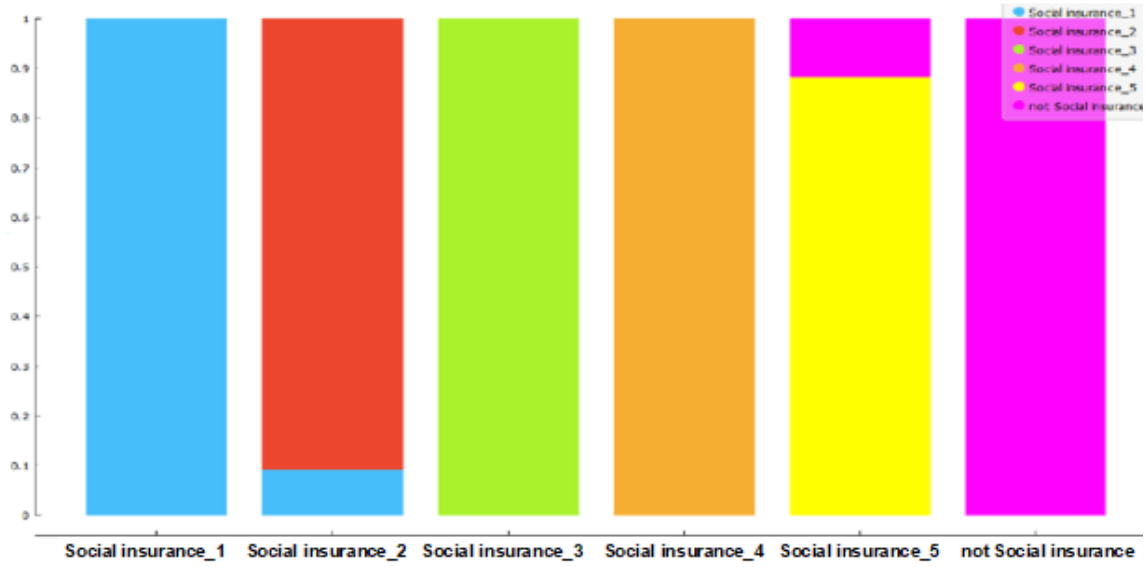


Figure 15: Distribution of Random Forest

After selecting the random forest algorithm that showed the highest accuracy, the system predicts the needs of the social welfare department, and then shares the results via the department's website in online way, as shown in Figures 16 and 17:

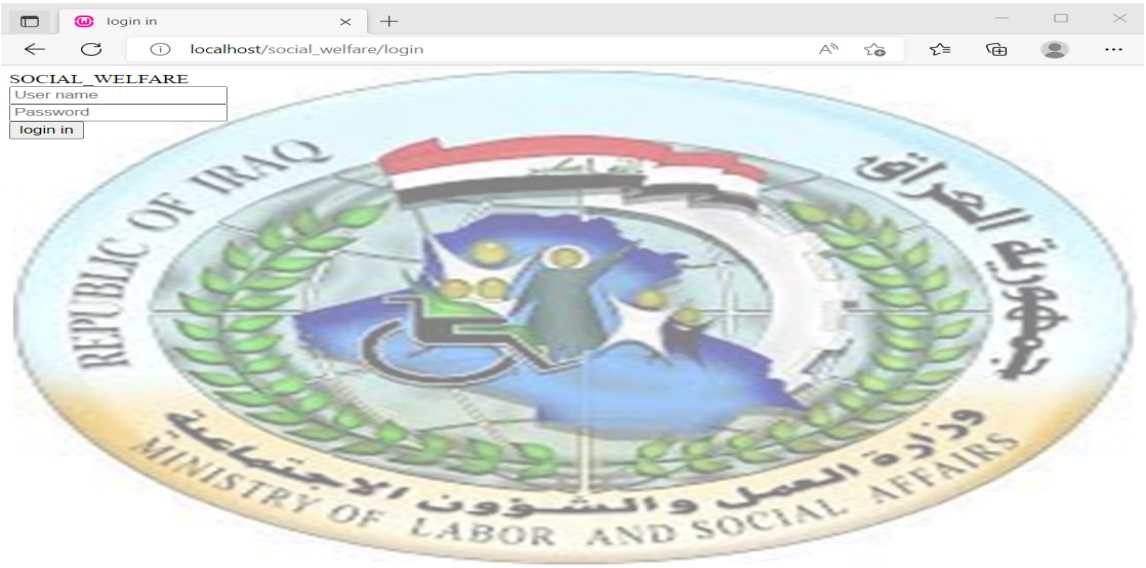


Figure 16: Website login to the social welfare department

	Name	Mother_name	Birth_date	Living_governorate	Class_social_welfare
1	Jinan Abdul-Hussein Idris Al-Battat alknani	rwan abdalhusiain nabil	1994	Basrah	Social insurance_1
2	Ibtisam Odah Hazal Mansour Sayed	kukab abdallatif riyad	1981	Basrah	Social insurance_1
3	hyat Hato Ghaji Habash Al-Aboudi	narjis abd yaseen	1989	Basrah	Social insurance_1
4	Anwar Fadel Jaafar Muhammad alknani	nsreen faris alkarim	2001	Basrah	Social insurance_1
5	Janan Laiby Handal Merz albhadli	hind loii gfar	1964	Basrah	Social insurance_2
6	Maan Salman Abdul hamid Razzaq	rfil gfar yaseen	1954	Babylon	Social insurance_2
7	ahlam Jassim Mohammed Mahmoud alharoon	hager abdalkarim abdalabas	2002	Maysan	Social insurance_2
8	somia Ali Hussein Sharqi Al	hnadi nabil abdalabas	1961	Basrah	Social insurance_2
9	Anwar Fouad Yaqoub Taher Al-Mansour	narjis rahim abdallatif	1998	Karbala	Social insurance_2
10	ruqai Mukheiber Jibril Sanafi Al-Rubaie	omalbnin allatif kamil	1999	Basrah	Social insurance_2
11	Qasimia Mohammed Saeed Areed Al	rfil jasm yaseen	1992	Basrah	Social insurance_3
12	Nour Al-Huda Abdul-Kadhim, the gesture	ruidi abdalkarim abdalabas	2001	Baghdad	Social insurance_3
13	Ibrahim Abdul-Zahra Abdul-Hassan Butti Al-Bahadli	ruidi kamil nashit	2001	Stockholm	Social insurance_3

Figure 17: Data website the social welfare department

4.3. Statistics and Planning

Offline way, The SVM algorithm shows higher accuracy (100%) Compared to other algorithms in the government statistics and planning department, as shown in Figures 18, 19, and 20:

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

Figure 18: Evaluation Results

		Predicted										
		baby boy	baby girl	man	no	old man	old woman	teenage boy	teenage girl	woman	young man	young woman	Σ
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
	Σ	50	30	55	90	60	60	40	20	40	185	120	750

Figure 19: Confusion Matrix of SVM

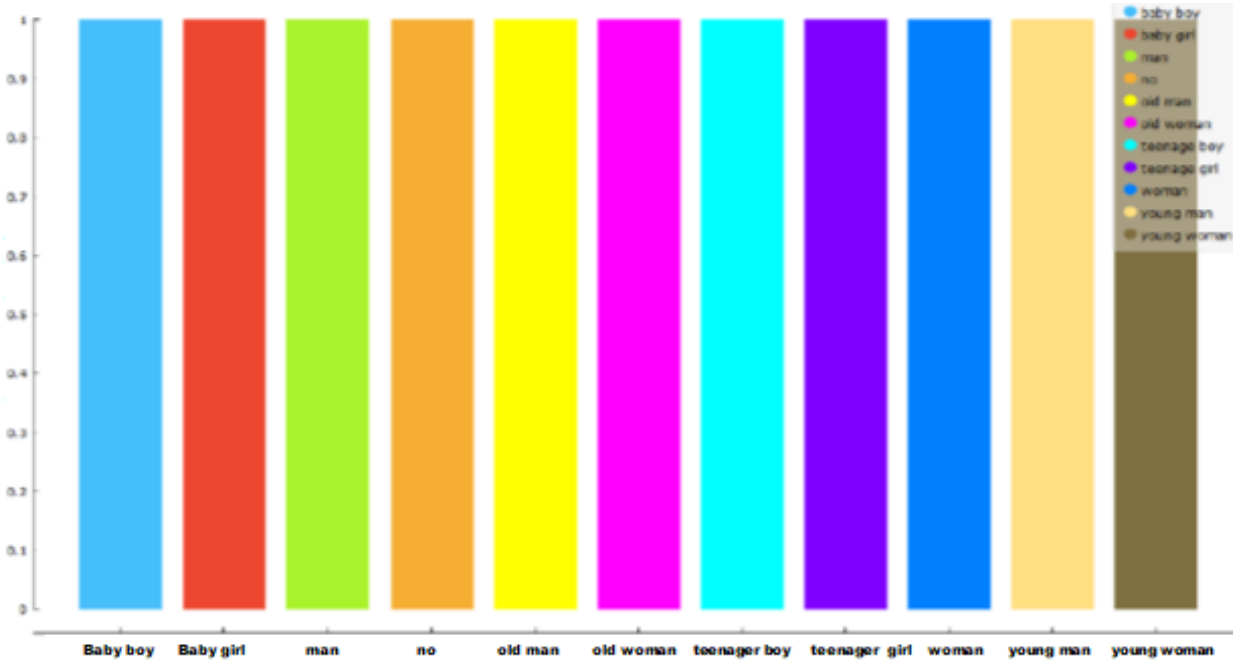
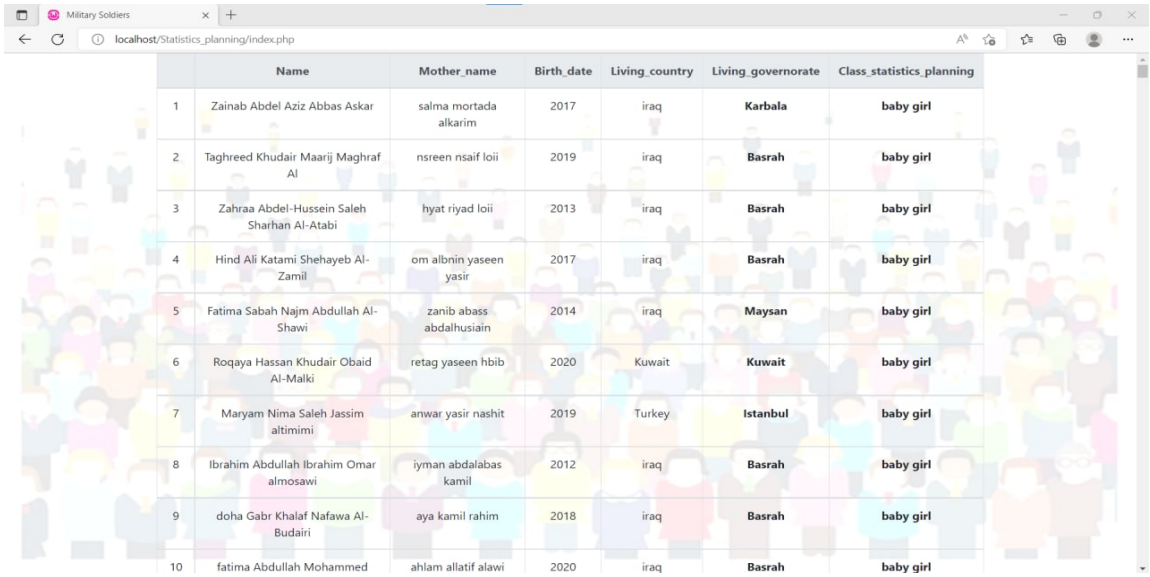


Figure 20: Distribution of SVM

After selecting the SVM algorithm that showed the highest accuracy, the system predicts the needs of the statistics and planning department, and then shares the results via the department's website in an online way, as shown in Figures 21 and 22:

Figure 21: Website login to the statistics and planning department



	Name	Mother_name	Birth_date	Living_country	Living_governorate	Class_statistics_planning
1	Zainab Abdel Aziz Abbas Askar	salma mortada alkarim	2017	iraq	**Karbala**	**baby girl**
2	Taghreed Khudair Maarij Maghraf Al	nsreen nsaif loii	2019	iraq	**Basrah**	**baby girl**
3	Zahraa Abdel-Hussein Saleh Sharhan Al-Atabi	hyat riyad loii	2013	iraq	**Basrah**	**baby girl**
4	Hind Ali Katami Shehayeb Al-Zamil	om albnin yaseen yasir	2017	iraq	**Basrah**	**baby girl**
5	Fatima Sabah Najm Abdullah Al-Shawi	zanib abass abdalhusiain	2014	iraq	**Maysan**	**baby girl**
6	Roqaya Hassan Khudair Obaid Al-Malki	retag yaseen hbib	2020	Kuwait	**Kuwait**	**baby girl**
7	Maryam Nima Saleh Jassim altimimi	anwar yasir nashit	2019	Turkey	**Istanbul**	**baby girl**
8	Ibrahim Abdullah Ibrahim Omar almosawi	iyman abdalabas kamil	2012	iraq	**Basrah**	**baby girl**
9	doha Gabr Khalaf Nafawa Al-Budairi	aya kamil rahim	2018	iraq	**Basrah**	**baby girl**
10	fatima Abdullah Mohammed	ahlam allatif alawi	2020	iraq	**Basrah**	**baby girl**

Figure 22: Data website the statistics_ planning department

5. CONCLUSION AND FUTURE WORK

the process of discovering the hidden knowledge in the data know DM Including ECR data[28]. 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 in the national identity 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. For future work, it seems that the real data of the ECR can be used and shared with all government departments, i.e. under the need of each department. 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 and M. S. S. Sherekar, “Comparative Analysis of Data Mining Tools and Techniques for Evaluating Performance of Database System,” Int. J. Comput. Sci. Appl., vol. 6, no. 2, pp. 2–7, 2013, [Online]. Available: www.researchpublications.org.

[2] C. Alexopoulos, V. Diamantopoulou, Z. Lachana, Y. Charalabidis, A. Androutsopoulou, and M. A. Loutsaris, "How machine learning is changing e-government," ACM Int. Conf. Proceeding Ser., vol. Part F1481, pp. 354–363, 2019, dio: 10.1145/3326365.3326412.

[3] K. Altmemi and I. S. Alshawi, “Enhance Data Similarity Using a Fuzzy Approach,” Journal of Positive School Psychology, vol. 6, no. 5, pp. 1898–1909, 2022.

[4] M. R. Rajagopalan and S. Vellaipandiyan, "Big data framework for national E-governance plan," Int. Conf. ICT Knowl. Eng., 2013, dio: 10.1109/ICTKE.2013.6756283.

[5] M. D. Aljubaily and I. S. Alshawi, “Energy sink-holes avoidance method based on fuzzy system in wireless sensor networks,” Int. J. Electr. Comput. Eng., vol. 12, no. 2, pp. 1776–1785, 2022, doi: 10.11591/ijece.v12i2.pp1776-1785.

[6] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Elsevier Science, 2011.

[7] I. S. Alshawi, Z. A. Abbood, and A. A. Alhijaj, “Extending lifetime of heterogeneous wireless sensor networks using spider monkey optimization routing protocol,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 20, no. 1, pp. 212–220, 2022, doi: 10.12928/TELKOMNIKA.v20i1.20984.

[8] N. Indumathi, R. Ramalakshmi, and V. Ajith, “Analysis of risk factors in the Firework Industries: Using Decision Tree Classifier,” in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 811–814, doi: 10.1109/ICACITE51222.2021.9404726.

[9] H. H. Al-badrei and I. S. Alshawi, “Improvement of RC4 Security Algorithm,” Adv. Mech., vol. 9, no. 3, pp. 1467–1476, 2021.

[10] I. Alshawi, H. Allamy, and R. Z. Khan, “Development Multiple Neuro-Fuzzy System Using Back-propagation Algorithm,” Int. J. Manag. Inf. Technol., vol. 6, pp. 794–804, 2013, doi: 10.24297/ijmit.v6i2.736.

[11] Z. A. Abbood, I. S. Alshawi, A. A. Alhijaj, and F. P. Vidal, “Automatic sound synthesis using the fly algorithm,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 18, no. 5, pp. 2439–2446, 2020, doi: 10.12928/TELKOMNIKA.V18I5.15665.

[12] A. Mir and S. N. Dhage, “Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare,” Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697439.

[13] M. Saritas and A. YASAR, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,”

Int. J. Intell. Syst. Appl., vol. 7, 2019.
[14] P. Thakar, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," vol. 110, no. 15, pp. 60–68, 2015, [Online]. Available: http://arxiv.org/abs/1509.05176.
[15] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," Big Data Min. Anal., vol. 3, no. 2, pp. 85–101, 2020, doi: 10.26599/BDMA.2019.9020015.
[16] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review," IEEE Access, vol. 6, pp. 56046–56058, 2018, doi: 10.1109/ACCESS.2018.2872784.
[17] N. O. Alsrehin, A. F. Klaib, and A. Magableh, "Intelligent Transportation and Control Systems Using Data Mining and Machine Learning Techniques: A Comprehensive Study," IEEE Access, vol. 7, pp. 49830–49857, 2019, doi: 10.1109/ACCESS.2019.2909114.
[18] I. Journal, "A Study Some Data Mining Classification Techniques," Int. J. Mod. Trends Eng. Res., vol. 4, no. 1, pp. 210–215, 2017, doi: 10.21884/ijmter.2017.4031.zt9tv.
[19] A. H, Q. A., M. N., and E. M., "A Comparison Study between Data Mining Tools over some Classification Methods," Int. J. Adv. Comput. Sci. Appl., vol. 1, no. 3, 2011, doi: 10.14569/specialissue.2011.010304.
[20] I. Garcia-Magarino, G. Gray, R. Lacuesta, and J. Lloret, "Survivability Strategies for Emerging Wireless Networks with Data Mining Techniques: A Case Study with NetLogo and RapidMiner," IEEE Access, vol. 6, pp. 27958–27970, 2018, doi: 10.1109/ACCESS.2018.2825954.
[21] M. S. Basarslan and I. D. Argun, "Classification of a bank data set on various data mining platforms | Bir Banka Müşteri Verilerinin Farkli Veri Madenciliği Platformlarinda Siniflandirilmasi," 2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018, pp. 1–4, 2018.
[22] S. A. W. Saddam, "Secure mining of the cloud encrypted database," J. Basrah Res. Journal of Basrah Researches (Sciences), vol. 43, no. 2A, pp. 44–57, 2017.
[23] M. Z. Al-Faiz, A. A. Ali, and A. H. Miry, "A k-nearest neighbor based algorithm for human arm movements recognition using EMG signals," EPC-IQ01 2010 - 2010 1st Int. Conf. Energy, Power Control, no. May 2019, pp. 159–167, 2010, doi: 10.37917/ijeee.6.2.12.
[24] M. A. Aljabery and S. Kurnaz, "Applying datamining techniques to predict hearing aid type for audiology patients," J. Inf. Sci. Eng., vol. 36, no. 2, pp. 205–215, 2020, doi: 10.6688/JISE.202003_36(2).0002.
[25] N. Noori and A. Yassin, "Towards for Designing Intelligent Health Care System Based on Machine Learning," Iraqi J. Electr. Electron. Eng., vol. 17, no. 2, pp. 120–128, 2021, doi: 10.37917/ijeee.17.2.14.
[26] B. Vrigazova, "The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems," Bus. Syst. Res., vol. 12, no. 1, pp. 228–242, 2021, doi: 10.2478/bsrj-2021-0015
[27] S. Kodati and R. Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka," Glob. J. Comput. Sci. Technol. C Softw. Data Eng., vol. 18, no. 1, pp. 16–22, 2018.
[28] S. Kurnaz and M. A. H. Aljabery, "Predict the type of hearing aid of audiology patients using data mining techniques," ACM Int. Conf. Proceeding Ser., pp. 2–6, 2018, doi: 10.1145/3234698.3234755.