Forest Fires Detection Using Machine Learning Techniques

Ahmed M. Elshewey
Computer Science Department,
Faculty of Computers and Information, Suez University, Egypt.
elshewy86@gmail.com

Amira. A. Elsonbaty
Communication & Electronics Department,
Higher institute of engineering and technology,
New Damietta, Egypt, 34517
dr_eng.amira@yahoo.com

Abstract- Nowadays, forest fires became one of the foremost important problems that cause damage to several areas around the world. The paper displays machine learning regression techniques for predicting forest fire-prone areas. The data set used in this paper is presented within the UCI machine learning repository that consists of climate and physical factors of the Montesinos park in Portugal. This research proposes three machine learning approaches, linear regression, ridge regression, and lasso regression algorithm with data set size 517 entries and 13 features for each row. This paper uses two versions, all features are included in the first, and 70% of the features were included in the second. The paper uses a training set which is 70% of the data set, and the test set is 30% of the data set. The accuracy of the linear regression algorithm gives more accuracy than ridge regression and lasso regression algorithms.

Keywords- Machine Learning, Linear Regression, Ridge Regression, Lasso Regression, Forest Fires, Algorithms.

I. INTRODUCTION

One of the most extremely occurring disasters in recent times is forest fires (wildfires). Due to these wildfires, a lot of acres of forest area are getting destroyed. The significant reasons that lead to the occurrence of forest fires are warming due to the increase in the average temperature of the earth [1], and human negligence. Dynamic Integrated Model of Climate and the Economy (DICE) indicates that the economy will lose about $23 trillion in the next 80 years due to the change in climate [6]. In Africa, South America, Southeast Asia, and New Zealand, forest fires occur due to human factors like husbandry of animals and agriculture [7]. Nowadays, there are various technologies for fire modelling to predict the spread of fires, such as physical models and mathematical models [8]. These models depend on data collection during forest fires, simulation, and lab experiments to specify and predict fire growth in many regions. Recently, simulation tools have been used to predict forest fires, but simulation tools faced some problems such as the accuracy of input data and simulation tool execution time [7]. Machine learning is a sub-branch of Artificial Intelligence (AI) to learn computers aspect. Machine learning can be divided into two classes: supervised, unsupervised and reinforcement. In supervised learning, a supervisor is existed to give insights to the learning algorithm on how a decision or an action is bad or good. In supervised learning, the whole the data set is labelled completely. Supervised machine learning algorithms are as linear regression, Support Vector Machine (SVM), Artificial Neural Networks (ANN) and decision trees. In unsupervised learning, the data set is not labelled. This leads that the algorithm must define the labels. The structure of the data set and the relationship between the features will be learned by the algorithm. Unsupervised machine learning algorithms are as k-means clustering and Self-Organizing Map (SOM). In reinforcement learning, the learning algorithm gets punished in case of a wrong action and gets rewarded in case of correct action. Fig 1 shows the areas of machine learning.

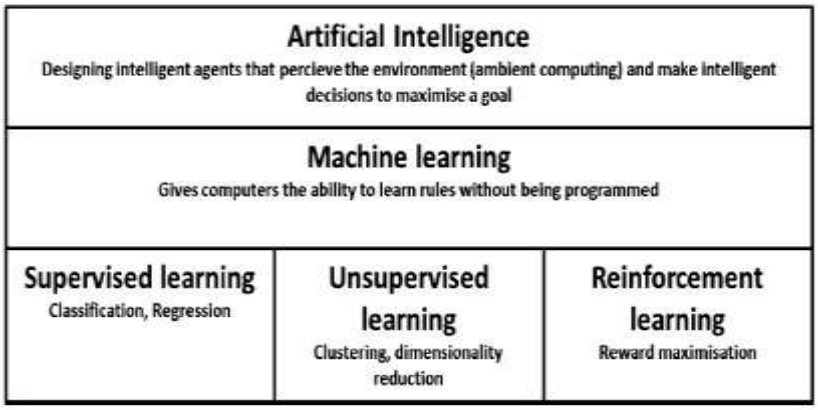

Figure1. Machine learning areas

Data mining is one of the most significant approaches such as forest fires can be predicated upon their occurrences [9, 10]. Data mining requires real and clean data for making a prediction. If the data set contains many unknown values, then these values must be ignored or imputed before using them in the modelling. The workflow of data mining goes through several steps. These steps are data collection, cleansing, transformation, aggregation, modelling, predictive analysis, visualization and dissemination. Figure 2 demonstrates the steps of data mining.

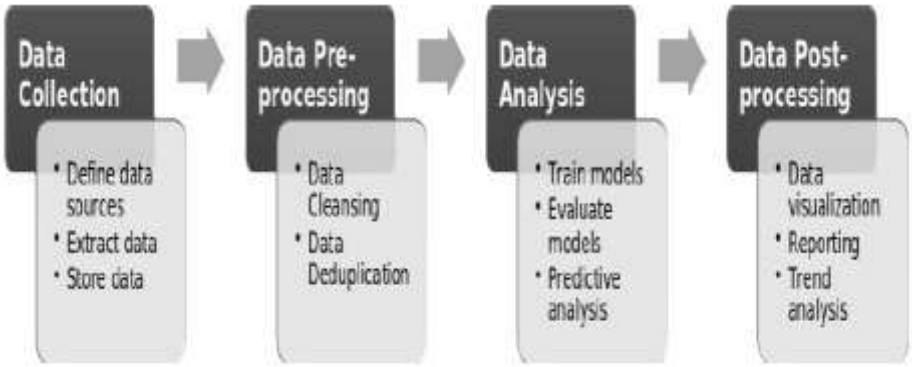

Fig 2. Steps of data mining.

The rest of the paper is organized as the following: Literature Survey section that discusses the related works, Methodology section that displayed the proposed algorithms, Results and Analysis section that discusses the results, and a Conclusion section that contains a summary of the paperwork.

II. LITERATURE SURVEY

Burned areas in forest fires were predicted using estimation methods as the Multilayer Perceptron (MLP), SVM, Radial Basis Function (RBF) networks and fuzzy logic [11]. The results indicate that MLP gives more accurate results. Available, and Reliable Storage for an Incompletely Trusted Environment (FARSIGHT) simulator was used to predict forest fires spread in the Euro-Mediterranean countries [12]. The outputs of FARSIGHT were obtained by two models, custom fuel model and standard fuel model. The experimental results showed that the accuracy of the custom fuel model was better than the standard fuel model. An intelligent system called geometric semantic genetic programming to predict burned areas [13]. The results obtained using that intelligent system were better than using standard genetic programming. A novel system called forecast to predict the spread of forest fires in the future [14]. The forecast is a system that combines Artificial Intelligence (AI) and Geographic Information Systems (GIS). The forecast obtained more accurate results when compared to other random prediction models. A machine-learning algorithm based on Wireless Sensor Networks (WSN) to predict

forest fires [15]. A fire prediction tool called Disjunctive Normal Form (DNF) model to predict forest fires [16]. The results obtained from the DNF model were compared with other machine learning models as naive Bayes, decision tree, SVM, RBF, and polynomial kernel functions. The DNF model gave the highest average accuracy with 97.8% among the other machine learning models. An algorithm that depends on SVM to predict forest fires [17]. SVM used two class predictions of fire risk. The results demonstrated that the accuracy of SVM was approximately 96%. ANN model was used to predict the size of burned areas of forest fires in southern Spain [18]. ANN was used in two stages: classifying forest fires size and evaluation of the burned surface areas. The results mentioned that the process of prediction was over 60%, prediction can reach more than 70% in some central areas. A probabilistic model was used to predict forest fires [19]. There were three steps to design the probabilistic model. In step 1, the probabilistic model of forest fires was built from data of weather forecast and historical satellite. In step 2, the prediction of forest fires was produced using the data of the weather forecast as an input in the model of forest fires. In step 3, the warnings of forest fires were transported on different levels based on the need of the user. Machine learning models to predict the size of forest fires at the time of their inflammation [20]. Decision trees, random forests, and MLP models were used in the process of prediction. The decision tree model predicted that 40% of the inflammation led to a large number of fires, and this per cent is about 75% of the total burned area. Random forests and MLP models were tested, but they did not perform the accuracy as the decision tree model. Different machine learning models to predict forest fires in Slovenia [21]. Logistic regression, decision tree, random forests, bagging, and boosting of decision tree models were used to predict forest fires in Slovenia. These models were applied to these three data sets: Kras region, Primorska region, and continental Slovenia. From the experimental results, the bagging decision tree model obtained the best accuracy for all the data sets. Semiparametric models were used to predict forest fires [22]. Two semiparametric models that depend on time series were used to predict the burned area every week per year. The experimental results obtained show that the first semiparametric model accuracy in results was better than the second semiparametric model, where the errors were lower in the first semiparametric model. Several machine learning models to predict forest fires [23]. SVM, decision tree, regression, ANN, etc. Models were used for the prediction of forest fires. The accuracy of regression was better when compared to the other machine learning models. Five machine learning models to predict forest fires, namely, MLP, RBF, SVM, Polynomial Neural Network (PNN), and Cascade Correlation Network (CCN) [24]. The Principal Component Analysis (PCA) model was used to find the best patterns in the data set and the Particle Swarm Optimization (PSO) model was used to make segmentation the fire regions. The experimental results showed that the SVM model was more effective than other machine learning models.

III. METHODOLOGY

Machine learning models play a major role in the process of evaluation and prediction. Prediction is often done by using the available variables within the data set. Through the available variables within the data set, machine learning models can make predictions for the long term [26]. In this section, linear regression, ridge regression, and lasso regression are presented.

3.1. Linear Regressions

Regression analysis is the process of statistical analysis to evaluate the relationship between various variables. Nowadays, regression analysis models are being widely used for prediction in the field of machine learning. The concept of regression analysis is to show how the dependent variable value varies when one independent variable value changes, where the other variables are restricted [27]. Also, regression analysis is used to compute the dependent variable, the average value when the independent variables are restricted. The linear regression model is one of the most significant predictive analysis models. The linear regression model is a statistical model that explains the relationship between one dependent variable (or outcome variable) and one or more independent variables (or predictor variables). The main idea of the regression is to check two significant things: first, the performance of the independent variables while predicting the dependent variable. Second, the independent

variables are important for the dependent variable. If one independent variable has a linear relationship with one dependent variable, then the regression is called simple linear regression. If two or more independent variables have a linear relationship with one dependent variable, then the regression is called a multiple linear regression. In the linear regression model, if there is one independent variable, then the regression function is a straight line, if there are two independent variables, then the regression function is plane and if there are n independent variables, then the regression function is hyper-plane with n-dimensional. If there is fitting between the actual values and the predicted values, then the actual values will be similar to the predicted values. But if there is a difference between the actual values and the predicted values, this difference is called cost, loss, or error.

The regression function $\hat{y}$ dependent on n independent (predictor) variables $x_1, x_2, \ldots, x_n$ can be expressed as in Eq. 1:

$$\hat{y} = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n + b. \quad (1)$$

Eq. 1 represents how the value of $\hat{y}$ changes with the independent $x_1, x_2, \ldots, x_n$. $w_0, w_1, \ldots, w_n$ are called feature weights (model coefficients) and b is called a constant bias term (intercept).

An important concept in regression is Ordinary Least Squares (OLS), which is a statistical method that calculates the relationship between one dependent variable and one or more independent variables, the method calculates the relationship through minimizing the sum of the squares in the difference between the actual values and the predicted values of the dependent variable that represent a straight line. Also, OLS easily applied to multivariate models that contain two or more independent variables. OLS finds w and b that minimizes the Residual Sum of Squares (RSS) over the training data between the actual values and the predicted values. RSS can be expressed as in Eq. 2:

$$RSS\ (w, b) = \sum_{i=1}^{N} (y_i - (w x_i + b))^2. \quad (2)$$

3.2. Ridge Regression

Ridge regression is used to analyze data that is multiple regression, these data contain multicollinearity (independent variables are highly correlated). Ridge regression is a technique to reduce the complexity of the model and to avoid overfitting. Prediction of new values done by ridge regression technique gives good results when there is a correlation between the predictor variables [30]. Ridge regression learns the parameters w, b through using the same criterion of the least-squares with the addition of adding a penalty term to make a big variation in the parameter of w. The penalty term is called regularization, which restricts the model to prevent overfitting, and also regularization methods are used to control the coefficients of the regression, this will help to minimize the variance and reduce the sampling error [31]. Ridge regression uses L2 regularization, which minimizes the sum of the square of the coefficients [31]. L2 regularization has analytical solutions, thus L2 regularization is computationally efficient. RSS for ridge regression can be expressed as in Eq. 3:

$$\text{RSS}\ (w, b) = \sum_{i=1}^{N} (y_i - (w x_i + b))^2 + \alpha \sum_{j=1}^{p} w_j{}^2, \quad (3)$$

Where α is called a penalty term, the higher alpha refers to a simple model and more regularization. The penalty term α adjusts the parameters when the parameters take large values, then the optimization function is penalized. So, ridge regression minimizes the parameters to reduce the complexity of the model and multicollinearity.

3.3. Lasso Regression

The word LASSO stands for (Least Absolute Shrinkage and Selection Operator). Lasso regression is another form of regularization that uses the L1 regularization penalty for training [30]. L1 regularization minimizes the sum of the coefficient absolute values. RSS for lasso regression can be expressed as in Eq. 4:

$$\text{RSS}\ (w, b) = \sum_{i=1}^{N} (y_i - (w x_i + b))^2 + \alpha \sum_{j=1}^{p} |w|, \quad (4)$$

Where α is called a penalty term that controls the amount of L1 regularization. When the value of α increases, the value of the bias will increase and the value of variance will decrease. L1 regularization penalty affects some coefficients to be zero, this is called a sparse solution (feature selection) [31], hence, L1 regularization performs feature selection. When the value of α increases, some of the coefficient values will be zero. So, lasso can give good results when there are few coefficients.

3.4. Data Scaling

The method of data scaling is one of the most significant steps in machine learning during the process of preprocessing. This method is very effective in the process of normalizing the variables of the data [32]. In this paper normalize method is used to perform the normalizing process on the data, it normalizes the rows to unit norm. Each row with non-zero components is rescaled individually by its norm (L1, L2, or max). L1 norm is the sum of the absolute values of the row, the L2 norm is the square root of the sum of the squared values of the row, and the max norm is the maximum values of the row.

IV. RESULTS AND ANALYSIS

The implementation of the linear regression, ridge regression, and lasso regression algorithms are done using the Google Collab notebook. Google Collab notebook helps to write and execute Python in the browser, where it is open-source and widely used for the implementation of machine learning algorithms such as regression, classification, and clustering.

4.1. Implementation Using all Features

Linear regression, ridge regression, and lasso regression machine learning algorithms are implemented in the forest fires data set that is presented in the UCI machine learning repository. Accuracy score, MAE, MDAE, and MSE were calculated for these algorithms. The accuracy score on the training data set is 1, 0.98, and 0.88 on linear regression, ridge regression, and lasso regression, respectively. The accuracy score on the testing data set is 1, 0.95, and 0.81 on linear regression, ridge regression, and lasso regression, respectively. MAE, MDAE and MSE on linear regression are 2.25e-16, 2.22e-16 and 6.46e-32, on ridge regression are 0.0044, 0.0027, and 4.58e-05 and on lasso regression are 0.0089, 0.0051 and 0.0002, respectively.

Table 1. Accuracy score, MAE, MDAE, and MSE on linear regression using all features.

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
1	1	2.25e-16	2.22e-16	6.46e-32

Table 2. Accuracy score, MAE, MDAE, and MSE on ridge regression using all features

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
0.98	0.95	0.0044	0.0027	4.58e-05

Table 3. Accuracy score, MAE, MDAE, and MSE on lasso regression using all features

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
0.88	0.81	0.0089	0.0051	0.0002

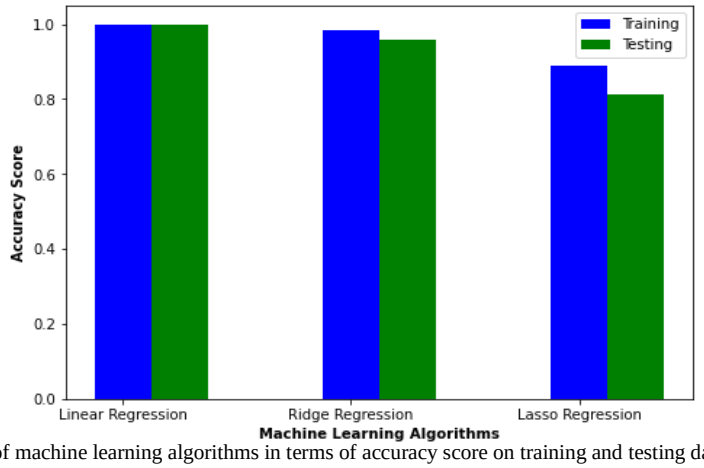

Figure3. Comparison of machine learning algorithms in terms of accuracy score on training and testing data set using all features

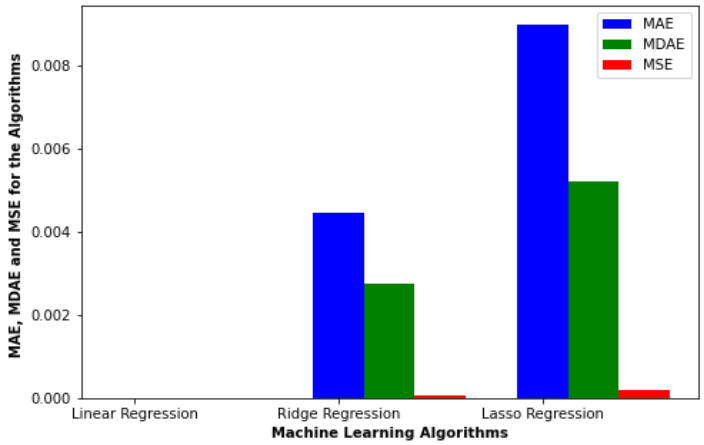

Figure 4. Comparison of machine learning algorithms in terms of MAE, MDAE and MSE using all features

So, from these results, linear regression gives better accuracy. Tables 1, 2, and 3 show these results. Fig 3 demonstrates a comparison of these algorithms in terms of accuracy score on training and testing data set using all features. Fig 4 demonstrates a comparison of these algorithms in terms of MAE, MDAE, and MSE uses all features.

4.2. Implementation Using 70% of the Features

The accuracy score on the training data set is 0.99, 0.76, and 0.84 on linear regression, ridge regression, and lasso regression, respectively. The accuracy score on the testing data set is 0.99, 0.79, and 0.87 on linear regression, ridge regression, and lasso regression, respectively. MAE, MDAE and MSE on linear regression are 0.0023, 0.0014 and 1.30e-05, on ridge regression are 0.0093, 0.0056 and 0.00037 and on lasso regression are 0.0083, 0.0050 and 0.00022, respectively.

Table 4. Accuracy score, MAE, MDAE, and MSE on linear regression 70% of the features

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
0.99	0.99	0.0023	0.0014	1.30e-05

Table 5. Accuracy score, MAE, MDAE, and MSE on ridge regression using 70% of the features

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
0.76	0.79	0.0093	0.0056	0.00037

Table 6. Accuracy score, MAE, MDAE, and MSE on lasso regression using 70% of the features

Accuracy scores on the training data set	Accuracy score on the testing data set	MAE	MDAE	MSE
0.84	0.87	0.0083	0.0050	0.00022

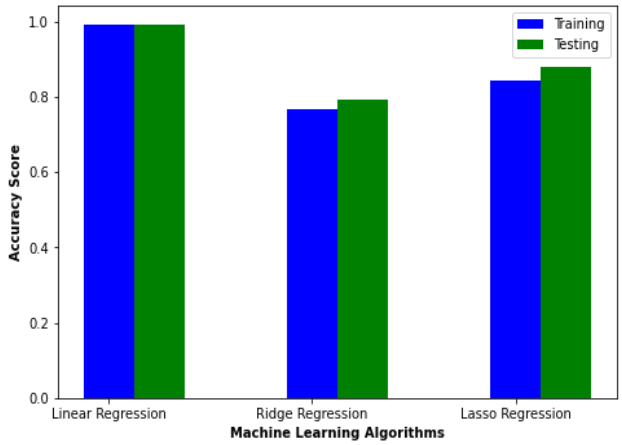

Figure 5. Comparison of machine learning algorithms in terms of accuracy score on training and testing data set using 70% of the features

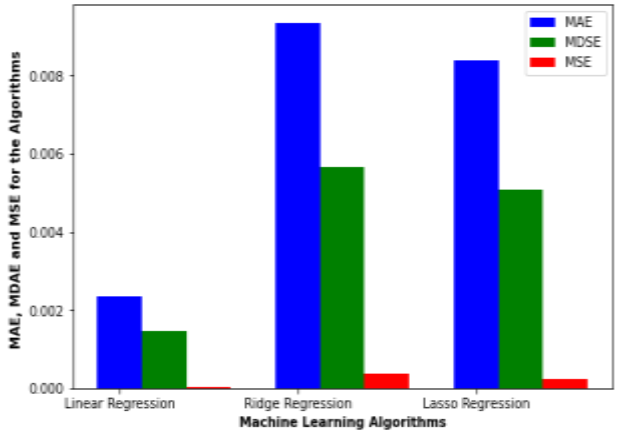

Fig 6. Comparison of machine learning algorithms in terms of MAE, MDAE and MSE using 70% of the features

So, from these results, linear regression gives better accuracy. Tables 4, 5, and 6 show these results. Figure 3 demonstrates a comparison of these algorithms in terms of accuracy score on training and testing data set uses 70% of the features. Fig 4 demonstrates a comparison of these algorithms in terms of MAE, MDAE, and MSE uses 70% of the features.

V. CONCLUSION

In this research, the main idea is to perform three machine learning algorithms to predict forest fires. The data set is presented in the UCI machine learning repository. The size of the data set is 517 instances, and some attributes are 13. Linear regression, ridge regression, and lasso regression machine learning algorithms are

implemented to perform the prediction process. The three algorithms are applied using two scenarios. In the first scenario, all attributes of the data set were included and in the second scenario, 70% of the attributes were included. The training set is 70% of the data set and the test set is 30% of the data set in the two scenarios. Accuracy score was calculated on both training and testing data set, in the training data set it was 1, 0.98 and 0.88 on linear regression, ridge regression, and lasso regression, respectively, in testing data set it was 1, 0.95 and 0.81 on linear regression, ridge regression, and lasso regression, respectively. The experimental results demonstrated that the linear regression algorithm presented the best result.

VI. REFERENCES

1- László F, Rajmund K, "Characteristics of forest fires and their impact on the environment", Academic and Applied Research in Military and Public Management Science, vol.15, pp.5-17, 2016.
2- Barker T, "The economics of avoiding dangerous climate change. An editorial essay on The Stern Review", Climatic Change, vol. 89, pp. 173-194, 2008.
3- Mote T, Singh A, Prasad M, Kalwar P, "Predicting burned areas of forest fires: an artificial intelligence approach", International Journal of Technical Research and Applications, vol. 43, pp. 56-58, 2017.
4- Lin Z, Liu HH, Wotton M, "Kalman filter-based large-scale wildfire monitoring with a system of UAVs", IEEE Transactions on Industrial Electronics, vol. 66, pp. 606-615, 2018.
5- Agarwal S, "Data mining: data mining concepts and techniques", In International Conference on Machine Intelligence and Research Advancement, IEEE, 2013.
6- Cortez P, Morais AD, "A data mining approach to predict forest fires using meteorological data, Environmental Science, 2007.
7- Özbayoğlu AM, Bozer R, "Estimation of the burned area in forest fires using computational intelligence techniques", Procedia Computer Science, vol.12, pp. 282-287, 2012.
8- Salis M, Arca B, Alcasena F, Arianoutsou M, Bacciu V, Duce P, Duguay B, Koutsias N, Mallinis G, Mitsopoulos I, Moreno JM, "Predicting wildfire spread and behaviour in Mediterranean landscapes", International Journal of Wildland Fire, vol. 25, pp. 1015-1032, 2016.
9- Castelli M, Vanneschi L, Popovič A, "Predicting burned areas of forest fires: an artificial intelligence approach", Fire Ecology, vol. 11, pp. 106-118, 2015.
10- Radke D, Hessler A, Ellsworth D, "Forecast: leveraging deep learning to predict wildfire spread", In Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019.
11- Zhu H, Gao D, Zhang S, "A perceptron algorithm for forest fire prediction based on wireless sensor networks", JIOT, vol. 1, pp. 25-31, 2019.
12- Deng L, Perkowski M, Saltenberger J "A novel forest fire prediction tool utilizing fire weather and machine learning methods", In Proceedings for the 5th International Fire Behavior and Fuels Conference, International Association of Wildland Fire, 2016.
13- Sakr GE, Elhajj IH, Mitri G, Wejinya UC, "Artificial intelligence for forest fire prediction", In International Conference on Advanced Intelligent Mechatronics, IEEE, 2010.
14- Pérez-Sánchez J, Jimeno-Sáez P, Senent-Aparicio J, Díaz-Palmero JM, Cabezas-Cerezo JD, "Evolution of burned area in forest fires under climate change conditions in southern Spain using ANN", Applied Sciences, vol. 9, pp. 4155, 2019.
15- Dacre HF, Crawford BR, Charlton-Perez AJ, Lopez-Saldana G, Griffiths GH, Veloso JV, "Chilean wildfires: probabilistic prediction, emergency response, and public communication", Bulletin of the American Meteorological Society, vol. 99, pp. 2259-2274, 2018.
16- Coffield SR, Graff CA, Chen Y, Smyth P, Foufoula-Georgiou E, Randerson JT, "Machine learning to predict final fire size at the time of ignition", International Journal of Wildland Fire, vol. 28, pp. 861–873, 2019.
17- Stojanova D, Panov P, Kobler A, Džeroski S, Taškova K, "Learning to predict forest fires with different data mining techniques", In Conference on Data Mining and Data Warehouses, SiKDD, 2006.
18- Boubeta M, Lombardía MJ, González-Manteiga W, Marey-Pérez MF, "Burned area prediction with semiparametric models", International Journal of Wildland Fire, vol. 25, pp. 669-678, 2016.
19- Kansal A, Singh Y, Kumar N, Mohindru V, "Detection of forest fires using machine learning technique: A perspective", In Third International Conference on Image Information Processing (ICIIP), IEEE, 2015.
20- Al_Janabi S, Al_Shourbaji I, Salman MA, "Assessing the suitability of soft computing approaches for forest fires prediction", Applied Computing and Informatics, vol. 14, pp. 214-224, 2018.
21- Garrard P, Rentoumi V, Gesierich B, Miller B, Gorno-Tempini ML, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse", Cortex, vol. 55, pp. 122-129, 2014.
22- Aleksandar P, Silvana P, Valentina ZP, "Multiple linear regression model for predicting bidding price", Technics Technologies Education Management (TTEM), 2015.
23- James G, Witten D, Hastie T, Tibshirani R, "An introduction to statistical learning", New York: Springer, vol. 112, pp. 3-7, 2013.
24- Pereira JM, Basto M, da Silva AF, "The logistic lasso and ridge regression in predicting corporate failure", Procedia Economics and Finance, vol. 39, pp. 634-641, 2016.
25- Singh BK, Verma K, Thoke AS, "Investigations on the impact of feature normalization techniques on classifier's performance in breast tumour classification", International Journal of Computer Applications, vol. 116, pp. 11-15, 2015.