Project Report

Prototype AI Project using Python and Anaconda

Burned Area of forests prediction using AI.

November 22, 2020

Contents

List of Figures

1 Abstract

A data set is selected for the analysis of AI techniques for solving real world problems. Data set is analyzed first and then research is carried out for techniques identification. Prediction results for different techniques are compared on the basis of common characteristics. The test data is cross validated to avoid overfitting of the model. At the end the result are discussed and techniques for deployment on real world data is discussed.

2 Introduction

Forest fire has been the major issue in countries like Australia and USA. It had create a loss of billion dollars to the economy and ruined the habitat of a lots of animal species. Because of this many of the species of wild animals become endangered. Detection of forest fire has been used for decades to overcome this issues. For this purpose the people use the following conventionnal methods.

- automatic tools
- local sensors
- weather stations
- meterological conditions
- satellite based infrared smoke sensors

Uptill now these techniques didn't work well and are not efficient to solve this problems. As once the fire started then it becomes difficult to control.
Due to recent advances in Data mining (DM) and Data Science it has become possible to solve this problems from the data collected from different scenarios. The novel Data mining approaches can be used to solve the problems by emphasing the use of real time data and non costly meteorological data.

Forestfire data set is used to train the AI models for predicting forest fire. The data set is imported in colab (an open source Google Cloud Server) and data set visualization tools provided by anaconda (an open source platform for AI) are used for analysis. A preprocessing technique is used to encode the data set for training provided by Scikit learn. By using cross validation on data set the data set is splited into training and testing data set. Four AI techniques were used which are Support Vector Regressor (SVR), Deep Neural Network (DNN), Decision Tree Regressor (DTR) and Random Forest Regressor (RFR). At the end the performance measures of all the techniques are compared.

3 Background

In the past various techniques were applied on the available meterological data has been used to solve the issues related to environmental conditions. The data has corporated to numerical indices which was used for prevention and fire management purposes. A Canadian based agency named Canadian Forest Fire Weather Index (FWI) developed a data base for forest fire data considering various meterological situations for tackling the issue. They developed the different indexes for predictiono of weather conditions. All the indices are mentioned in the data set description.
A method based on spatial clustering was adopted by Hsu et Hal to detect forest fire spots in satellite images. These images were feed to SVM for regression and 75% accuracy was observed.
Upon analysis it was cleared that it is a clear regression task and can be handled from the techniques available. A common Regressor can be stood best fit for the solution of this dataset.

4 AI Techniques

4.1 Literature Survey

Upon studying the literature the following approaches are available in the literature to solve the problems of regression. Although there are numerous others but they are suitable to our problem for the reasons which are discussed under this section. [1]

1. Support Vector Regression (SVR)
2. Deep Neural Network (DNN)
3. Decision Tree Regression (DTR)
4. Random Forest

4.2 Support Vector Regression

Support vector regression works similar to suppor vector machine SVM which is a binary classifier and mostly used in the supervised machine learning tasks. The working principle of SVR is little bit different from SVM as it is a regression algorithm thus it can be used with continuous data. In SVR the error differnce is tried to fit under the minimum threshold. The working principle of support vector regressor is as follow:

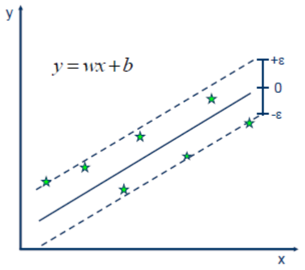


Figure 1: Working principle of support vector regressor.

4.3 Deep Neural Network

Deep Learning is one of the advanced technique of Machine learning and is also very powerful. It can be applied to any of the raw form of data and provide a significant results. In our study numerous regression problems have been solved using deep learning. The key idea behind the deep learning working is the modeling of human brain. As human also learn by using the visual data by tuning its parameters.

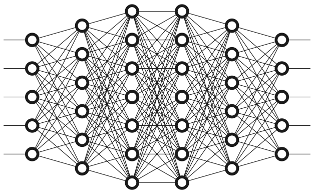


Figure 2: Working principle of deep learning.

4.4 Decision Tree

Decision tree as from the name means that the using tree like structure to make decisions about the problems. It is one of powerful tool which is used in cluster type data to make the prediction. It has numerous applications in regression problems and particularly in the cases where data is formatted in the form of cluster and single decision is affected by the numerous factors.

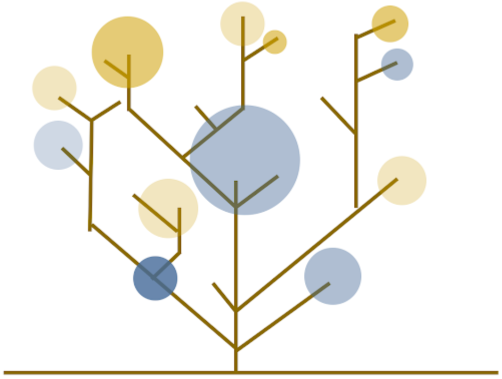


Figure 3: Working principle of decision tree.

4.5 Random Forest

Random forest is an ensemble learning method for classification and regression. It works by construction of multilevel decision tress during the training session. It works by prediction the class of decision tree for detection of label of test data set. Basically a random forest is a cluster of decision trees which tries to hold the final decision by spiliting the classification task into the different branches.

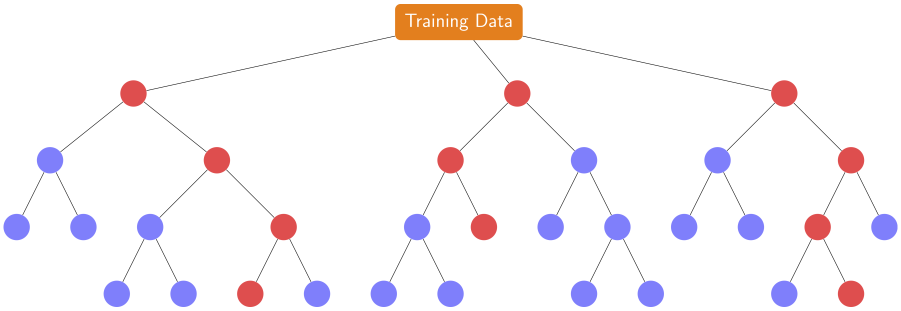


Figure 4: Working principle of Random Forest.

5 Software tools

The following software tools were used for solving the dataset: [2]

1. Python Libraries
   - Numpy
   - Pandas

- Scikit-learn
- Keras
- Tensorflow

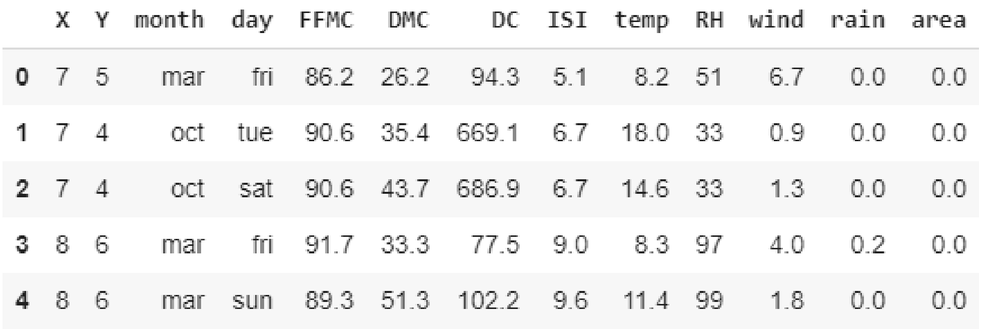
2. Google Colab

3. Python (Scripting language for scientific computation)

4. Anaconda (a package manager for supporting Machine learning libraries)

6 Data Set Descriptive Analysis

The dataset has 13 attributes which fully describes its different characteristics. There is no need of using any data preprocessing techniques. The data set has no missing values and ambiguities. A simple glimpse of data is shown below which is visualized by using pandas.



	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

Figure 5: Visualization of dataset.

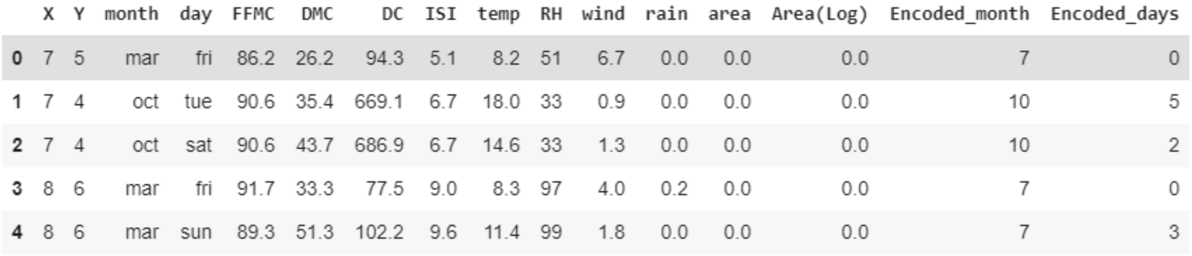
6.1 Features Identification of Dataset

The data set has the following features: [3]

- **(X,Y)** defines the spatial coordinates with reference to Montesinho park.
- **month** months of the year
- **day** week days
- **FFMC** Fine Fuel Moisture Code range (18.7 - 96.20).
- **DMC** Duff Moisture Code range (1.1 - 291.3).
- **DC** Drought Code range (7.9 - 860.6).
- **ISI** Initial Spread Index range (0.0 - 56.0).
- **temp** Temperature range (2.2 - 33.30) degree celcius.
- **RH** Relative Humidity range (15.0 - 100) %.
- **wind** wind speed range (0.40 - 9.40) $\frac{km}{h}$.
- **rain** amount of rain range (0.0 - 6.4) $\frac{mm}{m2}$.
- **area** Burned area due to fire range (0.0 - 1090.84) (ha).

6.2 Dataset Preprocessing and feature extraction

In the data set we have the names of months and days thus they cannot be feed directly to the Machine learning models. As most of the models like Neural Networks and SVR used the data in the discrete form. Thus these data features are encoded into numerical forms before feeding into the model. These features are encoded using the built in functionality provided by the Scikit library. After encodation the data set looks like this.



	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area	Area(Log)	Encoded_month	Encoded_days
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0	0.0	7	0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0	0.0	10	5
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0	0.0	10	2
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0	0.0	7	0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0	0.0	7	3

Figure 6: Encoded data set after preprocessing.

7 Problem Nature Identification

From the data set it is cleared that it is a **complex regression task**. Based upon the different features we have to predict the area burned due to forest fire. As we have number of independent variables and only one dependent variable which is area to be predicted in this case. Thus we have to use those machine learning tools which are efficiently predict the output using regression. In the next section different AI techniques which are suitable to solve this problem and available in literature are studied and discussed.

8 Evaluation Method

The evaluation procedure for this particular problem is REC (Regression Error Characteristics Curve). As the task is related to regression thus the different regression function can be evaluated depending on the REC curve of different regression functions.

An AI approach is analysed by numerous factors but as far as our problem is concerned we can use the following evaluation parameters to check the validity of our results.

1. Accuracy
2. Small Prediction Error
3. Favorable for Real time application
4. Computational intensity

8.1 Support Vector Regression

As support vector regression is a linear regression. It has the following powerful features that motivates us to use this in solving our problem.

8.1.1 Pros

- It is easy and straightforward to explain and understand.
- In order to avoid overfiting regularization is possible.

8.1.2 Cons

- Under non linear relations it behave poorly
- It is unable to capture complex pattern

8.2 Deep Learning Approach

It is one of the powerful tool to learn complex patterns yet it has many pros and cons

8.2.1 Pros

- Its backpropagation algorithm makes it powerful enough to analyse the pattern in data
- It alleviate the burden of data preprocessing

8.2.2 Cons

- It requires a large amount of data
- Requires expert knowledge to train and tune its hyperparameter.

8.3 Decision Tree

It works by divide and conquer based strategy to get the most of the benefit. It has the following pros and cons.

8.3.1 Pros

- It has the ability to learn non linear relation
- Perform well enough in practice

8.3.2 Cons

- It may subjected to overfitting
- It takes time to form pattern for learning

8.4 Random Forest

Random forest works by ensembling different decisions tree for the problems. Thus it works by making classes and subclases inside the data set.

8.4.1 Pros

- It makes the use from the power of decision trees by clustering the data in classes.
- It result are superior to decision trees.

8.4.2 Cons

- Training time is comparatively higher.
- May subject to overfitting if data is not well subjected to cross validation.

9 Results

As we have the complex regression task so generally the results of regression task are analyzed using Regression Error Characteristics (REC).

9.1 REC

In REC a cumulative distribution function of the error is generated by plotting the error tolerance level on horizontal axis and predicted data points on the vertical axis.
REC for three of the models are shown below:

9.2 SVR

The REC curve for SVR is as follow:

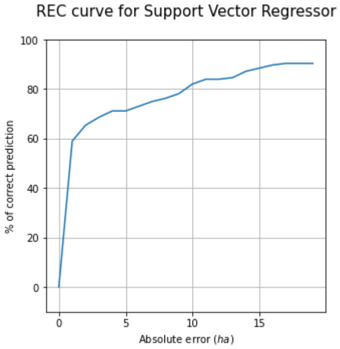


Figure 7: REC curve for SVR.

9.3 Decision Tree

The REC curve for Decision Tree is as follow:

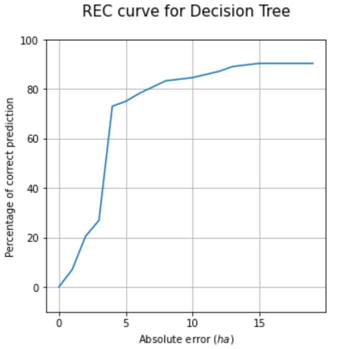


Figure 8: REC Curve for decision tree.

9.4 Deep Neural Network

The REC curve for Deep Neural Network is as follow:

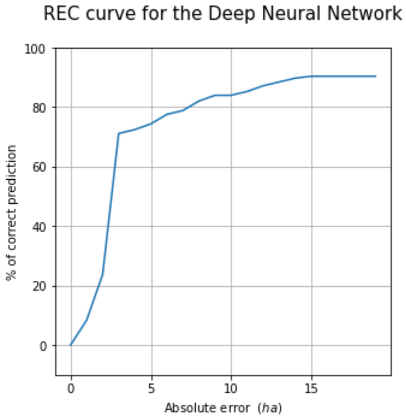


Figure 9: REC curve for deep neural network.

9.5 Random Forest

The REC curve for Random Forest is as follow:

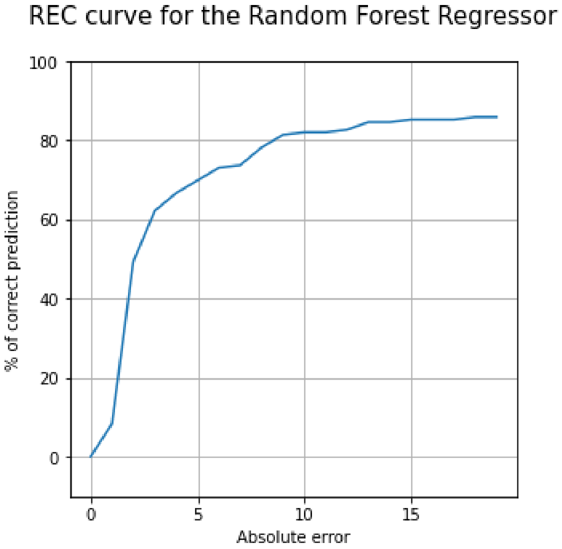


Figure 10: REC curve for random forest.

9.6 Performance Comparison

The overall performance comparison is shown below for the models

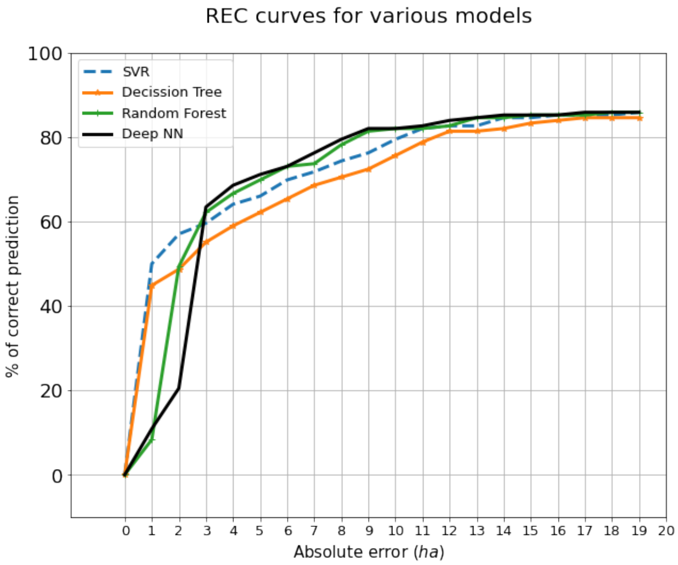


Figure 11: Performance comparison for all the models.

It can be seen from the above curves that the SVM tends to produce the best prediction. It also shows the non relevance of spatial and temporal variables on SVM. Thus it is better to use weather scenarios for SVM. Thus SVM is the best for all of the regression model from this result.
The overall results shows that all the classifier best fits for the performance accuracy of around 85 percent which is considered as good. The performance can be improved by using further techniques of data preprocessing which are as follows:

1. Principal Component Analysis (PCA).

2. Dimensionality Reduction

10 Conclusion and Future Directions

SVR provides best of its accuracy for predicting small fires which constitutes a large portion and one of the major concerns. Its accuracy is relatively low for large fires. Thus it can be deployed for normal use for fire fighters and good for real time detection. DNN takes time and require huge computation for prediction thus it takes a strong machine which must be GPU enabled for deploying. Decision Tree also takes time larger than the SVR but it is suitable as it takes less computation. Random forest and DNN curves are closly related to each other. Their behavior is quite similar.

The system can be developed for practical applications and can be depolyed on any real time data available. The regressor used in the solving the dataset can be best suitable if data is preprocessed and feed to the regressors in perfect conditions. Various other technological aspects like real time weather monitoring and weather forcast can be used to get the best of the results from this system.

References

[1] towardsdatascience, “Regression,” https://towardsdatascience.com/, Nov 2017, accessed on 2020-05-18.

[2] anaconda, “keras and tensorflow,” https://www.anaconda.com/, Sep 2018, accessed on 2020-05-19.

[3] P. Cortez and A. d. J. R. Morais, “A data mining approach to predict forest fires using meteorological data,” 2007.