ASSIGNMENT 4 SIT720

Abstract : This report aims to predict the burned area caused due to forest fires in the northeast region of Portugal. We will be producing the same RMSE results based on the research paper by Paulo Cortez and Anibal Morais . We consider five Machine Learning models – Support Vector Machines (SVM), Random Forest, Decision Tree, Multiple Regression and Neural Network in this paper, along with four feature selection setups via spatial , temporary, weather features and FWI components.

Explanation of Forest Fire Data –

According to the Fire Weather Index (FWI), there are six components for rating fire danger which are :

Fine Fuel Moisture Code (FFMC) -
Duff Moisture Code (DMC)
Drought Code (DC)
Initial Spread Index (ISI)
Buildup Index(BUI)
FWI

The FWI index indicates the fire intensity and combines the ISI and the BUI indices. High values of every code and index suggest that there are more severe burning conditions.

In the table below, there are pre-processed selected variables with the description. The authors selected four spatial and temporal variables. X and Y denotes the geographic location of where the fires occurred because of low vegetation quality. There is a distinct **month** based weather conditions, and **day** of the week might influence forest fires. The four FWI variables are : FFMC code, DMC code, DC code and ISI index. The four weather *conditions* are : temperature, RH (relative humidity), wind and rain. Additionally, the area represents the total burned area (in *ha).*

Since the burned area distplot (Figure 1) illustrates a highly positive skew with most fires covering a smaller area, in order to reduce the skewness and improve the symmetry of the distribution, we use log transformation of the area variable (Figure 2).

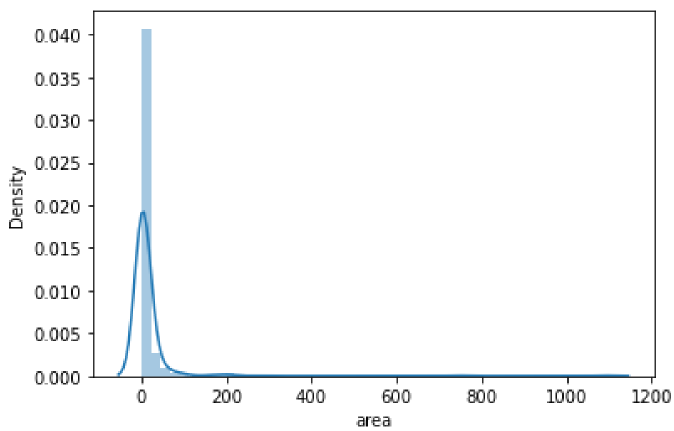

Figure 1

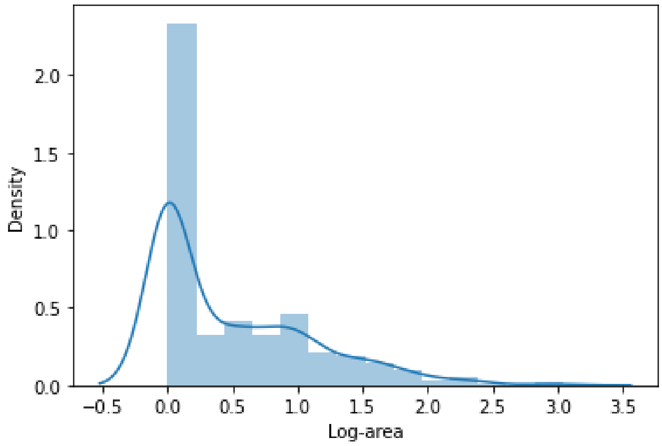

Figure 2

We also display boxplots of Log area grouped by day and month. We can observe that there are a few outliers in the second boxplot.

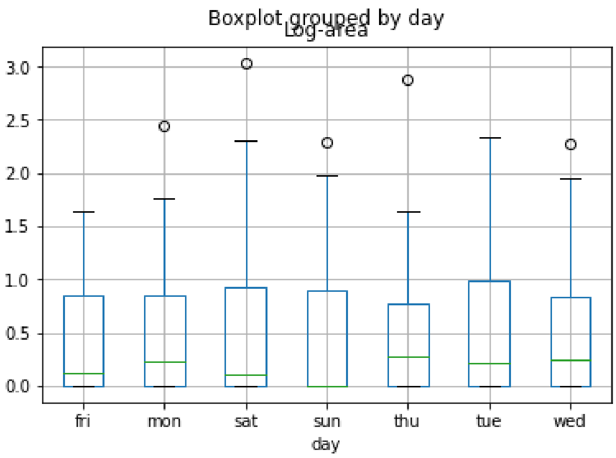

Figure 3

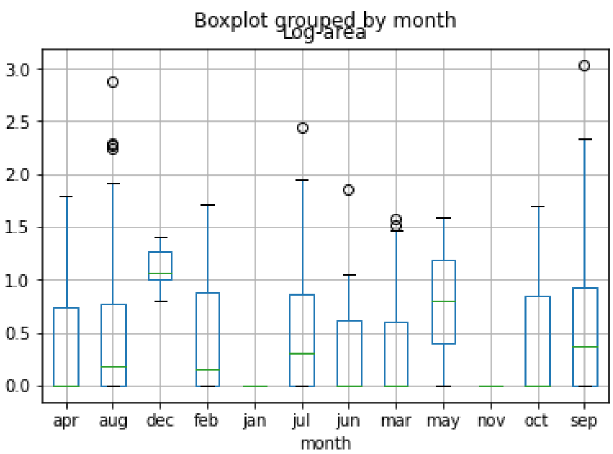

Figure 4

Machine Learning Models -

Multiple ML models have been implemented to achieve the task of predicting the burned area.

1) **Multiple Linear Regression** – This is the most standard and classical approach; however, it can only learn linear functions. It uses the Ordinary Least Squares method for prediction. It explains the relationship between one continuous DV (area in hectares) and multiple Independent variables. However, it learns on linear mapping, hence, to address this drawback, we use alternative machine learning models.
2) **Decision Tree** - These are non-parametric supervised learning method, wherein a tree is a piecewise constant approximation. The DT predictor learns from the data to approximate a sine curve with a set of IF-THEN decision rules, implying that the more the depth of the tree, the fitter the model is. Additionally, the cost of using the tree is logarithmic to train the tree via the data points.
3) **Random Forest Regressor** – This ensemble model is a meta-estimator which fits multiple DTs on various sub-samples (using random feature selection via bootstrap training examples of the dataset) and averages the result to improve the predictive accuracy score and control over-fitting. This is a significant improvement from the single Decision Tree .
4) **Neural Network** – It is a series of procedures that endeavours to recognize the underlying relationships between the data attributes via a process that imitates the way a human brain operates. NN basically refers to a system of neurons, both organic or artificial. Multilayer perceptron (MLP) is used here, which is a fully connected class of feedforward artificial neural network (ANN) . The process will consider a hidden layer *H*, and nonlinear activation function and an output layer. It utilizes a backpropagation technique for training. Thus, having the lowest penalised error, the NN will depend on the hidden nodes *H*.
5) **Support Vector Machine** – SVMs are generally used in high-dimensional spaces using non-linear mapping and uses a subset of training points in the decision function (known as support vectors), having memory efficiency. SVM then evaluates the best linear hyperplane in the feature space. The kernel used here is Radial Basis Function kernel, which can be mathematically represented as :

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

SVM implements the Sequential Minimal Optimization (SMO) (A Library for Support Vector Machines, 2022) which works with a Lagrange multiplier α_1 and a second multiplier α_2 to optimise and repeat until convergence point. It has two hyperparameters associated with it : 'C' for SVM and 'γ' for the RBF kernel. So, finding the right γ is crucial in order to achieve the best Bias-Variance trade off

Experiment Protocol

Data Pre-processing and Steps taken

The nominal variables: **month** and **day**, were first transformed to numerical attributes using OneHotEncoding and LabelEncoder() function for SVM, NN and MR.

All the attributes were also standardised using a Standardscaler() function in Python, with a mean equal to zero and one standard deviation.

The test size used in this process is 0.4. We use the train_test_split function to split the dataset into training and test dataset for further evaluation, ratio of 60:40 samples.

Furthermore, for REC estimation, we define a function rec, presenting the absolute deviation against the percentage of accurate prediction, i.e., burned area. The maximum tolerance limit for the REC curve is set to be 20.

A GridsearchCV is employed to search for the best hyperparameters for SVM and NN (10-fold grid search) on the training dataset.

We then fit the regressor models to the training dataset. The parameters used for each model is presented in the table below, for SVM, RF and NN produced with the parameter grid search cross-validation approach.

Support Vector Regressor	**Decision Tree Regressor**	**Random Forest Regressor**	**Neural Network**	**Multiple Linear Regressor**
C : 3	max_depth : 10	max_depth : 50	epochs : 100 *	Normalize= True
Gamma : as provided in the screenshot below	criterion : squared error	max_leaf_nodes : 2	batch_size =10 *	OLS method
kernel : rbf	min_samples_split:4	min_samples_leaf : 10	verbose = 0 *	
	min_samples_leaf:2	min_samples_split : 2	Learning_rate : constant	
	min_weight_fraction_leaf : 0.01	n_estimators (T) : 500	Hidden_layer_sizes :4	
		random_state = 42	Activation: logistic	
			learning_rate: 0.001	
			solver : lbfgs	
			max_iter : 200	

```
Best parameters obtained via Grid Search: {'C': 3, 'gamma': 0.5, 'kernel': 'rbf'}
Best parameters obtained via Grid Search: {'C': 3, 'gamma': 0.00195, 'kernel': 'rbf'}
Best parameters obtained via Grid Search: {'C': 3, 'gamma': 0.03703, 'kernel': 'rbf'}
Best parameters obtained via Grid Search: {'C': 3, 'gamma': 0.00195, 'kernel': 'rbf'}
Best parameters obtained via Grid Search: {'C': 3, 'gamma': 0.03703, 'kernel': 'rbf'}
```

Figure 5

The best hyperparameters for NN and SVM (median values)

	Feature Selection Setup			
DM Model	STFWI	STM	FWI	M
NN	4	6	4	4
SVM	2^{-5}	2^{-3}	2^{-3}	2^{-3}

Figure 6

The best SVM hyperparameters were selected based on a 10-fold GridsearchCV, where cv=10 used on the training data.

After fitting the ML models, the RMSE scores were postprocessed using the inverse of the logarithmic transformation. In some cases, there was a negative transformation which was set to zero, similar to the research paper.

Feature selection -

The feature selection setup for these processes is : **STFWI** : which uses spatial, temporal and the four FWI components (FFMC, DMC, DC, ISI) ; **STM** : which uses spatial, temporal and four weather attributes (rain, temp, wind, RH) ; **FWI** : uses only the four FWI components; and **M** – four weather variables.
Each of these distinct feature selection setups were modelled for the ML algorithms.

Results with variation – RMSE results

We use the RMSE metric here which is represented by this formula :

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 / N}$$

A lower value of RMSE is always considered better in predictive models, although, RMSE is highly sensitive to high errors. We also use the Regression Error Characteristic (REC) curve as a metric, plotting the absolute deviation (x-axis) against the percentage of accurate predictions (y-axis). Thus, the ideal regressor model should give the output of REC area nearer to 1.0.

There is slight variation in the RMSE results mainly due to the different train_test_split ,test size of 60:40 ratio. Although, the random state equals 42 to avoid any randomness of the split, variation arises due to highly complex models with multiple features here.

NB:

I tried installing TensorFlow package in this environment, however, when I run the below code it comes up as a dead kernel. I therefore, uninstalled and reinstalled my Anaconda Navigator and Python version - 3.9.1.

After installing the new version, I created a new environment named tensorflow to install the tensorflow packages and the keras package, however, when I open the Jupyter notebook(tensorflow), I am able to install keras and tensorflow package but not other various packages such as matplotlib.pyplot or sklearn.model_selection.

Additionally, I tried running the below code in Google collab which works fine since it is on Cloud, however, I am not able to upload my dataset and run the neural network on the dataset.

From my experience, downloading tensorflow package, my kernel was not supporting it, thus, I couldn't get the required RMSE result for Neural Network. Although, I have still provided the REC curve for Deep Neural Network to compare against other models.

If possible, when you download this .ipynb file, if you have tensorflow package in your anaconda environment, you can see the RMSE results.

However, in addition to that, I have solved the model using the MLP regressor for comparison purposes and produced the results.

Models	Multiple Linear Regressor	Support Vector Regressor	Decision Tree Regressor	Random Forest Regressor	Neural Network
RMSE(Naïve)	0.622	0.809	1.213	0.798	**0.714**
RMSE(STFWI)	0.626	1.316	1.128	0.628	1.040
RMSE(STM)	0.620	0.624	0.771	0.664	0.815
RMSE(FWI)	0.615	1.363	**0.671**	**0.617**	1.203
RMSE(M)	**0.614**	**0.619**	0.771	0.636	1.350

Thus, Naïve average predictor is set as a benchmark for model comparison. We can observe that, comparing the RMSE results with the research paper, we can observe similar to the research paper, that SVM has the best configuration for M feature set, however, in this scenario, MR model performs equally great, although it could be because there is variation in the train and test split. However, these values are very close to the errors in the report, **with admissible errors of ranging between 2.85 – 3.50.**
In effect, the feature set containing four weather conditions is far better to use for the SVM model, likewise with MLR model. It is an intriguing outcome, as the weather-related attributes can be obtained directly from the weather sensors, with no requirement for aggregated calculations.

REC Curve :

The REC curve below represents the percentage of accurate predictions against the absolute error (tolerance) in prediction ($).

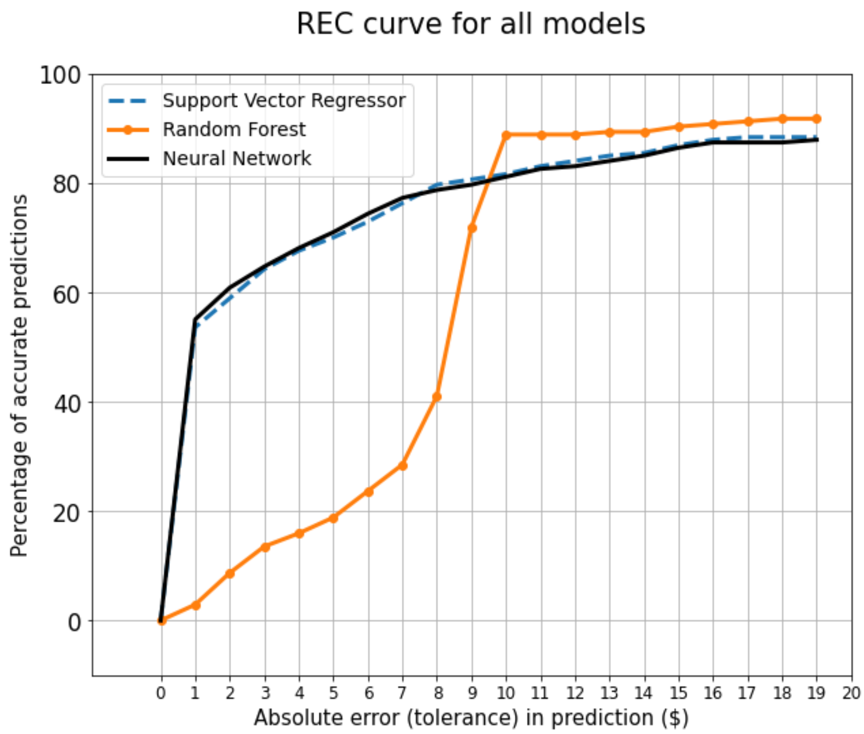

Figure 7

Q-2

a) In this section, I have implemented L2 Regularised Linear Regression using the same feature selection set-up and received a figuratively lower RMSE score. The motivation behind this is because in ridge regression, the cost function is changed by adding a penalty which is equivalent to square of the magnitude of the coefficients. It can be mathematically presented as:

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

Cost function for ridge regression

Therefore, ridge regression constraints the coefficients (*w*). The penalty term (aka lambda) regularizes the coefficients in a such way that if the coefficients end up taking large values, the optimisation function is then penalized. Hence, ridge regression shrinks the coefficients and aims to reduce the model complexity and its relevant multi-collinearity between the attributes (1.1. Linear Models, 2022)

b) The proposed ML solution of ridge regression is different than other models since it aids in overcoming over-fitting problem in the models, by adding a Penalty term. It is termed as regularization as it assists in keeping the parameters regular. The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage , meaning, the larger the value of α, the larger the amount of shrinkage and therefore, the coefficients become more robust to collinearity.

c) The first steps to conduct the above ML model is to import Ridge from sklear.linear_model. We set the Ridge function with an alpha of 0.003 and normalise = True. These are the hyperparameters. The feature selected here are the four feature selection setups. And our target variable to predict is area, in this study.

d) **Experiment protocol** - We then fit our model to the training dataset. Following that, we predict on our unseen test dataset to produce the output of predicted values. The predicted results are flattened to an array to convert dimensionality. The RMSE is then calculated by taking the squareroot of the squared difference between the actual and the predicted values.

e) **Evaluation Metrics –** The metric score being evaluated here is RMSE score which is

f) **Visualisation of tables and graphs -** The graph below shows the Actual Area burned against the error in a scatter plot. The second graph shows the histogram for the prediction errors against the prediction error. And lastly, the REC curve for Ridge Regression, there is a maximum area under the curve, hence, Ridge regression performs extremely better than Multiple Linear Regression as well.

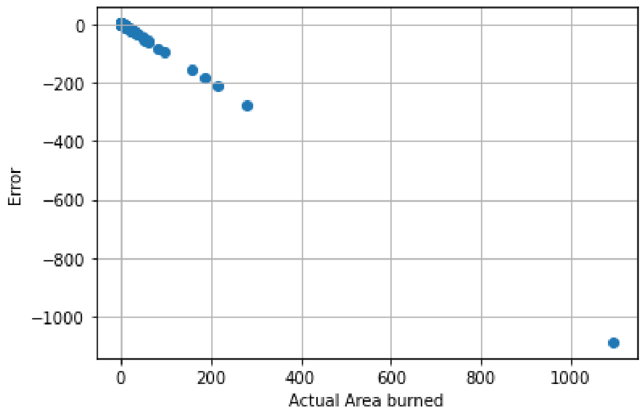

Figure 8

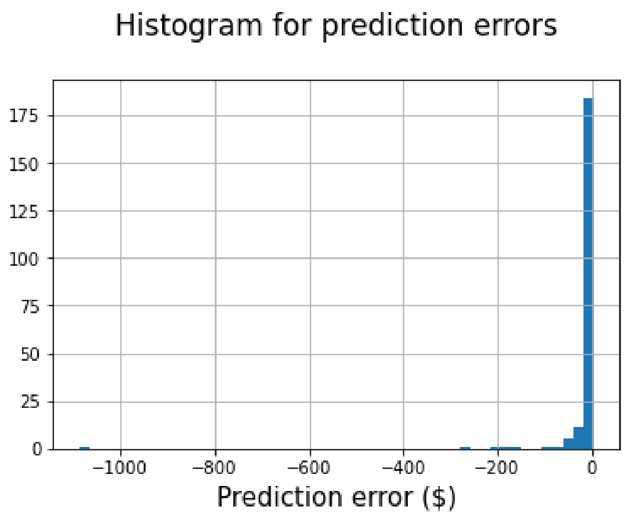

Figure 9

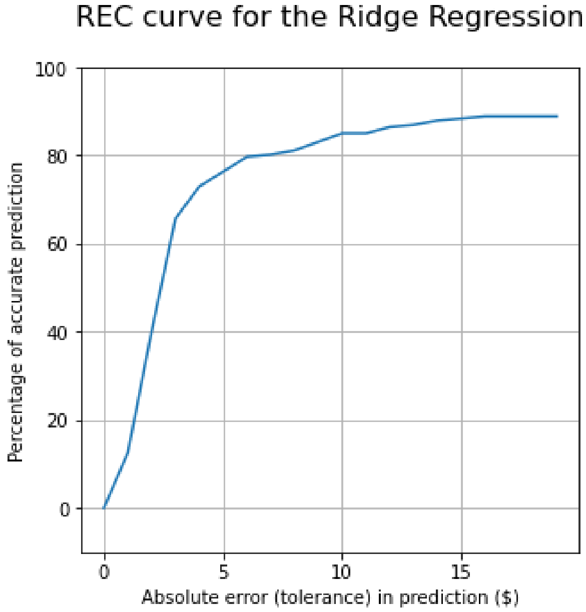

Figure 10

g) We can clearly observe that Ridge regression performs better than Multiple Linear Regression when the feature selection is **STFWI and STM.** It provides the same output when there are only weather variables or FWI variables of 0.615.

	Multiple Linear Regression	**Ridge Regression**
RMSE(Naïve)	0.622	0.621
RMSE(STFWI)	0.626	0.618
RMSE(STM)	0.620	0.617
RMSE(FWI)	0.615	0.615
RMSE(M)	0.614	0.615

REFERENCES –

(IJCAI), Montreal, Quebec, Canada, August 1995.
2004. *AGU Fall Meeting Abstracts*, pages B853+, December 2005.
679, May 2000.

1. A.Flexer.Statisticalevaluationofneuralnetworksexperiments:Minimumrequirementsand
2. A.SmolaandB.Scholkopf.Atutorialonsupportvectorregression.TechnicalReportNC2-
1. B. Arrue, A. Ollero, and J. Matinez de Dios. An Intelligent System for False Alarm Reduc- tion in Infrared Forest-Fire Detection. *IEEE Intelligent Systems*, 15(3):64–73, 2000.
3. C. Hsu, C. Chang, and C. Lin. A Practical Guide to Support Vector Classification.
4. C. Vega-Garcia, B. Lee, P. Woodard, and S. Titus. Applying neural network technology to

current practice. In *Proceedings of the 13th European Meeting on Cybernetics and Systems*
5. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge,
6. D. Stojanova, P. Panov, A. Kobler, S. Dzeroski, and K. Taskova. Learning to Predict Forest
7. D. Viegas, G. Biovio, A. Ferreira, A. Nosenzo, and B. Sol. Comparative Study of various methods of fire danger evalutation in southern Europe. *International Journal of Wildland*
8. D.Mazzoni,L.Tong,D.Diner,Q.Li,andJ.Logan.UsingMISRandMODISDataForDe- edition, 1999.
9. European-Commission. Forest Fires in Europe. Technical report, Report N-4/6, 2003/2005.

Fire, 9:235–246, 1999.

Fires with Different Data Mining Techniques. In D. Mladenic and M. Grobelnik, editors, *9th*

http://www.csie.ntu.edu.tw/˜cjlin/papers/guide/ guide.pdf, July, Dep. of Comp. Science and

human-caused wildfire occurence prediction. *AI Applications*, 10(3):9–18, 1996.

10. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*

in coastal eastern Spain. *Climatic Change*, 38:345–357, 1998.

Inference, and Prediction. Springer-Verlag, NY, USA, 2001.

Information Eng., National Taiwan University, 2003.

International multiconference Information Society (IS 2006), Ljubljana, Slovenia, 2006.

2. J. Bi and K. Bennett. Regression Error Characteristic curves. In *Proceedings of 20th In- ternational Conference on Machine Learning (ICML)*, pages 43–50, Washington DC, USA, 2003.
11. J. Zhao, C. Lu, and Y. Kou. Detecting Region Outliers in Meteorological Data. In *Pro- ceedings of the 11th ACM International Symposium on Advances in Geographic Information*
12. J.TerradasJ.PinolandF.Lloret.Climatewarming,wildfirehazard,andwildfireoccurrence *Journal of Modern Physics*, 13(8):1017–1031, 2002.
13. K. Malarz, S. Kaczanowska, and K. Kulakowski. Are forest fires predictable? *International*
3. L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. *Classification and Regression Trees*.
4. L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

ligent Information Systems, 19(1):7–23, 2002.

MA, 2001.

Mining. MIT Press, 1996.

5. P. Cortez. RMiner: Data Mining with Neural Networks and Support Vector Machines using

Pharmaceuticals with Neural Networks. *IEEE Transactions on Neural Networks*, 11(3):668–

project.org, ISBN 3-900051-00-3.

14. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL: http://www.R-

R. In R. Rajesh (Ed.), *Introduction to Advanced Scientific Softwares and Toolboxes*, In Press.

15. R. Kewley, M. Embrechts, and C. Breneman. Data Strip Mining for the Virtual Design of
16. R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model
    Research, volume 2, pages 1005–1008, Vienna, Austria, 1996.
17. S. Haykin. *Neural Networks - A Compreensive Foundation*. Prentice-Hall, New Jersey, 2nd
18. S. Menard. *Applied Logistic Regression Analysis*. SAGE, 2nd edition, 2001.
19. S. Taylor and M. Alexander. Science, technology, and human factors in fire danger rating:
    Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*
    SVM Regression. *Neural Networks*, 17(1):113–126, 2004.
    Systems, pages 49–55, New Orleans, Louisiana, USA, 2003.
20. T.Hastie,R.Tibshirani,andJ.Friedman.*TheElementsofStatisticalLearning:DataMining,*
    tection and Analysis of Smoke Plume Injection Heights Over North America During Summer
    the Canadian experience. *International Journal of Wildland Fire,* 15:121–135, 2006.
    TR-1998-030, University of London, UK, 1998.
21. U.Fayyad,G.Piatetsky-Shapiro,andP.Smyth.*AdvancesinKnowledgeDiscoveryandData*
    V. Cherkassy and Y. Ma. Practical Selection of SVM Parameters and Noise Estimation for
22. W. Hsu, M. Lee, and J. Zhang. Image Mining: Trends and Developments. *Journal of Intel-*
    Wadsworth, Monterey, CA, 1984.
    with Java Implementations. Morgan Kaufmann, San Francisco, CA, 2005.
23. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
24. Marquardt, D.W. and Snee, R.D., 1975. Ridge regression in practice. *The American Statistician*, *29*(1), pp.3-20.
25. Scikit-learn. 2022. *1.1. Linear Models*. [online] Available at: <https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression> [Accessed 6 June 2022].
26.