IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

Salary Prediction in It Job Market

Navyashree M[1*], Navyashree M K[2], Neetu M [3], Pooja G R [4], Arun Biradar[5]

[1,2,3,4,5]Department of Computer Science, EWIT, Bangalore, India

DOI: https://doi.org/10.26438/ijcse/v7si15.7884 | Available online at: www.ijcseonline.org

***Abstract*—**In this study, Random Forest Regression machine learning rule is applied to predict salary levels of individuals supported their years of expertise. Random forest regression is employed since it gave higher accuracy compared to decision tree regression and Support Vector Regression classifier. Choosing the foremost effective machine learning rule so as to unravel the problems of classification and prediction of data is the most vital part of machine learning which depends on dataset likewise. The predictive accuracy of the Random forest regression on test data is 97%, while the accuracy of Decision tree regression and Support Vector Regression is 85% and 90% respectively. The model has been used on the training data to predict dependent variables and to extract features with highest impact on salary prediction.

***Keywords*:** Machine learning, Random forest regression, Decision tree regression, Support vector regression, salary prediction**.**

I. INTRODUCTION

Accurate recruitment of employees is a key element in the business strategy of every company due to its impact on companies' productivity and competitiveness. At present recruitment processes have evolved into complex tasks involving rigorous evaluations and interviews of candidates, with the goal of hiring the best suited professionals for each company's needs. With the advent of Internet and the web, e-Recruitment has become an essential element of all hiring strategies. Many websites, such as *CareerBuilder*, *Monster* or *Tecnoempleo*, or even Social Networks, like *LinkedIn*, help companies and jobseekers to find the best possible matches.

On the recruiter side, recent applications include: a framework for candidate ranking and resume summarization to improve recruiter's performance (Ref. 7), a tool for automatically evaluating candidates' CVs(Ref.9), and a system to screen candidates and score them, thus enabling candidate filtering and reducing the workload of recruitment officers' (Ref. 10).

In more detail, the main contributions of this paper are:

• An investigation on which fields in a job post have greater influence on salary and how they are interrelated;

• The discovery of the main data-driven profiles obtained through skill-set based aggregation;

•The formulation of the salary prediction problem as a classification task in order to have a better accuracy by focusing on discrete ranges instead of continuous salary values;

• The comparison of several classifiers, including SVM, Decision tree, random forests, in order to find the model with the best accuracy in predicting the salary range. This model

can be effectively employed by an e-Recruitment website to provide an automatic categorization of job posts by salary range.

As a last paragraph of the introduction should provide organization of the paper/article (Rest of the paper is organized as follows, Section I contains the introduction of the paper , Section II contain the related work of Prediction model, Section III explains the methodology with flowchart and different regression techniques, Section IV contain the results and discussions, section V contains the conclusion and future scope.

II.RELATED WORK

Himanshi, Komal Kumar Bhatia carried out a research study to determine the Prediction Model for Under-Graduating Student's Salary Using Data Mining Techniques.

Data mining is an essential concept as it provides the effective and efficient learning to the students and researchers [2]. Recommender systems are used to produce different aspects of the recommendations. Data mining is used to study the behavioral patterns of learners. Effectiveness can be improved by designing the individualized recommendation system for the students [3]. PornthepKhongchai, PokpongSongmuang in [4], proposes a salary prediction system for increasing students‟ motivation in studying. A decision tree technique is used to generate a prediction model with seven features. The system has improved by including more features suggested by users or academic staffs. To make a more effective system, occupations related to other studying fields can be added to the training dataset.

In [5], students are classified based on the marks scored by them using the Bayesian network. Using the training data-set a model is built to compare the relative performance of test data-set. 10-fold cross-validation is used for model evaluation. Ordinary least-squares (OLS) regression model is used by Jerrim in [6], to build prediction models for determining the future salaries of the student in American college based on family background data and social profiles. Whereas Karla et al [7] used hierarchal linear regression to build a prediction model with students and program characteristics as control variables and salary as the predictor variable. There are some issues in the above prediction model as 1) the prediction model only predicts the salary of the group of students not for the individual students, and 2) The results of the prediction model require extensive statistical knowledge to fully comprehend.

III . METHODOLOGY

The methodology follows the best practices in the literature and in the industry, including different phases:

1. *Data collection*: a Python-based web crawler‡ is developed to parse and gather the necessary information from the website. The crawler was run on a daily basis from December2015 to April 2016 and duplicates were removed.
2. *Data cleaning*: posts with missing values are removed and possible conflicts in the data format (e.g. text encoding) are fixed.
3. *Manual feature engineering*: irrelevant features are discarded and others are standardized (e.g. converted into numerical features) by exploiting the domain knowledge.
4. *Dataset description*: statistical tools and simple models are used in order to provide a preliminary and compact description of the data.
5. *Automatic feature selection:* feature selection algorithms are employed to automatically select the most informative features in the dataset with respect to the output class.
6. *Model selection:* a grid search is performed to find the optimal hyper-parameters for a set of well-known machine learning models.
7. *Model training and validation*: the selected models are trained and cross-validated in order to find the classifiers that best describe the data and are able to predict the output variable with the highest scores.
8. *Model comparison*: each model is compared to the others with respect to standard scores and curves like the classification accuracy, the F1 score, the ROC curve, the Precision-Recall curve etc.

In statistical modelling, regression analysis could be a set of statistical processes for estimating the relationships among variables. It includes several techniques for analysing many variables, once the main focus is on the connection between a dependent variable and one or a lot of freelance variables. Figure 1 shows the Predicting modelling.

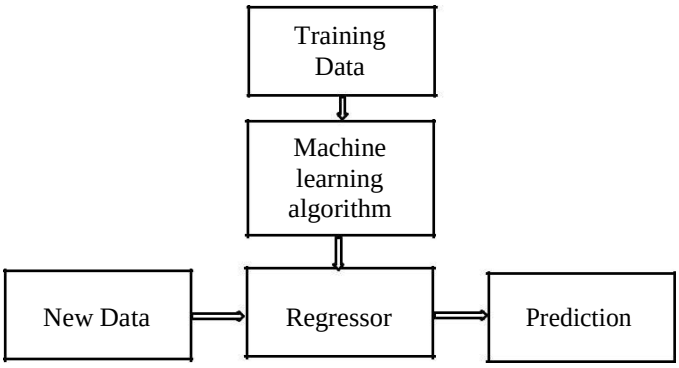

Figure 1. Predicting modelling

Types of regression techniques:

A. Random Forest Regression

Random Forest or random decision forests are an ensemble learning methodology of classification, regression and alternate tasks that operates by constructing a large number of decision trees at training time and outputting the category that's the mode of the classes(classification) or mean prediction(regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Figure 2 shows Random forest builds multiple decision trees and merges them along to urge a lot of correct and stable prediction. Random Forest could be a assortment of Decision Trees, but there are some variations. If you input a training dataset with features and labels into a decision tree, it'll formulate some set of rules, which can be accustomed build the predictions.

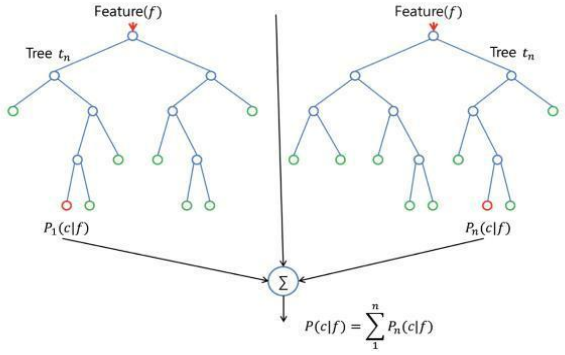

Figure 2: Basic Representation of Random Forest

Algorithm 1: Random forest regression

Input: Training data T, parameters{λ,δ,k,s,c}
Output: Model with evaluation
1. **Ensemble-RF**(T,λ,δ,k,s,c)
2. **for** i < −1 to s **do**
3. (train, test)←randomSplit(T,λ)
4. split ←bootstrap(train,δ,k)
5. model ←RandomForest.train(split,c)
6. score←evaluate(model,test)
7. out[i]← (model,score)
8. **end for**
9. **return** out

B. Support Vector Regression

Support Vector machine support linear and nonlinear regression which will be referred as SVR. Rather than attempting to suit largest Potential Street between two categories whereas limiting margin violations, SVR tries to suit as several instances as possible on the street whereas limiting margin violation. SVR tries to suit as several instances as potential on the street whereas limiting margin violation. The dimension of the street is controlled by hyper-parameter Epsilon.

Operating of SVR:

1. Collect a training set
2. Select a kernel and its parameters moreover as any regularization required.
3. Form the correlation matrix, K
4. Train your machine, precisely or about, to get contraction coefficients
5. Use that coefficient, produce your estimator.

We are arriving our correlation matrix and we are evaluating the kernel

$$. = (\Sigma_{|} - |^{2)+\epsilon} .$$

The main a part of the algorithmic program is $\vec{} \vec{} = \vec{}$

is the vector of the values like the training set,

$\vec{}$ is the correlation matrix, $\vec{}$ is that the set of unknowns we would like to solve for, $\vec{} = \vec{} -1\vec{}$

$$= (\Sigma_{|} - *|^{2)}$$

Once a parameter is thought from – estimator, we will use the coefficient we tend to found throughout the improvement step and also the kernel we started off with. To estimate the worth of y* from a test point, $\vec{}$ *- compute the correlation vector $\vec{}$, y*=$\vec{\vec{}}$. $\vec{}$.

C. Decision Tree Regression

Decision tree builds classification or regression models. Decision trees classify instances by sorting them down the tree from the foundation to some to some leaf node, that provides the classification of the instance. It's powerful and largely used modelling technique.

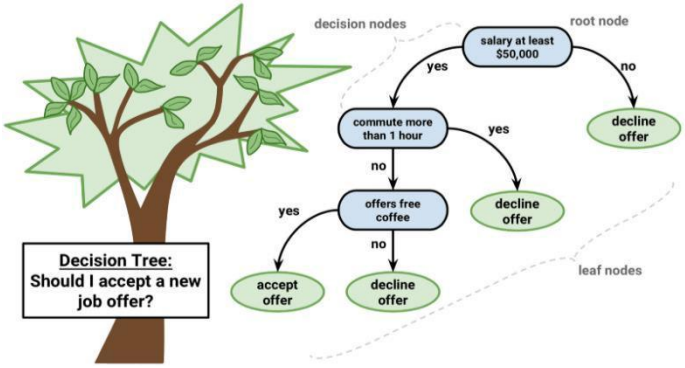

Figure 3: The illustrations of the decision tree model

Figure 3 shows the illustrations of the decision tree model is a binary tree. This is often your binary tree from algorithms and data structures, nothing too fancy. Every node represents a single input variable (x) and a split purpose thereon variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) that is employed to form a prediction. Predictions are created by walking the splits of the tree till incoming at a leaf node and output the category worth at that leaf node. Trees are quick to be told and extremely quick for making predictions. They're also often accurate for a broad range of issues and don't need any special preparation for our data.

PSUDOCODE

1. Set the current working directory.
2. Importing the dataset.
3. Splitting the dataset into the Training set and Test set.
4. Feature Scaling.
5. Creating the regression and fitting theregression to the dataset.
6. Predicting a new result.
7. Visualizing the Regression results.

IV .RESULTS AND DISCUSSION

To select the best data mining techniques for the data of employee salary of a company, we set up an experiment for comparing the performance measures of 3 regression techniques Random Forest, Support Vector Regression and decision tree on the selected data. The data include position, level, salary class. Data mining techniques were compared to predict salary using the data mining tool WEKA.

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

Figure 4: Dataset of position_salaries

Evaluating Regression Model performance

We use the backward elimination(Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss to fit) method to construct a multiple linear regression using our data and through and we constructed three separate model. we are looking at R-squared and adjusted R-squared and they will help us to improve our backward elimination method. Here, R-square tells us the goodness of fit R-squared is not greater than one and it as close as possible to one. The closer is one, the better is to fit the model. Here R-square and adjusted R-squared is very similar because adjust r is penalized factor. So, basically, if the variable that we added does not make adjusted or does not make R-squared grow that much like for instance. It only grew by fraction. Now, we use adjusted R to the goodness of fit all of our model and how it changes. Here, adjusted R-squared equation.——

$$\text{Adj } R^2 = 1-(1-R^2)\frac{}{}\frac{}{}^{-1}$$

p is the number of regressors, n is the sample size, p is the number of independent variables.

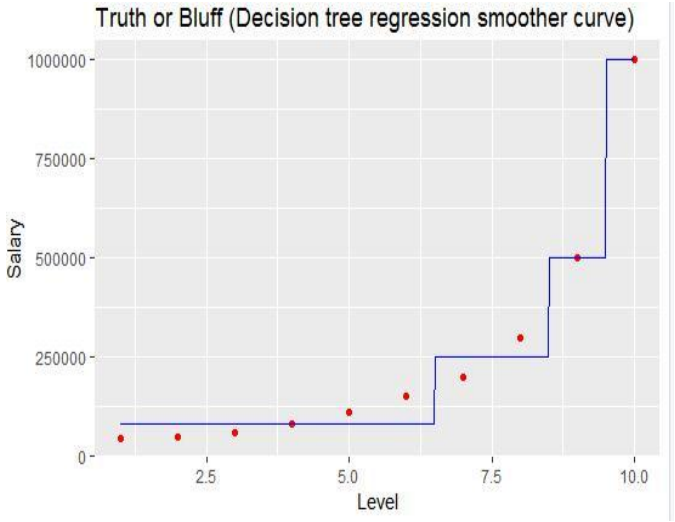

Figure 5: Truth or bluff obtained using decision tree

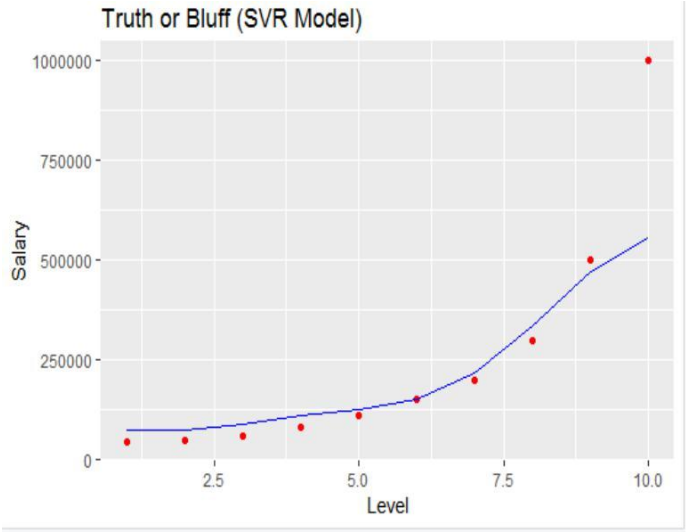

Figure 6: Truth or bluff obtained using Support Vector Regression (100 trees). Regression

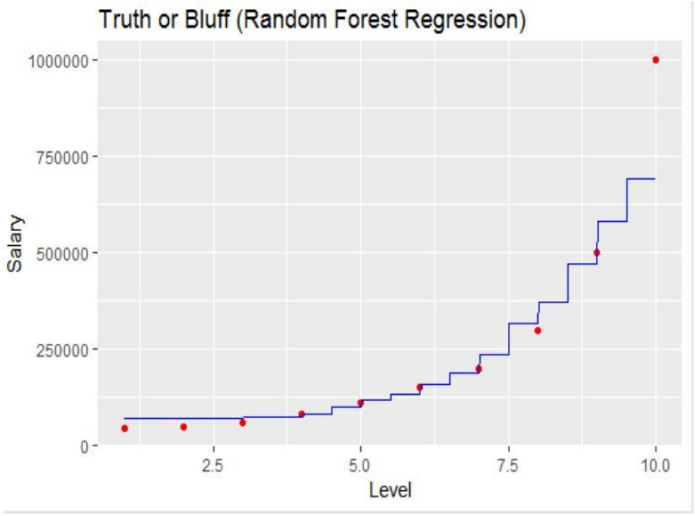

Figure 7: Truth or Bluff obtained using Random Forest

In figure 5 and 6 shows a red point are the real observation points that is real salaries associated. And all the point on this blue curve. Here is the prediction point. So for this level, the real observation point corresponds to this red point here and corresponds to this real salary. The prediction point is actually a projection of this red point on the blue curve. And that corresponds to the predicted salary.

In figure 5 we can see a graph, where we plotting a prediction of 10 salary corresponding to 10 level. It is nonlinear and non-continuous regression model decision. In visualization code, we apply the decision tree regression. After executing this code, we get a clear non continuous graph. By using a graph, we can easily predict the 6.5 level positions. In figure 6 we can see a graph, predicting the 6.5 level salary of our future potential employee. And now we need to include the 6.5 in this transform method of the sc_X scaler object. Here, this is an input here and the transfer method must be an array and the 6.5 is the numerical value. For this numerical value an array, we need to add np.array. We use the object of sc_Y, for the prediction of y and also use the inverse transform method.

After prediction, we get the result of $170,000.This is a better prediction because first, we get only $160k. So this is a better model. We got a good prediction of our 6.5 level salary.

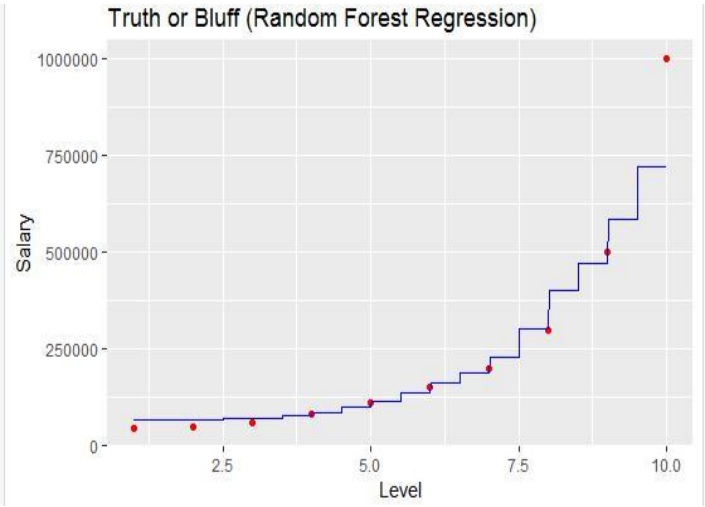

Figure 8: Truth or Bluff obtained using Random Forest Regression (500 trees).

Figure 7 shows the prediction of the output $237218 for 100 trees. This is very good prediction. So, now we say, these employees are not bluffing and it's an employee that we are happy to hire because not only is it good but also he is honest. In figure 8, we will get the output for 300 trees. After executing, we get the output $160k for 300 trees which is similar to the real value. The visualization Random Forest code, we can see a straight horizontal graph. It fit the data efficiently when compare to other two regressor models mentioned above. So this is a great model. We got a good prediction of our 6.5 level salary.

V .CONCLUSION AND FUTURE SCOPE

The paper presents a salary prediction system using data mining techniques. The system compares the employee's years of experience with the salary per annum. We compared data mining techniques that perform best in the task. An experiment was conducted using the 10000-organization data. The result indicated that Random forest regression gives the best result as compared to Decision Tree and Support Vector Regression. Random forest is a powerful tool for numerical prediction and fits through data points more efficiently. It can also be used to predict discrete classes also. Due to time and computation constraints, this model can be used for prediction.

Additionally, we would like to explore the possibility of further optimizing our feature extraction and selection techniques. In future work we would like to include features based on document statistics (document length, average number of words, Flesch-Kincaid readability statistics, etc.) as well as features based on similarities between documents (perhaps using latent dirichlet allocation, an unsupervised learning technique which assumes that the distribution of words in a document is indicative of underlying categories).

REFERENCES

[1] A. Parkavi1, K. Lakshmi "Predicting the Course Knowledge Level of Students using Data Mining Techniques ",IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), (2017).

[2] A. M Shahiri, W. Husain and N.A Rashid, "A Review on Predicting Students‟ Performance using Data mining Techniques," in The Third Information System Information Council

[3] R.A. Huebner, "A survey of educational data mining research," Research in Higher Education Journal.

[4] Improving Students‟ Motivation to Study using Salary Prediction System Pornthep Khongchai, Pokpong Songmuang Department of Computer Science Faculty of Science and Technology

[5] S. Anupama Kumar Vijayalakshmi M.N. "Inference of Naïve Baye‟s Technique on Student Assessment Data", R.V. College of Engineering, Bangalore, India

[6] John Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", Education Economics, 23:2, p 162-179, 2013.

[7] Karlar Hamlen and William A. Hamlen," Faculty Salary as a predictor of student outgoing salaries from MBA programs", Journal of Education for Business

[8] "Efficient Classification of Data Using Decision Tree" by Bhaskar N. Patel, Satish G. Prajapati, and Dr. Kamaljit I. Lakhtaria.

[9] Kayah, F. "Discretizing Continuous Features for Naive Bayes and C4. Classifiers". University of Maryland publications: College Park, MD, USA.

[10] S. Taruna, Mrinal Pandey, "An empirical analysis of classification techniques for predicting academic performance", IEEE Advances Computing Conference (2004).All references arranged in the following format for and remove website (URL) references or replaced by Journal references.

[11] F. Amato, R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, V. Moscato, F. Persia, and A. Picariello. Challenge: Processing web texts for classifying job offers. In Semantic Computing (ICSC), 2015 IEEE International Conference on, pages 460– 463, Feb 2015.

[12] Y. Abbound, A. Boyer, and A. Brun. Predict the emergence: Application to competencies in job offers. In Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on, pages 612– 619, Nov 2015.

[13] Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman. Knowledge discovery techniques for talent forecasting in human resource application. World Academy of Science, Engineering and Technology, 50:775–783, 2009.

[14] Chen-Fu Chien and Li-Fei Chen. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with Applications, 34(1):280 – 290, 2008.

[15] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. Choosing the right crowd: expert finding in social networks. In Proceedings of the 16th International Conference on Extending Database Technology, pages 637–648. ACM, 2013.

[16] R Schlogel, I Marchesini, M Alvioli, P Reichenbach, ¨ M Rossi, and J-P Malet. Optimizing landslide susceptibility zonation: Effects of dem spatial resolution and slope unit delineation on logistic regression models. Geomorphology, 2017. Faruk Bulut and Mehmet Fatih Amasyali. Locally adaptive k parameter selection for nearest neighbor classifier: one nearest cluster. Pattern Analysis and Applications, 20(2):415–425, 2017.

[17] Antonio Alarcon-Paredes, Gustavo Adolfo Alonso, ´ Eduardo Cabrera, and Rene Cuevas-Valencia. Simul- ´ taneous gene selection and weighting in nearest neighbor classifier for gene expression data. In International Conference on Bioinformatics and Biomedical Engineering, pages 372–381. Springer, 2017.

[18] Binh Thai Pham, Dieu Tien Bui, Hamid Reza Pourghasemi, Prakash Indra, and MB Dholakia. Landslide susceptibility assesssment in the uttarakhand area (india) using gis: a comparison study of prediction capability of na¨ıve bayes, multilayer perceptron neural networks, and functional trees methods. Theoretical and Applied Climatology, 128(1-2):255–273, 2017.

[19] Yudong Zhang, Yi Sun, Preetha Phillips, Ge Liu, Xingxing Zhou, and Shuihua Wang. A multilayer perceptron based smart pathological brain detection system by fractional fourier entropy. Journal of medical systems, 40(7):173, 2016.

[20] Jie Xu, Xianglong Liu, Zhouyuan Huo, Cheng Deng, Feiping Nie, and Heng Huang. Multi-class support vector machine via maximizing multi-class margins. In The 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), 2017.

[21] Ying Zhou, Wanjun Su, Lieyun Ding, Hanbin Luo, and Peter ED Love. Predicting safety risks in deep foundation pits in subway infrastructure projects: Support vector machine approach. Journal of Computing in Civil Engineering, 31(5):04017052, 2017.

[22] F Provost, C Hibert, and J-P Malet. Automatic classification of endogenous landslide seismicity using the random forest supervised classifier. Geophysical Research Letters, 44(1):113–120, 2017.

[23] Siddharth Hariharan, Siddhesh Tirodkar, Alok Porwal, Avik Bhattacharya, and Aurore Joly. Random forest-based

prospectivity modelling of greenfield terrains using sparse deposit data: An example from the tanami region, western australia. Natural Resources Research, pages 1–19, 2017.

[24] Michael Sprenger, Sebastian Schemm, Roger Oechslin, and Johannes Jenkner. Nowcasting foehn wind events using the adaboost machine learning algorithm. Weather and Forecasting, 32(3):1079–1099, 2017.

Authors Profile

Ms.Navyashree. M studying in 8th sem CSE dept, East West Institute of Technology.Area of Intrests are Machine learning, Artificial Intelligence

*Ms.*Navyashree.M.K studying in 8th sem CSE dept, EasT West Institute of Technology.Area of Intrests are Machine learning, Artificial Intelligence

Ms.Neetu.M studying in 8th sem CSE dept, EasT West Institute of Technology. Area of Intrests are Machine learning, Artificial Intelligence

Ms.Pooja.G.R studying in 8th sem CSE dept, EasT West Institute of Technology.Area of Intrests are Machine learning, Artificial Intelligence

Dr.Arun Biradar, He is currently working as head of department,Dept of Computer Science,East West Institute of Technology. He has presented more than 50 reference papers. . His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education.

