Academic Journal of Nawroz University (AJNU), Vol.12, No.3, 2023

https://doi.org/10.25007/ajnu.v12n3a1547

An Ensemble Machine Learning Approach for Classifying Job Positions

Ayaz Kh. Mohammed [1], Abdullahi Aliyu Danlami [2], Dindar I. Saeed [3], Abdulmalik Ahmad Lawan[4], Adamu Hussaini [5], Ramadhan Kh. Mohammed [6]

[1] Computer System Department, Ararat Private Technical Institute, Kurdistan Region - Iraq
[2] Department of Engineering, National Agency for Science and Engineering Infrastructure (NASENI), Abuja, Nigeria
[3] Employee in Scientific Journal University of Zakho, University of Zakho, Kurdistan Region - Iraq
[4] Department of Computer Science, Kano University of Science and Technology, 713281 Wudil, Nigeria
[5] Department of Computer Science, Kano University of Science and Technology, Nigeria
[6] Computer System Department, Ararat Private Technical Institute, Kurdistan Region - Iraq

ABSTRACT

Machine learning is one of the promising research areas in computer science, with numerous applications in automated detection of meaningful data patterns. Several data-centric studies were conducted on evaluating competencies, detecting similar jobs and predicting salaries of various job positions. However, the hazy distinction between closely related job positions requires powerful predictive algorithms. The present study proposed an ensemble approach for accurate classification of various job positions. Accordingly, different machine learning algorithms were applied on 955 instances obtained from Glassdoor using web scraping. Furthermore, the present study classify various job positions based on average salary and other correlated explanatory variables that cover many aspects of job activities on the internet. The study result revealed the superior performance of heterogeneous ensembles in terms of precision and accuracy. The proposed data-centric approach produce strong models for researchers, recruiters, and candidates to assigned job positions and its competencies.

Keywords: Machine learning, heterogeneous ensemble, job position, multi-class classification

1. INTRODUCTION

Dictionary definitions of "career" implies a job or profession that someone does for a long time. It also means a period of time spent in a job or profession which can be described as an advancement in the form of a linear progression. Job seekers as well as recruiters demand accurate classification of career opportunities in terms of various parameters such as competencies, responsibilities and salaries. Accordingly, several data-centric studies were conducted on evaluating competencies, detecting similar jobs and predicting salaries of various job positions [1]–[4].

However, the hazy distinction between closely related job positions requires powerful predictive algorithms. Thus, ensemble modelling, that utilize a cooperation of models in solving complex problems, could be a vital approach for improving the accuracy of the classification task. Ensemble modelling is an essential approach for dealing with complex systems especially when there is dynamic change in states between the interrelated entities. Thus, ensemble modelling could yield promising results in modelling career-related parameters such as job positions [2], [3].

The present study proposed an ensemble approach for accurate classification of various job positions based on various explanatory variables on different aspects of internet-based job activities. Accordingly, different machine learning algorithms were applied on 955 cases extracted from Glassdoor using web scraping. In essence, we sought a Heterogeneous Ensemble by aggregating the model accuracies and computing the average mean. Specifically, Adaboost Classifier (ADC), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), XG Boost Classifier (XGB), Extra Trees Classifier (EXC) were evaluated based on their experimental performances on the dataset obtained.

2. RELATED WORK

Several studies have utilized demographic and empirical parameters as well as various analysis techniques in the prediction of job-related measures. For instance, Dreher [5] described the degree to which company-generated data could be utilized in the prediction of employees' salary satisfaction. Dreher examined varying predictors such as educational level, monthly salary, years of continuous services, annual performance rating, a measure of the most recent salary increase, an estimate of career potential, and employee gender. Similarly, Alexander [6] employed parameters related to cognitive ability for the prediction of long-term job performance and career goals amongst more than 3,000 employees of an international technology company. The study findings revealed that job performance measures, both objective and subjective, such as level of promotion, salary, and supervisory performance ratings significantly account for the change in the variance of aptitude test scores. Nonetheless, with inconsistent variations in relation to some of the factors considered, the authors suggested that aptitude test is not a sufficient indicator to be considered in the empirical prediction of long-term job performance and success. In the same vein, Martín [3] conducted an empirical study, using 4,000 data instances gathered from an IT recruitment portal in Spain, to identify the most demanded and rewarded competencies in job offers that are the watch ward for job seekers and recruiters. It was found that experience has higher precedence over educational level. Based on the required skills, the authors further identified five profile clusters and utilized tree-based ensembles in the development of accurate salary-range classifier. Furthermore, Mainert [1] examined the influence of complex cognitive parameters of general mental ability and problem-solving skills in predicting of job-related parameters such as job level, complexity, as well as salary. The study findings indicated that CPS is linked to job complexity and salary but offered no significant influence in predicting job level. Recently, Dutta [4] utilized a Kaggle dataset containing 17,876 instances in detecting fake job recruitment using machine learning approach. The binary classification problem was aimed at checking fraudulent job adverts from online platforms. Comparative experimental results of the study revealed the superior performance of ensemble classifiers over single classifiers. The present study approached job position multi-label classification problem using heterogeneous ensemble based on average salaries and some explanatory variables that cover the different career-related aspects.

3. METHODOLOGY

3.1 The Research Model

The present study adapted the cross-industry process for data mining (CRISP-DM) methodology in planning the structured data mining approach of the study. CRISP-DM is a robust, flexibility, and useful model when using analytics to solve complex business problem [7]. It idealize sequence of events or tasks that can be performed in a different order and might necessitate backtracking to previous tasks and repeat certain actions. All possible routes through the data mining process of the study are depicted with the help of Figure 1.

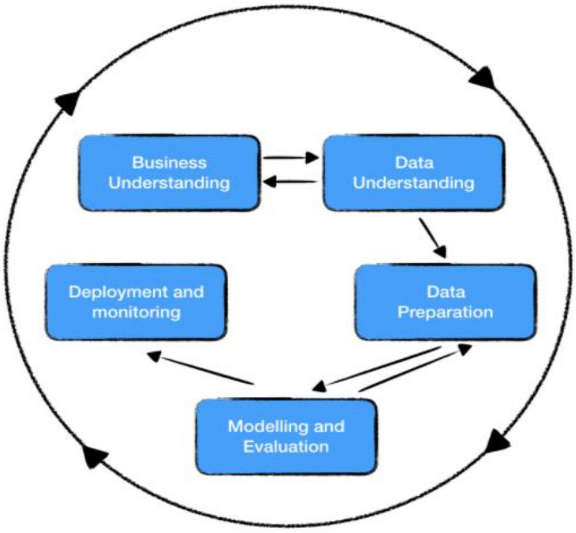


Figure 1: The research model

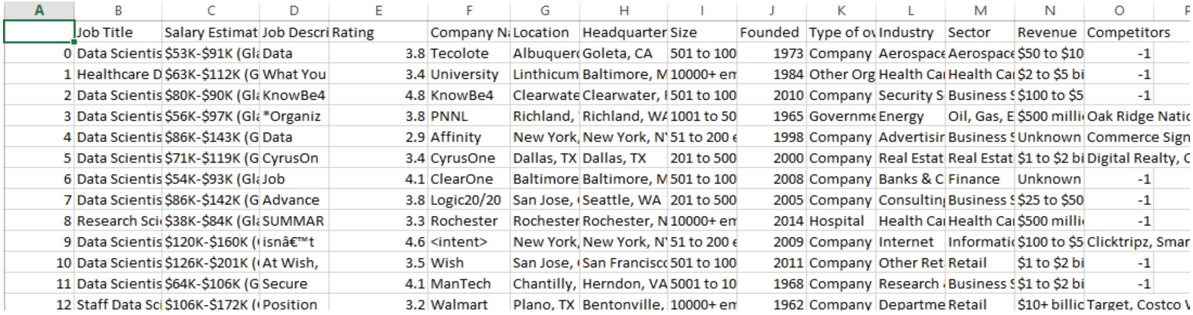
The research model constituted five stages with the aim of classifying job positions based on average salaries. The stages involved includes business understanding, data understanding, data preparation, modelling and evaluation.

3.2 Business Understanding

Business understanding involves uncovering the importance of "Job position" classifications based on competitive measures on average salaries, which could influence the outcome of the work [8], [9]. Accordingly, the present study considered various jobs positions compiled by investors within the datasets of the study [4]. Thus, the classification and detection of the job position were approached as a multi-class classification problem, with the help of machine learning algorithms and other data analytics techniques. Job positions were classified as "data scientist", "data engineer", "analyst", "machine learning", "manager", "director" and lastly all other job positions are level as "other job positions", so that the same algorithms and models could possibly be modified for larger datasets.

3.3 Data Understanding

The second stage of the CRISP-DM process of the study involves data collection from the Glassdoor website using web scraping (i.e. Selenium Web scraper) based on the procedure described by Sakarya [10]. The dataset used in the present study is open-sourced in a github repository; scraping-glassdoor-selenium [11]. The schematic structure of the dataset is depicted with the help of Figure 2.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Job Title	Salary Estimat	Job Descri	Rating	Company N	Location	Headquarter	Size	Founded	Type of o	Industry	Sector	Revenue	Competitors
0	Data Scientis	$53K-$91K (Gl	Data	3.8	Tecolote	Albuquer	Goleta, CA	501 to 100	1973	Company	Aerospace	Aerospace	$50 to $10	-1
1	Healthcare D	$63K-$112K (G	What You	3.4	University	Linthicum	Baltimore, M	10000+ em	1984	Other Org	Health Ca	Health Ca	$2 to $5 bi	-1
2	Data Scientis	$80K-$90K (Gl	KnowBe4	4.8	KnowBe4	Clearwate	Clearwater,	501 to 100	2010	Company	Security S	Business	$100 to $5	-1
3	Data Scientis	$56K-$97K (Gl	*Organiz	3.8	PNNL	Richland,	Richland, WA	1001 to 50	1965	Governme	Energy	Oil, Gas, E	$500 milli	Oak Ridge Natic
4	Data Scientis	$86K-$143K (G	Data	2.9	Affinity	New York,	New York, N	51 to 200 e	1998	Company	Advertisir	Business	Unknown	Commerce Sign
5	Data Scientis	$71K-$119K (G	CyrusOn	3.4	CyrusOne	Dallas, TX	Dallas, TX	201 to 500	2000	Company	Real Estat	Real Estat	$1 to $2 bi	Digital Realty, C
6	Data Scientis	$54K-$93K (Gl	Job	4.1	ClearOne	Baltimore	Baltimore, M	501 to 100	2008	Company	Banks & C	Finance	Unknown	-1
7	Data Scientis	$86K-$142K (G	Advance	3.8	Logic20/20	San Jose,	Seattle, WA	201 to 500	2005	Company	Consultin	Business	$25 to $50	-1
8	Research Sci	$38K-$84K (Gl	SUMMAR	3.3	Rochester	Rochester	Rochester, N	10000+ em	2014	Hospital	Health Ca	Health Ca	$500 milli	-1
9	Data Scientis	$120K-$160K (	isnâ€™t	4.6	<intent>	New York,	New York, N	51 to 200 e	2009	Company	Internet	Informati	$100 to $5	Clicktripz, Smar
10	Data Scientis	$126K-$201K (	At Wish,	3.5	Wish	San Jose,	San Francisc	501 to 100	2011	Company	Other Ret	Retail	$1 to $2 bi	-1
11	Data Scientis	$64K-$106K (G	Secure	4.1	ManTech	Chantilly,	Herndon, VA	5001 to 10	1968	Company	Research	Business	$1 to $2 bi	-1
12	Staff Data Sci	$106K-$172K (	Position	3.2	Walmart	Plano, TX	Bentonville,	10000+ em	1962	Company	Departme	Retail	$10+ billic	Target, Costco V

Figure 2: Schema structure of the dataset

The schema highlighted several parameters including six (6) numerical features/columns that were selected in conducting correlations analysis between dependent and independent variables of the study. The dataset constituted independent parameters such as salary estimate, size, revenue, job description, rating as well as the dependent variable job titles. Where, the salary ranges for a particular job position were defined with the float value – salary estimate. Size is an integer value that captures the number of active company staff. Revenue is a floating-point value that signifies the company's yearly turnout. While the integer value "Job description" stores a character description of the workload assigned to an employee. Lastly, the floating-point sentiment value named "Rating" stores number of the people that shows interest in the company's services. Furthermore, the dataset contains nine hundreds and fifty-five (955) unique cases including three hundred and fifty-eight (358) "data scientist", two hundred and thirty-eight (238) "other Job positions", hundred and fifty-eight (158) "data engineer", hundred and twenty-four (124) "analyst", thirty-six (36) "manager", twenty-six (26) "Machine Learning" and sixteen (16) "directors" categorized and utilized in the conducting the classifications task with "Job title" target attribute of the dataset. The distribution of the job positions is described pictorially with the help of Figure 3.

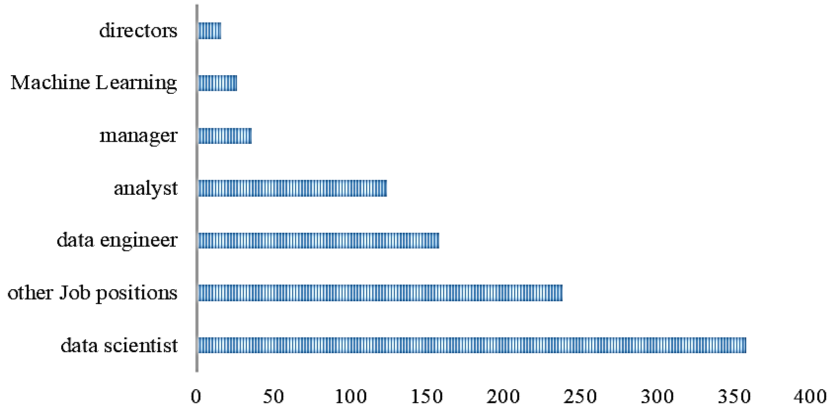


Figure 3: Simplified job position

3.4 Data Preparation

Data preparation also known as data pre-processing involves removing unwanted or noisy attributes from the dataset using data cleansing techniques. In the present study, unexpected delimiters were removed from some attributes, missing values were replaced using "Last Known Value" forward and backward filled, and outliers were identified using interquartile ranges and observations containing more than two (2) outliers were removed. We converted continuous attributes into bins, which significantly improve model performance while reducing class imbalances. We removed all columns where correlation matrix does not show any relationship with the targeted label, and lastly we used Label Encoding process in converting categorical labels to numeric values.

3.5 Modelling and Evaluation

Model Evaluation helps in finding the best model that represents the research data and how well the chosen model will work in the future. In the present study, to avoid over-fitted models training and testing datasets were created before building the individual models as well as the Ensembles. For each machine-learning algorithm, we tune model performance and build heterogeneous Ensembles with the best performing model. Furthermore, we check for variance in the results using k-fold cross-validation. Accordingly, we created appropriate training and testing splits for the classification models using the Holdout method of 80-20 data split; all the models were built using two splits of data to check variance and performance. The two splits of data were created twice with one data for AdaboostClassifier (ADC), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), XG Boost Classifier (XGB), Extra Trees Classifier (EXC), to form Heterogeneous Ensemble (Transformed and Normalized). The modelling and evaluation approach is further explained with the help of Figure 4.

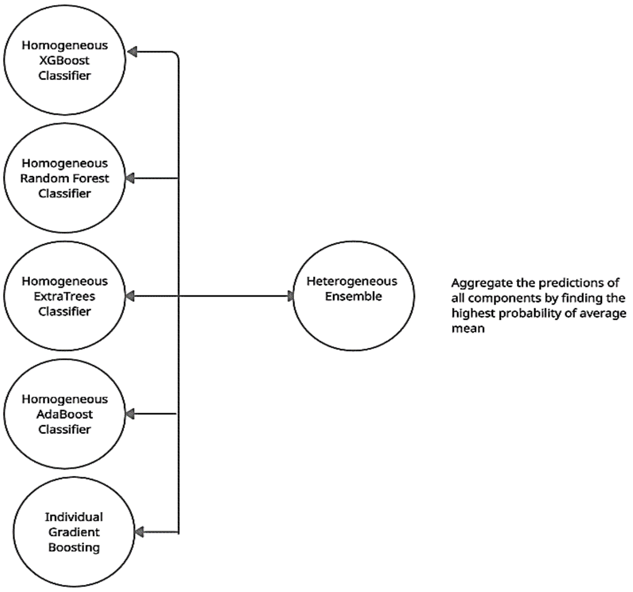

Homogeneous XGBoost Classifier
Homogeneous Random Forest Classifier
Homogeneous ExtraTrees Classifier
Homogeneous AdaBoost Classifier
Individual Gradient Boosting
Heterogeneous Ensemble
Aggregate the predictions of all components by finding the highest probability of average mean

Figure 4: Heterogeneous Ensemble

Confusing matrix and multiple evaluation metrics were employed in the multiclass classification problem to evaluate the performance of each of the models developed. The metrics utilized includes ROC, accuracy, recall, precision, F1-score, "mean squared error" (MSE), "root mean square error" (RMSE) as well as the Matthews's correlation coefficient.

Specifically, accuracy is calculated from true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as the proportion of the true results among the total number of cases examined, which is expressed as follows:

Accuracy = TP+TN/ (TP+TN+FP+FN).

Precision is the proportion of predicted positives that are truly positive:

Precision = TP/ (TP+FP)

Recall stands for the proportion of actual positives that is correctly classified:

Recall = TP/ (TP+FN).

F1-score refers to the mean of precision and recall expressed as follows:

F1-score = (2×Precision×Recall)/(Precision + Recall).

Furthermore, ROC is a probability curve which represents degree or measure of reparability. It tells how much model is capable of distinguishing between classes. However, the two statistical measures of MSE and RMSE measures the difference between the estimated values and what is estimated. Lastly, Matthews's correlation coefficient measures the quality of classification tasks both binary and multiclass by considering various relevant parameters including true positives, true negatives, false positives, and false negatives. In essence, MCC is nothing but a coefficient of correlations that lies between the values -1 and +1 and generally regarded as a balanced metric that can be relied upon even with significant differences in class size. Thus, MCC coefficients of -1, 0 and +1 denote inverse, average random and perfect predictions, respectively.

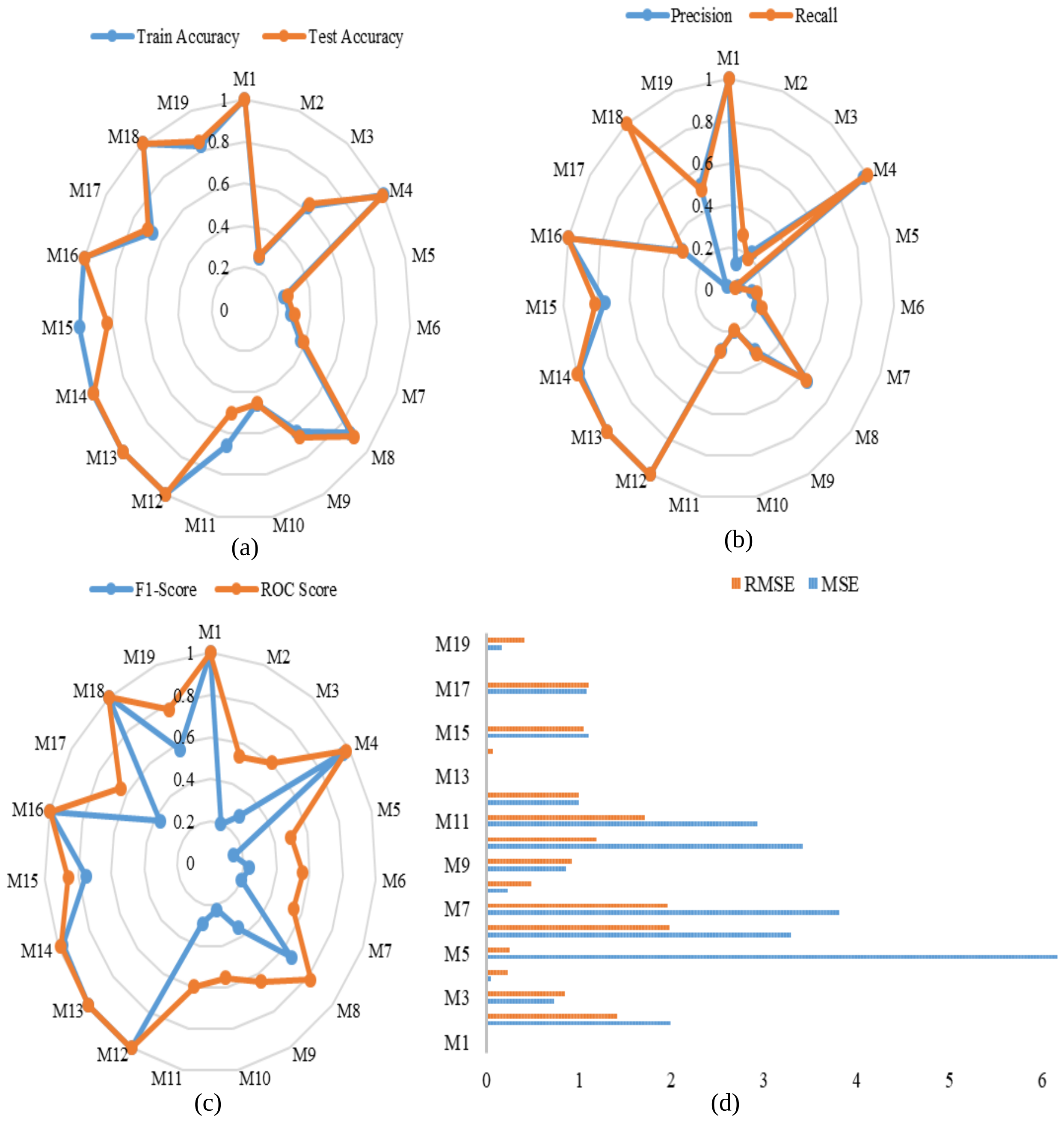
4. EXPERIMENTAL RESULTS

The empirical results of the modelling and analysis conducted in the present study is summarized with the help of tables and figures. The findings of the experimental approach applied on the dataset using nineteen (19) distinct classification algorithms is summarized with the help of Table 1. The classifiers that performed better on test accuracy were selected after performing cross-validation to form a heterogeneous Ensemble.

Table 1: Evaluation metrics for the various classifiers

Model ID	**MLA Name**	**Train Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC Score**	**MSE**	**RMSE**	**MCC**
M1	XGB classifier	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
M2	Sgd Classifier	0.2592	0.2708	0.1285	0.2717	0.1949	0.5346	1.9843	1.4086	0.2255
M3	Ridge Classifier CV	0.6204	0.6354	0.2223	0.1833	0.2857	0.6036	0.7291	0.8539	0.5134
M4	Random Forest Classifier	1.0000	0.9896	0.9726	0.9960	0.9535	0.9756	0.0520	0.2282	0.9861
M5	Quadratic Discriminant Anal.	0.2448	0.2656	0.0599	0.0379	0.1429	0.5000	6.1667	0.2483	0.0000
M6	Perceptron	0.2775	0.3021	0.1399	0.1679	0.2362	0.5583	3.2919	1.9804	0.0000
M7	Passive Aggressive Classifier	0.3652	0.3854	0.1824	0.2154	0.2042	0.5467	3.8073	1.9512	0.2154
M8	Logistic Regression CV	0.8665	0.8958	0.6435	0.6340	0.6611	0.8206	0.2344	0.4841	0.8609
M9	Linear SVC	0.6636	0.6927	0.3316	0.3524	0.3512	0.6409	0.8594	0.9270	0.6149
M10	Linear Discriminant Analysis	0.4568	0.4531	0.2024	0.1960	0.2247	0.5544	3.4115	1.1847	0.2165
M11	K Neighbors Classifier	0.6571	0.5000	0.2923	0.2989	0.2928	0.5974	2.9219	1.7094	0.3156
M12	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
M13	Gaussian NB	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
M14	Extra Trees Classifier	1.0000	0.9948	0.9831	0.9980	0.9714	0.9851	0.0052	0.0722	0.9930
M15	Extra Tree Classifier	1.0000	0.8299	0.7504	0.8075	0.7538	0.8606	1.1094	1.0533	0.7642
M16	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
M17	Bernoulli NB	0.6649	0.6979	0.3380	0.3289	0.3650	0.6491	1.0885	1.1043	0.6122
M18	Bagging Classifier	1.0000	1.0000	0.0168	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
M19	Ada Boost Classifier	0.8220	0.8443	0.5267	0.499	0.5714	0.7715	0.1719	0.4146	0.8052

The results of modelling and data analysis carried out on the dataset are pictorially depicted with the help of Figure 5(a)-(d) in relation to the various measures of performance.



(a)

(b)

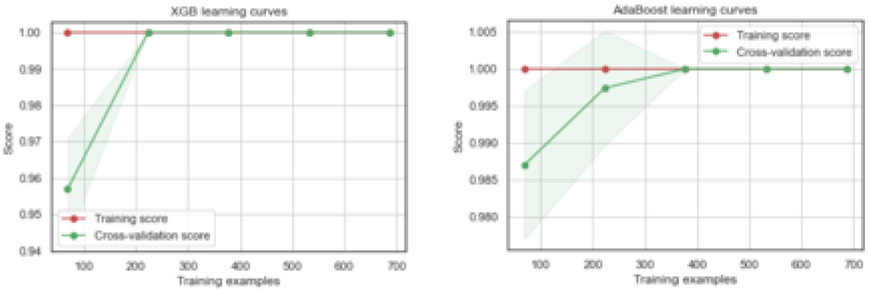
(c)

(d)

Figure 5(a)-(d): The result of various evaluation metrics from the data modelling

4.1 Models Cross-validation

The present study employed a stratified 10-fold cross validation in the comparison and evaluation of the



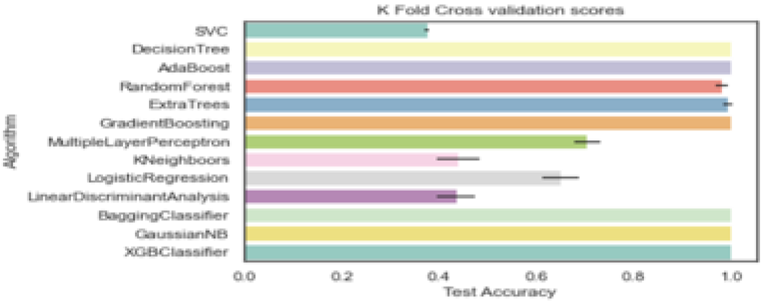


Figure 6: K-fold cross-validation of the models

Classification accuracies of the thirteen (13) distinct algorithms involved. Accordingly, Figure 6 highlighted the models involved in the stratified 10-fold cross-validation procedure. Consequently, the best performing models were selected after the cross-validation based on classification accuracies. Specifically, the five best performing algorithms selected were AdaBoost, Gradient Boosting, Random Forest, ExtraTrees, and XGBoost Classifier, which were further utilized in the conducting the Ensemble technique. Figure 7(a)-(e) depicted the models plotted for learning performances. Accordingly, the classifiers achieved improved performances in the learning curves.

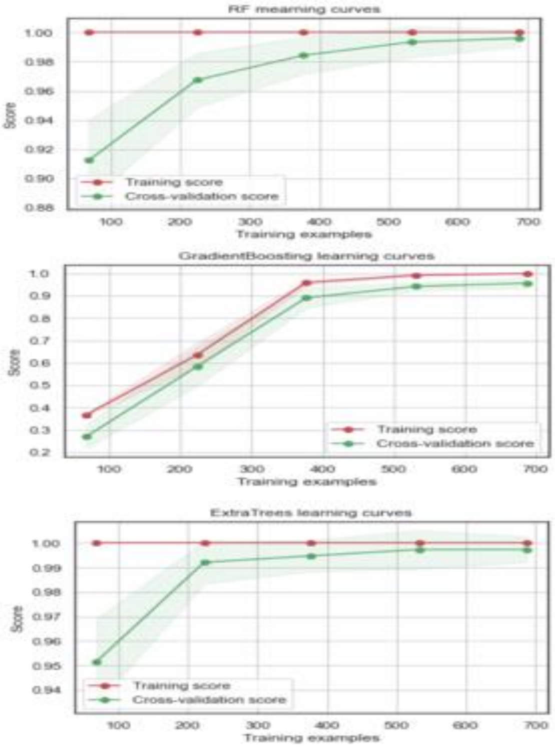


Figure 7(a)-(e): Learning curve of the various models

4.2 Hyperparameter Tuning for the Best Models

The present study involved conducting a grid search optimization for the best performing classifiers namely: Random Forest, AdaBoost, XGB Classifier, ExtraTrees, and Gradient Boosting. Accordingly, since four CPUs were involved in the process, the "n_jobs" parameter was set to the value of four (4), whose computation period approximately reached fifteen (15) minutes. Thus, the computation overhead was readily decreased as shown in Figure 8

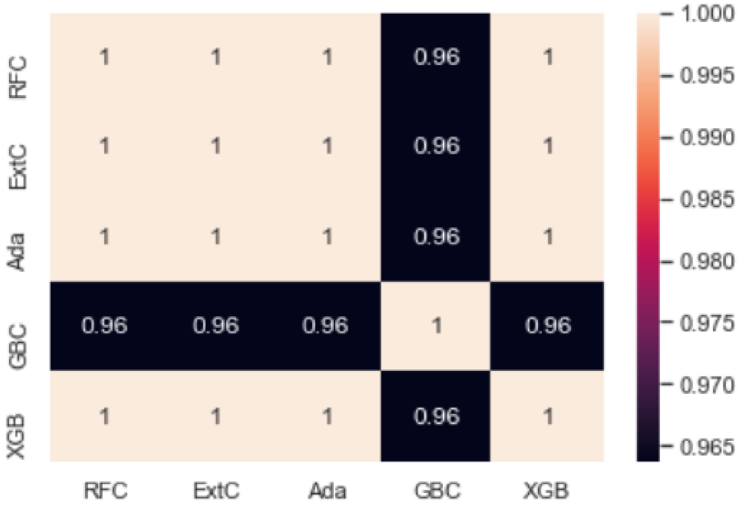


Figure 8: Tuning of the best models

The prediction appears to be similar for the five classifiers except for the gradient boosting classifier when compared to the other classifiers. The five classifiers give plus or minus the same prediction, but there are some differences. These differences between the five classifiers' predictions are sufficient to consider an Ensemble vote. We tendered to pass the argument "soft" to the voting parameter to take into account the probability of each vote. The metrics of the voting classifier are shown in Table 2.

Table 2: Heterogeneous ensemble based prediction

MLA Name	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	ROC Score	MSE	RMSE	MCE
Voting Classifier	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0

5.3 Actual and Predicted Estimates

The model was successfully developed, and the voting classifier was examined, that gave more test and train accuracies, precision, recall, and Matthew's correlations coefficient. The results of the first ten (10) instances were generated as shown in Table 3.

Table 3: Actual and predicted results

S/N	Actual	Predicted
0	3	3
1	2	2
2	2	2
3	0	0
4	0	0
5	4	4
6	0	0
7	3	3
8	3	3
9	3	3

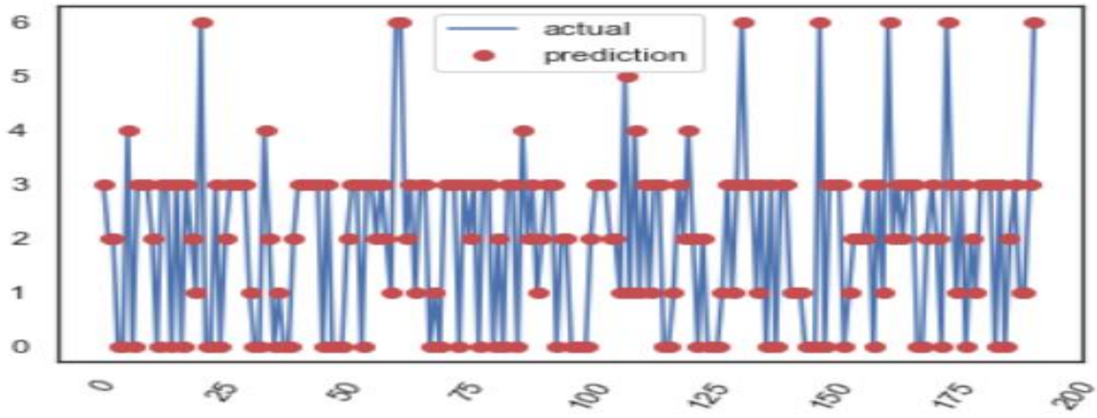


Figure 9: Actual and predicted results

The model exhibits good performance in both actual and predicted estimates as shown in Table 3 and Figure 3, hence it is excellent for Job Position detection.

6. Summary and Conclusion

This study focuses on the challenge of classifying job positions offered by companies based on average salaries and some explanatory variables that cover the different aspects of job parameters on the web. The multi-label classification approach of the study utilized 955 instances from a dataset containing multiple job positions obtained from Glassdoor website using web-scraping technique. Nineteen (19) machine-learning algorithms were applied on the dataset. Accordingly, assorted evaluation metrics were employed in evaluating the performance of the machine-learning algorithms. Consequently, five best performing algorithms to formed heterogeneous Ensemble. The comparative study results indicated the superior performance of heterogeneous ensembles in predicting job positions. The heterogeneous ensembles achieved an absolute accuracy of approximately 100% with soft voting. The study is limited by the amount of instance considered from the dataset. Thus, future studies are recommended to generate huge datasets to provide more comparative arguments on the performance of the machine learning algorithms.

REFERENCES

[1] J. Mainert, C. Niepel, K. R. Murphy, and S. Greiff, “The Incremental Contribution of Complex Problem-Solving Skills to the Prediction of Job Level, Job Complexity, and Salary,” *J. Bus. Psychol.*, vol. 34, no. 6, pp. 825–845, Dec. 2019, doi: 10.1007/s10869-018-9561-x.

[2] A. Petrakova, M. Affenzeller, and G. Merkurjeva, “Heterogeneous versus Homogeneous Machine Learning Ensembles,” *Inf. Technol. Manag. Sci.*, vol. 18, no. 1, Jan. 2015, doi: 10.1515/itms-2015-0021.

[3] I. Martín, A. Mariello, R. Battiti, and J. A. Hernández, “Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study,” *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, p. 1192, 2018, doi: 10.2991/ijcis.11.1.90.

[4] S. Dutta and S. K. Bandyopadhyay, “Fake Job Recruitment Detection Using Machine Learning Approach,” *Int. J. Eng. Trends Technol.*, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

[5] G. F. Dreher, “Predicting the salary satisfaction of exempt employees,” *Pers. Psychol.*, vol. 34, no. 3, pp. 579–589, Sep. 1981, doi: 10.1111/j.1744-6570.1981.tb00497.x.

[6] S. G. Alexander, “Predicting long term job performance using a cognitive ability test,” 2007.

[7] S. V. Europe, “What is the CRISP-DM methodology,” 2018. https://www.sv-europe.com/crisp-dm-methodology/ (accessed Feb. 16, 2022).

[8] A. Ikudo, J. Lane, J. Staudt, and B. A. Weinberg, “Occupational Classifications: A Machine Learning Approach,” *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3238563.

[9] I. M. Nasser and A. H. Alzaanin, “Machine Learning and Job Posting Classification: A Comparative Study,” *Int. J. Eng. Inf. Syst.*, vol. 4, no. 9, pp. 6–14, 2020, [Online]. Available: www.ijeais.org

[10] Omer Sakarya, “Selenium Tutorial: Scraping Glassdoor.com,” *Medium*, Oct. 14, 2019. https://mersakarya.medium.com/selenium-tutorial-scraping-glassdoor-com-in-10-minutes-3d0915c6d905 (accessed Feb. 27, 2022).

[11] scraping-glassdoor-selenium, “GitHub - arapfaik/scraping-glassdoor-selenium: Jupyter Notebook from Selenium Tutorial: Scraping Glassdoor.com".” https://github.com/arapfaik/scraping-glassdoor-selenium (accessed Jun. 20, 2023).