Estimation of Company sizes with Professions, Employment Types and on Employment Salaries.

|  |  |
| --- | --- |
| Rimsha Riaz | Syed Hamid Raza |
| *Dalarna University, Borlange, Sweden* | *Dalarna University, Borlange, Sweden* |
| *V22rimri@du.se* | *V22syera@du.se* |
|  |  |
|  |  |

***Abstract***

Human resource is the most powerful asset of every organization. They play a crucial role in the growth of the company and strengthen the organization. On other side, leading companies award different types of facilities to their employees with the handsome salaries packages. The salary package of the employee varies company to company. Different number of factors like company size, market value, physical location are the factors that directly impact the employee salary. In short, the employee salary and other factors can help to estimate the size of the company and market value of the company. The accurate estimation of company size will help the client on the time of project finalization. Here, we trained different machine learning model for the prediction of company size base on the different employee attributes. We also got the 0.79 accuracy with the fine-tuned Random Forest model.

***Index Terms - Correlation Analysis, Sales Predictions, Multiple Linear Regression, Descriptive Analysis.***

I. Introduction

It has always been important for us to estimate the impact of professions, employment types and salaries on company sizes. The variables which are affected the company size is dependent on the different variables like number of employees, employee salaries and employment types. The company size also replicates the market value, capital value and resource value of that company. The employee of an employee can translate the true value/size of that company. By considering all of these elements, we will predict the size of the company based on the employees and company related features. The company size will depend on the other variables that will drive independently.

Different studies have been published on the topic of salaries prediction using company profile. But the prediction of company using employee profile will open the new research gateway. The estimation of company size will also help the business persons, clients, stakeholders and other organization to deal with the company according to the company size. The proposed research will help to find whether the employee profile can directly estimate the company. We divided our paper in six sections. Section I as introduction, Section II as Literature review. Section III as problem statement, section IV methodology of our study. Section V as our study ‘s results. Last Section VI has conclusion of our study.

II. LITERATURE REVIEW

Since the majority of the industry depends on accurate future forecasting and prediction, models derived from the data-mining techniques are widely utilized in the industry. The most common types of forecast prediction analysis are categorization and extrapolation. Regression forecasts values and trends for unavailable data availability, whereas categorization forecasts the class labeling for the data. The classification forecasts the model for the categorical dataset, which is discontinuous and unordered, whereas the regression forecasts missing and unavailable numbered data. Segmentation is a component of prediction analysis that receives the most attention. In order to do this, the model which identifies and characterizes datasets must be derived.

Currently, corporate performance is regarded from the following angle: [1] noted which was critical to be capable to recognize the existing and potentially future profitable enterprises as well as pinpoint what motivates profitable businesses. It is also important for the start-ups and pre-revenue enterprises to comprehend which non-economic actions impact the survival chances, along with company achievement. Various fatality rates for businesses exist despite evaluating businesses with the same economic performance, suggesting either enterprise has a minimum feasible stage of economic performance and that non-economic performance is impacting the death risk. Thus, according to [2], factors influencing both economical and non-economic performance, as well as minimal viable stages of the performance, are exploited for the company's sustainability. According to the Gimero's comment mentioned above, it is crucial to look at the factors which affect both the economical and non-economic performance because they have an impact on the likelihood that company will succeed.

The two main categories of large company failure scenarios are the quantitative approaches and qualitative approaches. Although qualitative models depend on the internal organization assessments, quantitative models are dependent on the financial facts. The two various models look for characteristics—financial or otherwise—that can be used to distinguish between the firms that will survive and those that won't. Financial measures which distinguish among the surviving, no-surviving and formerly failed enterprises are identified by quantitative models. The foregoing are recognized financial indicators which are frequently used to pinpoint inability: Low profits related to assets and obligations; fewer equity returns, including dividends and investment; bad liquidity; excessive gearing; good income fluctuation. The qualitative system uses financial measurements that take into account qualitative and non-accounting characteristics to gauge the performance of the business. There have been multiple attempts to use the prediction to diagnose company success. Connection, induction or deduction, analysis and synthesis, charts of grading or evaluation, and other concepts were overused[3]. The program frequently foretells when the company is ready to declare bankruptcy. When selecting the points indicating a recognized subject or when awarding vital points, the critical portion is represented by the competence and competence of the budgetary inspector. He would be capable of evaluating qualitatively and quantitatively with objectivity and sufficiency thanks to all these. In order to capture the most recent company appraisal, which would then serve as the foundation for the decision-making, numerous institutions, credit bureaus, and governmental organizations use probabilistic modeling in the SWOT approach (to validate the quantitative technique). The main benefit of current model is that it helps a company identify that it is having the financial difficulties therefore, if identified in time, can prevent bankruptcies. However, the disadvantage of the current model is that, if the problem is not identified in time, the institution would not have the chance to survive. That model was created by [4] to diagnose international companies and assess company performance as well as risk exposure, encompassing collapse and financial distress. [5] showed in the study whether entrepreneurship aspirations are connected to a variety of human distinctions, including gender, whether it is female or a male, education, whether it is an MBA or an undergraduate, as well as a business person. They also showed why a proactive personality is associated with the company's prosperity. Considering three circumstances—markets with human-based forecasters; marketplaces with artificial intelligence-based networks; as well as markets with both humans and agents [6], they utilized expectations from marketplaces to complement predictions from diverse prognosticators.

To examine this study and the nature of the forecasting, a variety of indicators and parameters were used, particularly preciseness (measuring utilizing 3 simple scoring laws), Quant proportions (which are typically employed in finance to quantify compensation and danger performances, enabling us both to evaluate accuracy versus oscillations of the forecasting errors, making the comparison more illuminating). A website framework [7] was made available to company owners and innovators who wanted to assess their ventures and create company plans. In order to analyze investment applications and facilitate the evaluation of business plans, the study produced a framework of 32 factors. These parameters and their corresponding weights depend on the research into an investment plan used by investment firms for specific products and markets. Perhaps an entrepreneurial company, an individual, or a third party performs the examination. A self-evaluation, an entrepreneurial group, or an outside auditor does a review (e.g., the industry expert or the scientist). The achievement motivation scale was created as a predictor of entrepreneurial intention. They conducted research on executing company predictions through economic models that combined humans and artificial intelligence.

The assumptions are made by a hybrid humans-and-agents market. The benefit is that they offer a nice sharp ratio compromise between correctness and unpredictability of posterior probability. By creating a ROC model, hybrid markets offer a better exchange between wise and foolish choices. [8]. [9] created the data-mining method employing decision trees to predict potential company sectors for lending in consumer banking. In order to uncover repeatable tendencies and systematic correlations among parameters, the system employs consumer transactions, and data-mining methodology, and afterward validates the results. The evaluation Technique for Enhancing Practices in SMEs was created by [10]. The productivity of SMEs is tracked and evaluated using an evaluation technique.

III. PROBLEM STATEMENT

Company sizes are estimated using different parameters. So, it is known to required that which parameters are more accurately estimate the company size. For this problem, we perform the correlation analysis and classification of company size with full attributes of employee profile.

IV. METHODOLOGY

## Dataset

We downloaded the job salaries dataset from Kaggle. The downloaded dataset was based on the 11 features and 607 samples. The features of the dataset were belonged to the numerical and categorical values. The dataset was used to estimate the Company size based on the employee salary, employee profession and employment types. For the classification of the company size, the company size column in the dataset was selected as the target variable. Target column contain the three distinct values including the small, medium, large. In other words, we have the three types of company size that need to be classified. The rest of the features were used as the input variables. The detail description of the dataset is available in below Table 1.

Table 1: Detail Description of Dataset.

|  |  |
| --- | --- |
| Features | Work-year, Experience-Level, employment-type, job-title, salary, salary-currency, salary-in-USD, employee-residence, remote-ratio, company-location |
| Target | Company-Size |
| Classes | Small, Medium, Large |
| Samples | 607 |

## Data Exploratory Analysis

For the better understanding of the dataset, we analyse the dataset from different perspectives. The outcomes the exploratory analysis is presented in this section.

### Class Distribution Analysis

To estimate the weight of each class in company size column, we perform the class distribution analysis. Class distribution analysis count the frequency of each class in targeted column. Further, the count of each class was also plotted to graphical represented of class balancing. The class balance bar chart of employee salaries dataset is presented in below Figure 1.

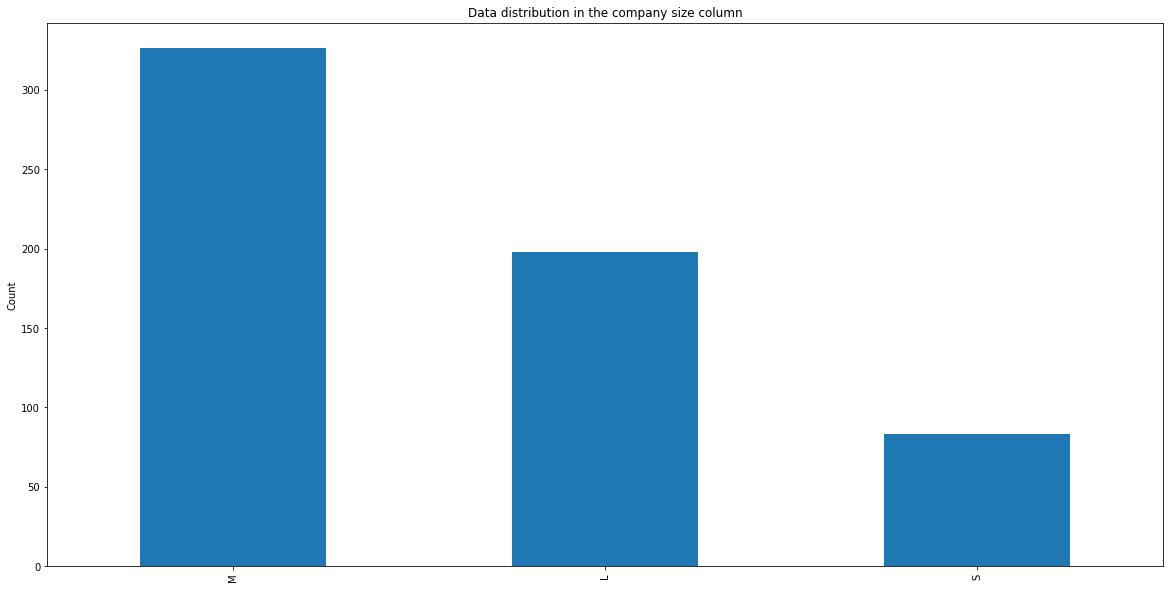


Figure 1: Class Distribution of Dataset.

### Analysis of Avg. Salary in Different Company Sizes

In this analysis, we are eager to know the avg salary of the employees in different sizes of the companies. In other words, this analysis will that how the salaries of the employee are varies with the size of the company. We calculate the avg salary of the employee for each class of target variable. The bar chart of average salary for different company sizes is presented in below Figure 2.

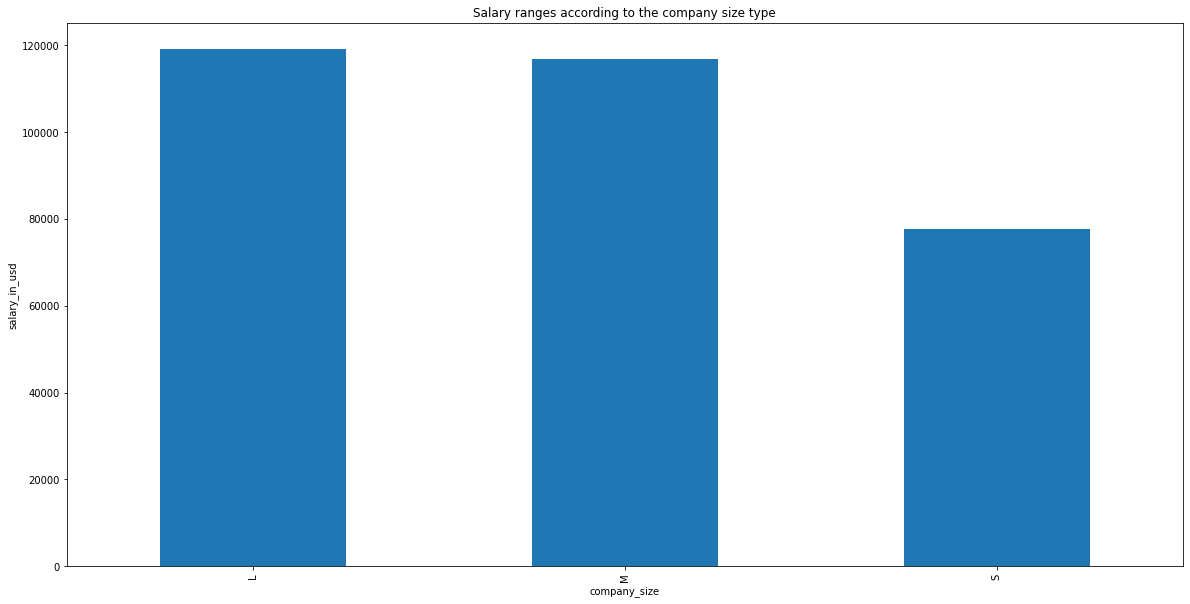


Figure 2: Average Salary of different Company Sizes.

### Analysis of Remote Employees in Different Company Sizes

We also analyse the variance in remote employees for small, medium and large companies. We Calculate the sum of remote employees for each type of company size and plotted in bar chat. The chart of remote employees in different seizes of companies in presented in Figure 3.

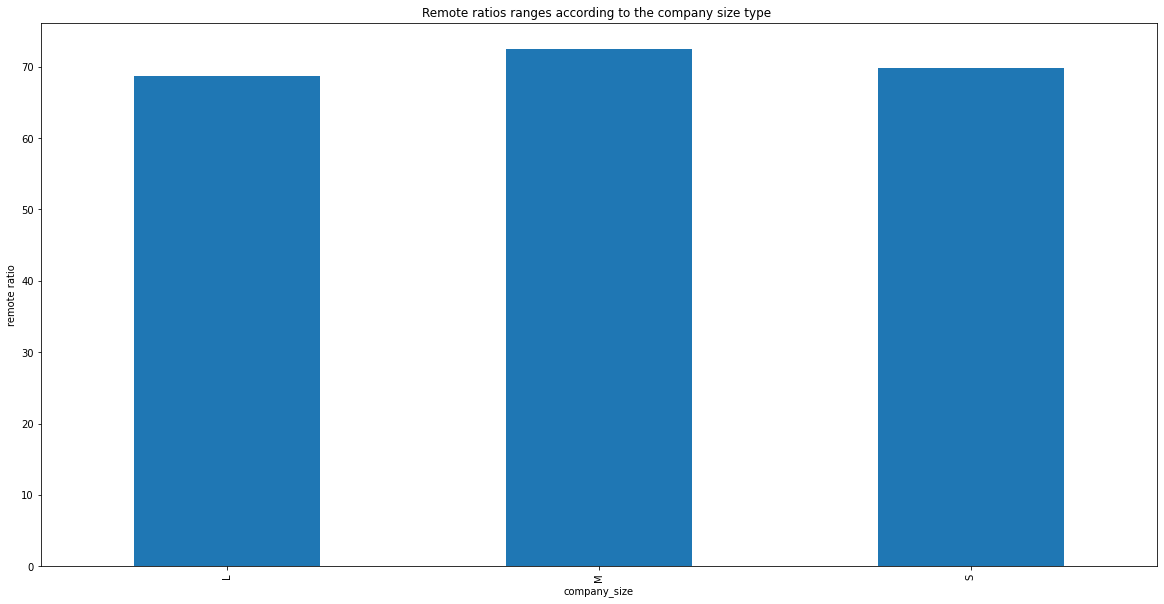


Figure 3: Analysis of Remote Employees.

### Correlation Analysis

For the effect of independent variables on the target column, correlation analysis was performed. It evaluates the effect of all column on the company size column. Salary currency, job title, and number of remote employees positively impact the company size while rest of the features negatively impact on company size. The value of each feature is presented in below Figure 4.

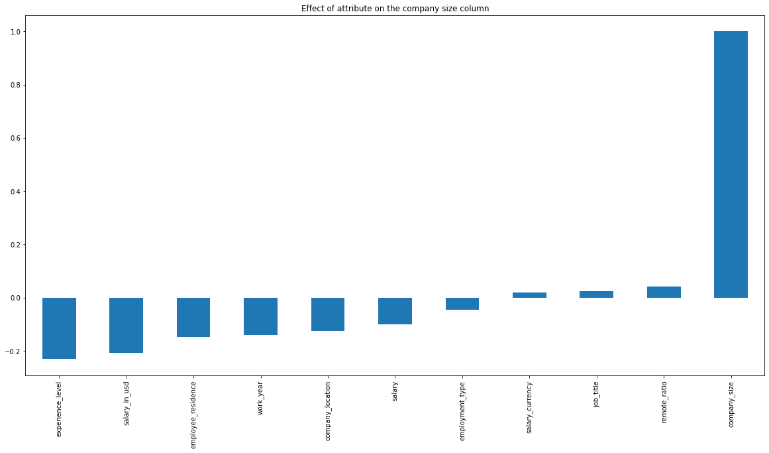


Figure 4: Correlation analysis with target Variable.

## Preprocessing

The first stage in preparing a dataset for machine learning and deep learning models is to clean truncate and transform the dataset. In this context, we employ a variety of data pre-treatment approaches, including data cleaning, handle null values and label encoding.

For the cleaning of the data, firstly we identify the null value and remove the samples with null values. Later the outliers and NAN values were identified and replace by the mean value of that column. As the features of the dataset was based on categorical values that need to be convert in numerical values. Label Encoding is a process of converting the categorical values in the numerical values. We convert the categorical features into numerical representation by using the label encoder. Lastly, we oversampling technique for balancing the dataset.

## Train Test Split

After the pre-processing and feature selection process, dataset need to be split in different subset for the training, testing and validation of the model. We used the built-in train-test-split function of scikit-learn library for dividing the dataset into three different subsets. The built-in train test split function randomly selects the samples from each class with some ratio for one subset and rest of samples for second subset. It did not override the samples into the divided subset. We also used the pre-processed features dataset and split into two different subsets. The extracted features dataset was dividing into training and testing set with the ratio of 70% and 10% respectively. After splitting the dataset, the train and test set contain the 655 and 323 samples.

Table 2: Train Test Sets.

|  |  |  |
| --- | --- | --- |
|  | Training Set | Testing Set |
| Number of Samples | 655 | 323 |
| Number of Features | 10 | 10 |

## Model Development and Training

For the classification of Company Size types with employee salaries data, we used different machine learning and deep learning models. From the machine leaning models, we used the Random Forest, Decision Tree, and K Nearest Neighbor algorithm for the classification of Company Sizes. We did not develop the machine learning model from scratch. We initialize the already developed machine learning models by using the scikit learn library of python. Decision Tree model was initialized with the entropy hyper parameter value of criterion. Random forest was initialized with the max-depth value of 5 and bootstrap value of 100. While the KNN model was initialize with the 5 n-neighbor value. Rest of the hyper parameters of machine learning models were used with their default values.

For the training of the machine learning models, simply the train set was passed to the initialized machine learning models. After the complete training of the models, all the trained models were evaluated using the 323 samples of test set.

## Model Evaluation

For the evaluation of the trained models, we used different evaluation measures including accuracy, precision, recall and f1-score. Recall or sensitivity is the ratio of real positive cases that correctly predict positive with the total real positive cases. Contrarily, precision or confidence refers to the percentage of predicted positive instances that are actually real positives. So, we can mention the recall means “how many samples of particular class you find over the all samples of that class," and the precision will be “how many are correctly classified among that class." The f1-score is the harmonic mean between precision & recall. The test set with 7423 samples was used for the evaluation of the trained model. The evaluation measures were calculated by using the formula of that measure. The equation of calculating the measures is presented in eq 1-4.

Eq. 1

Eq. 2

Eq. 3

Eq. 4

## Environmental setup

We will discuss the development environment, used libraries, programming language and all other related setting for the training of the proposed models. All the experiments were performed by using the python programming language. The python 3.7 version was use to perform the experiments. A Conda environment was established with python version 3.7 for the training of the models. All the essentials library were installed in the developed environment.

From the well-known framework for deep learning models, we used the TensorFlow framework for the development and the training of the proposed models. We installed different libraries in our established environment. The list of all core libraries with their version number is listed in below Table 3.

Table 3: Used libraries for Proposed Study

|  |  |
| --- | --- |
| Library Name | Version |
| TensorFlow | 2.3.0 |
| Matplotlib | 3.5.2 |
| scikit-learn | 1.10.1 |
| Pandas | 1.3.5 |
| Numpy | 1.19.0 |
| Seaborn | 3.6.6 |

V. RESULTS AND FINDINGS

For the classification of the company sizes, we used different machine learning models and a customized neural network deep learning model. In this section, we present the results of the trained models.

## Decision Tree

After the complete training of the Decision Tree model, random forest showed the 0.69% test accuracy on test set. The complete classification report of Decision Tree model on test set in presented in Table 4. The confusion matrix of Decision-Tree model on the 323 samples of test is also presented in Figure 5.

Table 4: Company-Size Classification - DT Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report for Decision-Tree Model** | | | | |
|  | precision | recall | f1-score | support |
| Small | 0.59 | 0.61 | 0.60 | 92 |
| Medium | 0.75 | 0.73 | 0.74 | 130 |
| Large | 0.73 | 0.73 | 0.73 | 101 |
|  |  |  |  |  |
| accuracy |  |  | 0.70 | 323 |
| macro avg | 0.69 | 0.69 | 0.69 | 323 |
| weighted avg | 0.70 | 0.70 | 0.70 | 323 |
| Accuracy (Train Set): 0.7017 Accuracy (Test Set): 0. 6966 | | | | |

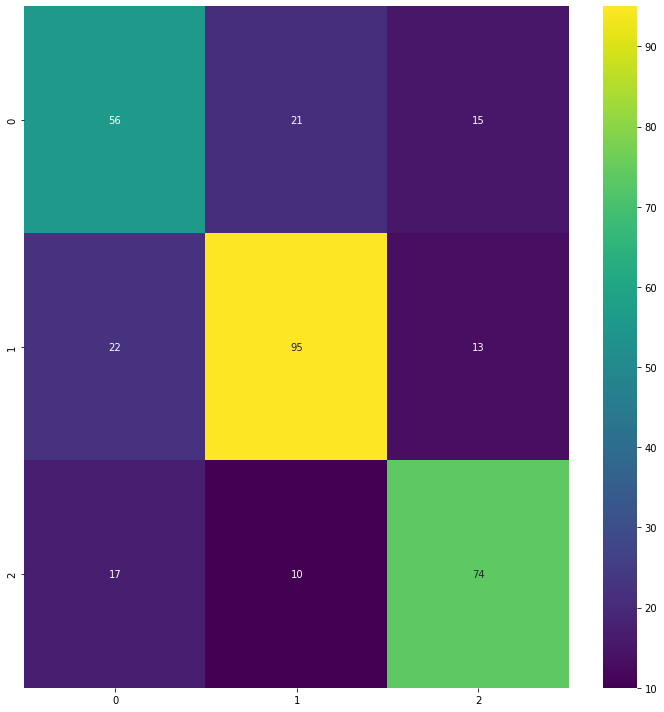


Figure 5: Company-Size Classification - DT Confusion Matrix

## Random Forest

After the complete training of the random forest model, random forest showed the 0.79% test accuracy on test set. The complete classification report of random forest model on test set in presented in Table 5. For the visual understanding of the trained model, the confusion matrix of the model on test set is also presented in Figure 6.

Table 5: Company-Size Classification - RF Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report for RF Model** | | | | |
|  | precision | recall | f1-score | support |
| Small | 0.68 | 0.79 | 0.73 | 92 |
| Medium | 0.87 | 0.75 | 0.80 | 130 |
| Large | 0.85 | 0.87 | 0.86 | 101 |
|  |  |  |  |  |
| accuracy |  |  | 0.80 | 323 |
| macro avg | 0.80 | 0.80 | 0.80 | 323 |
| weighted avg | 0.81 | 0.80 | 0.80 | 323 |
| Accuracy (Train Set): 0.8143 Accuracy (Test Set): 0.7988 | | | | |

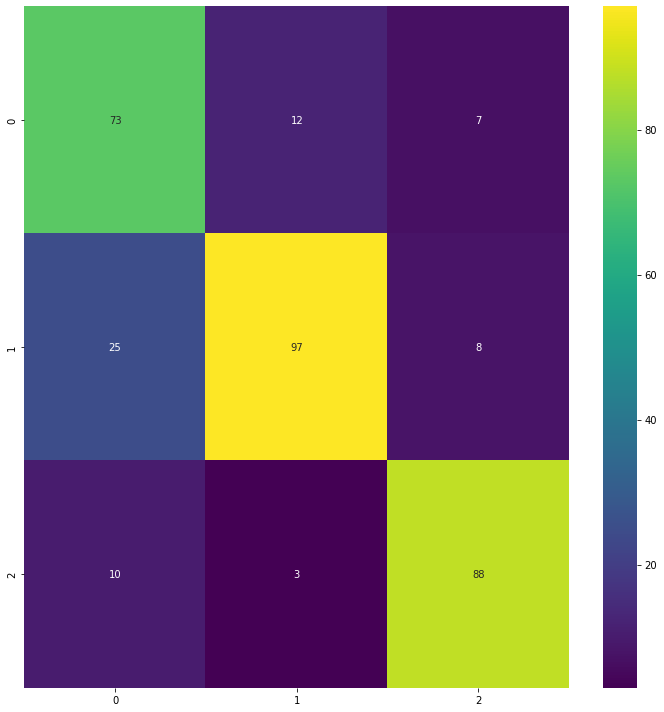


Figure 6: Company-Size Classification - RF Confusion Matrix

## KNN Model

KNN model showed the 0.43% test accuracy on test set after the complete training on train set. The complete classification report of KNN model is shown in Table 6. The confusion matrix of KNN model is also shown in Figure 7.

Table 6: Company-Size Classification - KNN Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report for KNN Model** | | | | |
|  | precision | recall | f1-score | support |
| Small | 0.34 | 0.60 | 0.43 | 92 |
| Medium | 0.44 | 0.27 | 0.33 | 130 |
| Large | 0.64 | 0.51 | 0.57 | 101 |
|  |  |  |  |  |
| accuracy |  |  | 0.44 | 323 |
| macro avg | 0.47 | 0.46 | 0.45 | 323 |
| weighted avg | 0.47 | 0.44 | 44 | 323 |
| Accuracy (Train Set): 0.4467 Accuracy (Test Set): 0.4396 | | | | |

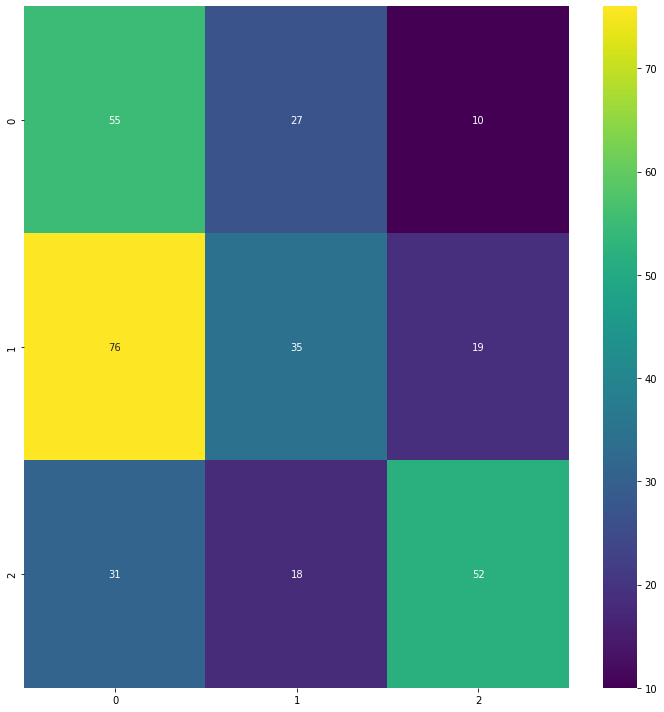


Figure 7: Company-Size Classification - KNN Confusion Matrix

## Comparative Results

Lastly, we perform the comparative study on the results of the trained model. After the training of the proposed model, the results were compiled on test set. The comparative view of all trained models is presented in Table 7.

Table 7: Comparative results of all models.

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | DT | RF | KNN |
| accuracy | 0.6965 | 0.7987 | 0.4396 |
| precision | 0.6896 | 0.7987 | 0.4729 |
| recall | 0.6907 | 0.8036 | 0.4606 |
| f1-score | 0.6900 | 0.7981 | 0.4459 |

Further, the comparison bar chart for all evaluation measure was plotted for the comparison of the results. The comparative bar chart of all trained models is presented in Figure 8.

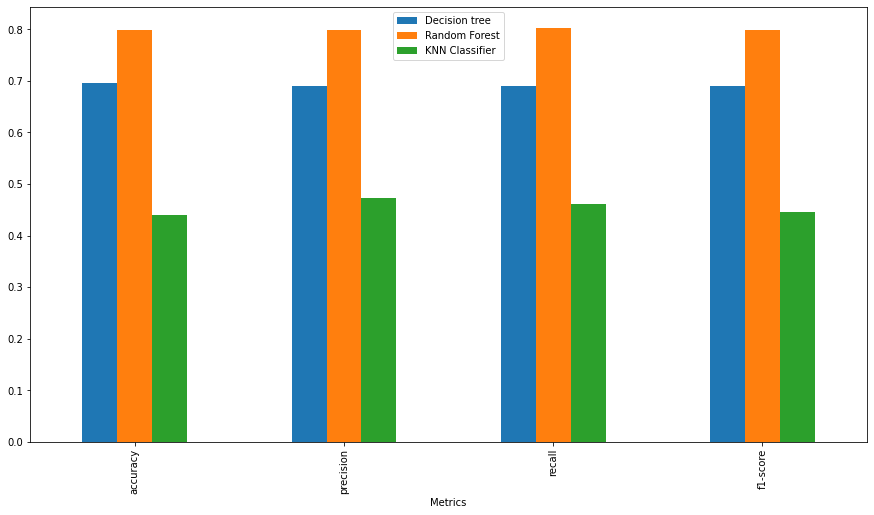


Figure 8: Evaluation Measures Repost - All Models.

VI. CONCLUSION

In this proposed study, the aim objective was to estimate the Company Size using employee salaries data. The more accurate classification of company size will help to deal the new business with company. For the classification of company size, we trained numerous machine learning models. By using the different evaluation measure like accuracy, precision, recall and f1-score, we evaluate all the trained models. As the selected dataset for the proposed study was converted into class balance dataset and the accuracy is the best evaluation measure for class balanced dataset, we select the accuracy as our base evaluation measure. On the basis of all model’s accuracy score, we discover that the Random Forest model is more robust for Company Size estimation related to other machine learning models. Random Forest model showed the 0.79% test accuracy that was the highest accuracy score compare to the other trained models. By analyzing the all-evaluation measures of proposed model, we hypothesized that our model is robust enough to deploy in real world environment for the classification of company size using employee salaries data.

References

[1] Fried, Harold and Tauer, Loren, (2015), An entrepreneur performance index, Journal of Productivity Analysis, 44, issue 1, p. 69-77, https://EconPapers.repec.org/RePEc:kap:jproda:v:44:y:2015:i:1:p:69-77.

[2] Gimeno, J., Folta, T. B., Cooper, A. C., & Woo, C. Y. 1997. Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. Administrative Science Quarterly, 42(750-83).

[3] Borlea, p. S. (2014). Diagnosing of Business and Predictive Model of Bankruptcy Risk- A Model Design.

[4] Borle, A. p., & Monica-Violeta, A. p. (2014). Diagnosing of Business and Predictive Model of Bankruptcy Risk- A Model Design.

[5] Buss, Arnold H. and Finn, Stephen E.(1987) Classification of personality traits. In Journal of Personality and Social Psychology, Vol 52(2), Feb 1987, 432-444

[6] Sharpe, W.F. (1966) Mutual fund performance. The Journal of Business, 39, 119-138. http://dx.doi.org/10.1086/294846

[7] Mensink, T. (2010). An empirical study about early-stage IT and internet venture capitalists in the Netherlands, Singapore and Sweden. Enschede.

[8] Nagar, Y., & Malone, T. W. (2011). Making Business Prediction by Combining Human and Machine Intelligence in Prediction Markets O., O. (2009). Impact of micro finance on enetrepreneurial development.

[9] Islam, M. R., & Habib, A. (2015). Data Mining Approach to Predict Prospective Business Sectors for Lending in Retail Banking Using Decision Tree morelli, t., shearer, c., & buecka, a. (2010). IBM SPSS predictive analytics: Optimizing decision at the point impact.

[10] Alaskari, O., & Ahmad, M. M. (2014). Development of Assessment Methodology for Improving Performance in SME’s.