**HILCOE**

**School of Computer Science and Technology**

**Data Mining Project**

|  |
| --- |
|  |

**Salary Prediction**

**Group Members:**

1. **Yoseph Tadesse - XR2277**
2. **Zinedin Abdulshkur - KZ9441**
3. **Naan Ahmed - TV0776**
4. **Naol Ebissa - HG1141**
5. **Yabsira Bekele - YY2819**

**Submitted To Dr. Eyob Negussie**

**Section DRB1902B**

**Date: January 28, 2023.**

**Table of Contents**

1. Introduction 3

1.1Problem 3

**1.2 Objective 3**

**1.3 Methodology 3**

2. Review related works 4

3. Data Preparation 4

4. Experimental setup 5

**4.1 Mining Methods5**

**4.2 Parameters Used6**

5. Summary of Experimental Results & Findings of the Study 6

6. Conclusion 7

**7. References8**

1. **Introduction**

**1.1 Problem**

The salary of an employee is one of the most important pieces of information for an employer. Salaries are one of the largest expenses for an organization, and the salary of a new hire can have a large impact on the overall budget. In addition, the salary of an employee can have a large impact on a company’s culture.

The first step in this assignment is to review the data and understand the format of the dataset. The second step is to use the dataset to build a model that can predict the salary of a new hire. The goal of this assignment is to predict the salary of a candidate for a job, using data from a large company. We used information from the company’s website to construct a model that can be used to make predictions about the salary of a candidate.

**1.2 Objective**

Objective of the experiment being able predict the salary of a person using the attribute like Age, work-class, education, education-num, marital-status, occupation, relationship, race, sex, hours-per-week, native-country. To predict what his/her attribute will get them and for companies to make a prediction about the salary of a candidate.

**1.3 Methodology**

We collected publicly available salary data using a dataset from the data science business Kaggle The data was then pre-processed before we applied the two classification algorithms by using the training data to generate a model and using test data for testing. The precision, accuracy and recall are used to check the result output of the 2 models. The models are then compared on the basis of their results.

1. Data collection: Used a publicly available data found on Kaggle that was extracted by Barry Becker from the 1994 Census database.

2. Data cleaning: posts with missing values are removed and possible conflicts in the data format are fixed.

3. Manual feature engineering: irrelevant features are discarded by exploiting the domain knowledge.

4. Create A Model: using training data to create model

4. Model comparison: each model is compared to the others with respect to precision and recall like the classification accuracy.

1. **Review of Related Works**

In Tarun Kumara’s and Rajiv Kumara’s (Bajaj form Institute of Technology and Management) data report of their Employee Salary Prediction System they have used the save data parameters to determine the salary prediction. The data split in training and testing part is in the 8:2 ratio. The Logistic Regression Classifier, Gaussian Naïve Bayes, and Support vector machines classifier algorithms use the training data to generate a model and uses test data for the testing. The precision, accuracy, recall and F-1 score are used to check the result output of all 3 models.

All the algorithms above give the following result in the training data: the first Logistic Regression classifier with accuracy of 83.17%, Recall with 0.86, Precision of 0.91, FI with 0.66. The second one, Support Vector classifier with accuracy of 83.0%, Recall 0.86, Precision 0.93, FI 0.68. Third one Native Bayes Classifier with accuracy of 59%, Recall 0.96, Precision 0.47, and FI 0.42

From the observed data they found that the results of Logistic Regression and support vector classifier are nearly same, but the support vector classifier take large amount of time for training. The reason for the bad results of the Gaussian naïve bayes is that it treats each attribute as independent of each other which is practically very difficult to achieve. So, they concluded that the Logistic Regression classifier is better because the training time required in Logistic Regression classifier is much less as compared to support vector classifier.

The other work we saw was made by International Journal of Computational Intelligence Systems . In their data they used Linear models (LM), Logistic regression (LR), K-nearest neighbors (KNN), Multi-layer perceptron’s (MLP), Support vector machines (SVM), Random forests (RF), Adaptive boosting with decision trees (AB), Ensembles of the previous models to determine the salary. Their accuracy results are: LR 58.6%, KNN 79.2%, MLP 59.1%, SVM 66.3%, AB 83.6%, RF 84%, VOTE 84.4%, and VOTE3 83.7%.

1. **Data Preparation**

After we collected the data, the second thing we did was cleaning up our data. When cleaning our data, our initial step was to remove rows that would not help with our categorization; as a result, when we found an empty (missing) cell we removed the row. After that, we removed redundant data, and then trimmed our data 30000 to 1000 because we only need 1000 rows of data to do the salary prediction. Figuring out which data is useful and which is redundant was one of the most difficult aspects of this phase. We split our data by 7/3 ratio.

1. **Experimental setup**

**4.1 Mining method**

* **Gini index (CART Algorithm)**

Classification And Regression Trees (CART) algorithm is a classification algorithm for building a decision tree based on Gini’s impurity index as splitting criterion. CART is a binary tree build by splitting node into two child nodes repeatedly. rules, which can be accustomed build the predictions. The algorithm works repeatedly in three steps:

1. Find each feature’s best split. For each feature with K different values there exist K-1 possible splits. Find the split, which maximizes the splitting criterion. The resulting set of splits contains best splits (one for each feature).

2. Find the node’s best split. Among the best splits from Step 1 find the one, which maximizes the splitting criterion.

3. Split the node using best node split from Step ii and repeat from Step 1 until stopping criterion is satisfied.

Description: Text

Description automatically generated

* **Logistic regression**

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, we used it to tell to predict if an employee will get a money greater than or less than 50k by using the data given.

We used Two mining methods to determine the accuracy form our salary prediction, this are the Gini Index and Logistic Regression where we used the following attribute or parameters to determine which algorism gave a high accuracy rate for our prediction.

**4.2 Parameters used**

Age, work-class, education, education-num, marital-status, occupation, relationship, race, sex, hours-per-week, native-country were used to determine if Gini index gave the better result or if Logistic Regression gave the better result and to determine the salary of the employee if they should be paid >50k or <=50k

1. **Summary of Experimental results & findings of the study**

The result was very close both the algorithm we used on our experiment had a accuracy of 0.7880 ~ 0.79 for the Gini index and 0.82 for the Logistic Regression. Precision for less than equal to 50 thousand is 0.85 and for the grater than 0.54 in the Gini index but for the Logistic Regression the less than equal to 50 thousand is 0.86 and for the grater than 0.59.

Recall for less than equal to 50 thousand is 0.88 and for the grater than 0.48 in the Gini index but for the Logistic Regression the less than equal to 50 thousand is 0.93 and for the grater than 0.39.

FI-score for less than equal to 50 thousand is 0.86 and for the grater than 0.51 in the Gini index but for the Logistic Regression the less than equal to 50 thousand is 0.89 and for the grater than 0.47.

* Confusion matrix for Logistic Regression

|  |  |  |
| --- | --- | --- |
|  | <=50k | >50k |
| <=50K | 186 | 14 |
| >50k | 31 | 20 |

* Confusion matrix for Gini index

|  |  |  |
| --- | --- | --- |
|  | <=50k | >50k |
| <=50K | 205 | 28 |
| >50k | 36 | 33 |

* Summary Result

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | FI score |
| Logistic Regression | 82% | 80% | 82% | 81% |
| Gini index | 79% | 78% | 79% | 78% |

1. **Conclusion**

From the observed data we have found the results of Logistic Regression and Gini index have almost he save result when compered same of the results but there was slight difference that made on better than the other, the accuracy rate of the Logistic Regression have given better result than that of the Gini index whereas precision, recall, fi-score differ making the <=50k have bigger result on Logistic Regression and >50k have better result on than the other.

We concluded that the Logistic Regression give better prediction than the Gini index for our salary prediction algorithm.

1. **References**
2. [Salary Prediction Classification | Kaggle](https://www.kaggle.com/ayessa/salary-prediction-classification)
3. <https://www.google.com/url?sa=t&source=web&rct=j&url=http://www.jetir.org/papers/JETIR2008084.pdf&ved=2ahUKEwitpvmf6IL4AhValf0HHVKrCrcQFnoECAwQAQ&usg=AOvVaw1u0YuGTBBIKe9Y_26dMp5W>
4. <https://www.researchgate.net/publication/327080220_Salary_Prediction_in_the_IT_Job_Market_with_Few_High-Dimensional_Samples_A_Spanish_Case_Study>
5. Full Project Link:

<https://github.com/naanahmed/Data_Mining-Final-Project.git>