Student ID/Registration no: 220147226
Name: Sudip Subedi
Canvas ID: **bi49rb**
Course of Study: BSC(HONS) Computer System Engineering

INTELLIGENT PROTOTYPE DEVELOPMENT

Module code: CET313 Artificial Intelligence

Contents

Introduction

Salary is a regular payment made by a company to an employee for his work and service which he provides for the company from a specified position. Salary is a part of the employment contract and might be changed during the terms and conditions of the Organization. The primary advantage of getting a wage is to plan for the future. This makes choosing how much we can afford, what types of automobiles to buy, where to invest and save for the future so it well be easier for the future generations also. Salary can be vary depending on our work and position. If a person works for the government in a specific role, he may be paid less than someone who works for a private corporation. But Government jobs provide more job security than those who work in private corporations. Employees continually change jobs to earn their promised salary. And this results in the company's loss. Everyone has higher expectations and goals in this competitive world. However, we cannot provide every individual with their expected salary. Depending on what has already been learned and investigation a simple to implement protype has been chosen. The chosen protype is very easy to understand.

Mission Statement

The main aim of this study is to predict an employee salary on the basis of their job and education based on their year of experience. In this project I am going to use Logistic regression, Support Vector Regression, K-Nearest Neighbors and Random Forest algorithm in machine learning whether a person makes over 50 thousand or not. In order to do this, I have downloaded the data set from the Kaggle website which is the best option for this model. For the coding part I have using the Jupyter Software.

Over view of e-Portfolio

Many topics have been learned during the semester to convey knowledge about the broad concept of Artificial intelligence, in additional to provide student with all of the necessary basic skills. The expertise and skills gained allow to firmly choose the type of protype they want to do. A digital portfolio was on of the tools available for us to learn about the machine learning throughout the class.

With the extended exercises and research topics provided through canvus, we managed to gather as groups at some point to solve and implement python programs and codes. We were able to understand more about artificial intelligence through group discussion and research. for example types of machine learning, types of clustering algorithm and search algorithm. It was introduced to us for the first time. I find it flexible to use Jupyter

Different algorithms were taught. Natural language processing, planning how to deal with ai, imperfect Knowledge and Logic. Overview to Machine learning and its types supervised and Unsupervised machine learning and Clustering furthermore Artificial Neural Networks which is used to find the fraud detection. The task excises on the e-portfolio were handover with Jupyter. Python is the language required for prototyping and it includes necessary libraries that include pandas, Numpy , seaborn and matplotlib and many others

The following link is the link where my e-portfolio files are found:

https://canvas.sunderland.ac.uk/eportfolios/12576?verifier=EVEUBuuADKL0ZxUbRS8FsmHSsmdfSw7iw73kJHOQ

Identification and Planning of Prototypes

Literature review of protype identification

(Das, Barik and Mukherjee, 2020) The work in their paper is focusing on Salary prediction using Regression model in machine learning with specific implementation of Linear Regression and Polynomial Regression. The methods in this paper includes the screening of datasets from the different organization and makes a graph through this information form the datasets. The methods in this paper include the segments of study Machine learning, Linear Regression, Polynomial Regression, curve fitting method etc. This paper uses relevant data such as years of experience, education level, and job role to build the model. The outcomes of the study highlight the significance of features and model evaluation in generating accurate prediction.

According to (Mukherjee and Satyasaivani, n.d.2022) salary is the main reason whether to stay in the company or not. If they won't get desired salary, every employee is ready to change the company. the main aim of this paper is to predict the Salary of the employee using Machine learning classification model with Supervised Learning and Linear Regression Algorithm. Firstly, the author has collected the dataset through different open-source platform. Following that, they do processing of data such as EDA to ensure that there is no any null value exist. Further more, they use the data visualization like scatterplot diagram to display the data in two attribute data year of experience and wage after which they sperate the data into the dependent and independent variable before splitting the data into training and test data. Furthermore they test the data before visualizing the expected and actual data.

(Chen, Mao and Yuan, 2022) The work in this paper is focusing on Salary prediction using classification model in machine learning using random forest with fundamental. The methods in this paper incudes the data from UCI Machine learning Repository with the sample size of 32561. The methods in this paper includes the segments of Data processing, Missing data, Data analysis, One-hot encoding and model construction. In this paper they have use 5 algorithm Decision Tree, KNN, Logistic Regression, Naïve Bayes and Random Forest to find the overall accuracy and Random Forest has the highest Number of accuracies among all the models. The limitations or research gaps in the research manuscript is focusing on improving these models by using a more mathematical optimization process and reduction to transform data into a low dimensional space. Furthermore, similar to the RF model, they examine developing successful models on nonlinear regression utilizing a neural network with an input layer in future study.

(Yadav and Kumar, 2021) The goal of this paper is to predict salary of a person using Regression Techniques. This paper methodologies include data form salary system from the different organization and makes a graph through this database. Different Experiments have been implemented through this predication (NumPy, matplotlib. pyplot, pandas, Linear Regression and Polynomial features) This application may be seen as a graphical representation and can be forecast any point from position and calculates it automatically. Many survey operators aim to collect the most important input data in every possible way. The first option may ensure major corporations, whilst the second option is mostly for comparable smaller corporations. The limitations or research gaps in this research manuscripts is focusing on improving move accuracy by implementing k-nearest regression. this model can also be utilized for a variety of AI projects and applications.

Another similar research by (Chen et al., 2023) to predict salary Based on the Dual-Adaboosting system in machine learning with implementation of Regression, Ensemble learning, Random Forest and features Filtering model. This paper has practical reference significance for human resources management and the improvement of Adaboosting algorithms. The methods in this paper includes the data from US data science field with the sample size of 742 observation. Adaboosting regression is used for combing numerous weak regressor. The training sample weights are modified based on predication outcomes, and the predications errors of each weak regressor are weighted during the combination phase. In this paper they present and apply the novel optimization technique Dual-Adaboosting system to salary predication. This method effectively reduces salary predication mistake and has significant application prospects. The research gap in this paper is running time for algorithms is longer than that of typical algorithms.

Reflection on prototype identification

Base on the review supervised machine learning system is very good in forecasting salary. Despite multiple algorithms outlined above for example Random Forest were valuable in predicting salary because it can be used in both classification and regression model. The method provides accurate results after the procedure. Random forest was the best model for me to develop this protype. Further research on this algorithm revealed that it can be applied to solve issue like to fraud detection, stock market forecasting and medical diagnosis.

From the fundamental lesson that explained the way to operate the Jupyter to the advanced lesson that described how to import all the required libraries. The instructional materials all played an important role in the production of the produced protype. Machine learning algorithms task was the main example that set the path. For example, correct data in order for obtaining the results accurately. As the features of solving problems, we can assume the size of dataset will also be the factor with regard to algorithm. We can also say that standardizing and cleaning of dataset is also important. If dataset is containing any null values, it should be grouped for improving accuracy of the result.

Protype Development

Data Munging (Process of data cleaning)

The first step is to load the dataset into the application so that operations can be carried out. To load the data first we have to upload the data set and mount the data by using **pd.read_csv("**

data.csv ") then we will check the number of column and rows. To display the rows and column df.shape is used. Monitoring for data that is lacking in a dataset is critical since processing the data will result in the loss of some column attributes.

Data Analysis/Visualization

Following data cleaning the next stage is data analysis. The process of visualizing data to reveal patterns, trends and correlations between all variables is Data visualization. Pair Plot, heatmap were used to show. Heatmap pair plot will indicate the strength and direction of the correlations and the numbers within the cells will provide specific correlation values for each pair of variables in the Data Frame.

Training/ Modelling

Now we are going to use various models on Training set and observer which one produce the highest accuracy. Models are Random Forest, K-nearest Neighbors, Logistic Regression and Svc algorithm. We will demonstrate and share screenshots of the outcomes from the models tested for training in a separate file. **I have attached a pdf file of code output with the zip of prototype and also, I have provided the link of code output .**

https://canvas.sunderland.ac.uk/courses/67753/assignments/118874/submissions/135765?download=11803107

Evaluation

Testing and evaluating the protype are two critical part of application development. evaluating is a way of determining whether an application meets the expected requirement while evaluation involves deciding on a programmed based on its performance.

All the necessary libraries were imported and then the dataset has been mounted in Jupyter. Based on the outcomes Random Forest achieved a high accuracy rate then other. The procedure begins with identifying the problems and determining the available solutions. The paper offered by previous researchers in tackling the same problem prompted me to assess and choose the best method for my protype. The literature study acquired with various writers was critical in determining the proper algorithm to solve the wage predication.

Data manipulation takes performed prior to algorithm testing. Data manipulation is the process of removing unwanted data form a dataset before conducting any operations on it. This approach significantly helped in data sorting for easier training and predication. It includes operations such as validating null values, analyzing data and exposing pattern correlations utilizing various graphic presentations.

Random forests accuracy was 0.85. the matrix of confusion was used to test the model’s accuracy and the outcomes indicated that the model had an accuracy rate of 0.85. The random forest had an accuracy of testing of 0.85. All of these results indicate that the protype constructed was effective. It was able to read, visualization, training and testing data.

Testing Approach

There are several ways for testing that may be employed. There are two sorts of testing approaches: black box and white box. Black box testing is a method for testing without understanding of the application’s internal workings, while white box testing is a deep analysis of the logic and structure of code. The black box testing method was used to test the application.

Black box testing

Advantage	Disadvantage
Accessing the code is not needed	Ineffective checking due to the users lack of understanding of the app
Suitable for lengthy code sections	Only certain test are running due to the limited coverage.

The users viewpoint is firmly separated from the developers perspective via explicitly defined responsibilities.	Blind coverage occurs when the tester is unable to target specific code-segments or error-prone locations.

Testing case design plane

- Test number
- Description
- Expected results
- Actual results

Software used for the development of the protype is Jupyter.

Hardware

- Acer laptop, win10(operating system), 16gb ram and 500gb SSD

Test Report

Salary predication				
Test No.	**Illustration**	**Results we expect**	**Expected unsuccessful results**	**Final outcomes**
1	Mount the csv dataset	All the data are shown	All the data are not shown	The data include in the csv file is shown.
2	Checking the null values	No null values is shown in the dataset	Failed to show the null value	No null value is shown
3	Data visualize	The data is represented in serval diagram (Heatmap and pair plot)	The data is not visualized	The data is visualized
4	Measuring the outlier	Successfully deleted the outlier	Failed to delete the outlier	Outlier has been deleted

5	Feature engineering	Successfully performed the features engineering on test data	Missing Values, Scaling, Categorical Variables and Data Leakage	Successfully performed the features engineering
6	Visualize the correlation	A heatmap display the correlation of each column	The relation of each column is not shown	The relation of each column is shown
7	selection of models	Best model has been selected	Failed to display the accuracy of the model	Best model has been selected.

Conclusion

Based on the sources that are provided on the review papers it may be possible to say that the controlled machine learning has a long range of application in estimating salary of the employee. We have to follow the application of many ideas through multiple sources, procedures and models are performed correctly for the execution of prototype.

Random forest model proved to be an excellent option for the protype. Random forest is one of the best algorithms because it can be used it both classification and regression model. Random forest model produced the best results with an accuracy rate of 85%. For training the algorithm I have to choose the classification pathway, it can also generate accurate and good result. After studying deeper into the random forest algorithm, I understood that it may be used in variety other protype such as health-related situation and many more.

The theory given during the semester was very much beneficial for understanding the (AI) and how it works. Artificial intelligence (AI) holds the future of machine learning. All of the expertise, which includes the research and knowledge of exercise, aided in remaking the type of protype to construct. The prototype chosen was developed and it was a success, it was able to achieve what it was required to do.

The protype development process brought doing things were clearly not discussed in class or in depth. It was both hard and exciting because a concept should be understood before that can be applied. This shows the need of research in all the aspects of development. Whether, the deployment of all algorithm and all models were not easy as some produced shows poor results. As an example, consider the DT model is very much difficult to categories data into accurate classifications, which results in very much low accuracy percentage.

The overall performance of prototype was strong with a high accuracy rate in identifying salary predication. Future developments would have focus on leveraging a great real-world dataset for the train and test components. If possible, an evaluation should be done in near future must be carried out in which human radiologist is pitted against an artificial intelligence algorithm. This is to determine whether a human or machine is more accurate. If there was more time, then a user interface would have been created.

References

Das, S., Barik, R. and Mukherjee, A. (2020). Salary Prediction Using Regression Techniques. *SSRN Electronic Journal*. Doi : https://doi.org/10.2139/ssrn.3526707.

Mukherjee, T. and Satyasaivani, B. (n.d.). Employee's Salary Prediction. *International Journal of Advance Research, Ideas and Innovations in Technology Impact*, [online] 8, pp.3–8. Available at: https://www.ijariit.com/manuscripts/v8i3/V8I3-1357.pdf.

Chen, J., Mao, S. and Yuan, Q. (2022). Salary prediction using random forest with fundamental features. *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*. [online] Doi: https://doi.org/10.1117/12.2628520.

Yadav, P.K. and Kumar, R. (2021). Salary Prediction Using Regression Techniques. *SSRN Electronic Journal*. Doi: https://doi.org/10.2139/ssrn.3990877.

Chen, Y., Zhan, K., Lin, S.-W. and Yao, K. (2023). Salary Prediction Based on the Dual-Ad boosting System. Doi: https://doi.org/10.4108/eai.26-5-2023.2334428.