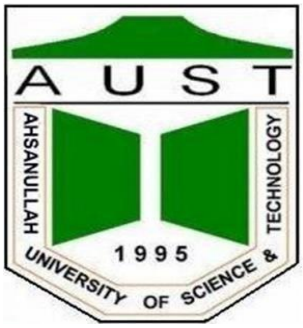
Ahsanullah University of Science and Technology



Department of Computer Science and Engineering

Program: Bachelor of Science in Computer Science and Engineering

Course No: CSE 4108

Course Title: Artificial Intelligence Lab

Project Report

On

Salary Prediction Using Machine learning language

Submitted to:

Mr. Md. Siam Ansary
Lecturer, Department of CSE, AUST.

Ms. Tamanna Tabassum
Lecturer, Department of CSE, AUST.

Group no : B1_01

Submitted by:

Name: Jesmin Akter
ID: 18.01.04.052

Name: Aysha Shiddika
ID: 18.01.04.054

Name: MD Shafique
ID: 18.01.04.059

- ## The description of the problem we worked on

  The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period. To predict salary we use in our project is regression models. By which we found mean absolute error, mean square error, root mean square error and r-squared error.

- ## A brief description of the dataset

  In our dataset , we take over 300 data. The features of our dataset are Gender, CGPA, Degree, Graduation, Year, Coding Skill, Problem solving skill, communication skill, Research, Years of experience, Salary. We create our dataset through google form. First, We take string type data for degree and gender. Then we encoded through label encoder to encode string to numeric. And we took numeric type data for other columns.

- ## Description of the used ML models

  1. **Multiple Linear Regression :** Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*.

  2. Logistic Regression : Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

  3. Polynomial Regression : Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance.

  4. Lasso Regression : Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

  5. Random Forest Regression : Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

6. Support vector Regression : Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

- Comparison of models performance

Accuracy	Multiple Linear Regression
MAE	9250.587279295962
MSE	125727471.3353709
RMSE	11212.826197501276
R-SQUARE	0.1752905066878534

Accuracy	Logistic Regression
MAE	11506.369426751593
MSE	196800955.41401273
RMSE	14028.576385863704
R-SQUARE	0.09554140127388536

Accuracy	Polynomial Regression
MAE	437370661968730.25
MSE	1.1812385304816415e+31
RMSE	3436915085482388.5
R-SQUARE	0.1752905066878534

Accuracy	Random Forest Regression
MAE	1382.708641569401
MSE	8498935.91621345
RMSE	2915.293452847148
R-SQUARE	0.9400320529505105

Accuracy	Lasso Regression
MAE	6899.365021449159
MSE	80438542.17290315
RMSE	8968.753657722078
R-SQUARE	0.490000

Accuracy	Support vector Regression
MAE	9702.290202196218
MSE	166561963.93233305
RMSE	12905.888730821021
R-SQUARE	-0.17525054101164161

After implementing 6 different regression models on this dataset, we have come to understand that , for predicting the salary of a person on the basis of 10 different features the best regression model is “Random Forest Regression Model” which gives an accuracy of 93.21% and the worst model for this dataset is “Support Vector Regression Model” which gives an accuracy of model does not follow the trend of the data, so fits worse than a horizontal line.

- ## Discussion/Conclusion

  Machine Learning is an application of artificial intelligence that provides systems the ability to improve from experience. Machine learning algorithms are often categorized as supervised or unsupervised. In regression analysis, curve fitting is the process of specifying a model that provides the best fit to the specific curve in the dataset. There are many regression techniques such as linear regression, polynomial regression, support vector regression etc.

- ## Percentage Of Contribution

  Jesmin Akter - 33%
  Aysha Shiddika Nuha - 34%
  Mohammad Shafique - 33%