A picture containing text, clock, sign

Description automatically generated

**ADTA 5340 – DISCOVERY AND LEARNING WITH BIG DATA**

**FINAL PROJECT EXECUTIVE SUMMARY**

**MEDIAN ANNUAL WAGE PREDICTION**

**BY**

**Vamshi Krishna Vallam**

**Introduction**

The use case I have used is to predict the median annual salary of different occupations in the IT industry. The main agenda of this project is to find out the median annual pay for various jobs based on several factors including employment, occupational opportunities, typical entry-level education, work experience in a similar occupation, and typical on-the-job training. Businesses can utilize this data to prepare in advance for hiring needs and hiring initiatives.

**Explanation of the Problem**

The problem being addressed here is to predict the median annual salary of various occupations in the IT industry. Coming to an example to explain this problem, we have many levels of jobs in the IT industry with different pay levels for each one. So, the employer/stakeholders need to attract the employees and retain the best talent with a better salary. This is a positive scenario. When coming to the negative scenario, if the salary for an employee is high which is not needed, then the employer cannot compete in the market and may need to cut some of the resources.

**Approach**

I have used a data science approach to work on the problem statement. Once I had the problem statement, I identified the data and the relevant techniques that would help in solving the problem using business understanding. The next step is data understanding where the collected data is assessed based on its quality. The features present in the data and how can these features be used in model evaluation. I then performed Exploratory Data Analysis to identify the trends and patterns in the data which helped us in gaining insights from the data. After identifying the trends in the data, I cleaned and preprocessed the data so that it is suitable for modeling. I have removed the null values, imputed dummy variables for the categorical features, selected the most important variables for modeling, and split the data into training and testing sets. Once the data is ready, I have developed two models using the processed data. The first model selected is Linear Regression and the second model selected was Decision Tree Regressor. The selection of the model depended upon the type of problem I was analyzing and the objectives of the project. After the models were developed, I evaluated them based on their performance. I have validated the model using the test data to ensure how useful it can be in predicting the new data. And finally, I have made predictions on new and unseen data.

**Outcome**

I have successfully developed two machine-learning models that were able to predict the median annual wage for different occupations. The decision tree model provided better predictions and the R-squared value obtained was 0.79 while compared to the linear regression R-squared value of 0.71. Apart from different occupations, I have predicted the expected median annual wage for the 3 of my friends who will be graduating this year.

**References**

* Das, S., Barik, R., & Mukherjee, A. (2020, February 04). *Salary Prediction Using Regression Techniques*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3526707
* ‌*Employment Projections*. (2018). Bls.gov. <https://data.bls.gov/projections/occupationProj>
* Tee, Zhen & Raheem, Mafas. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. 70-74.