Performance of Decision Tree Algorithm in Classification of Graduate Salary

Aman Hussain*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: aman.hussain2015@vit.ac.in*

***Abstract*—**

***Index Terms*—Decision Tree, Supervised Learning, Breast Cancer, Employability**

1. Introduction

Decision Tree algorithms are a class of supervised learning approaches that are widely popular due to its simplicity, efficiency and ease of training. It is easily comprehensible to layman and domain experts alike. The model can be understood by studying its' tree structure. Furthermore, the predictions made by the model can be manually arrived at by following the tree structure of the model. This makes it very practical for engineering and business purposes.

The performance of the Decision Tree model is evaluated on two very different datasets - Breast Cancer Wisconsin (Diagnostic) and Aspiring Minds' Employability Outcomes 2015. The Breast Cancer dataset aims to classify a cancerous tissue as malignant or benign given thirty numeric attributes computed from a digitized image of a fine needle aspirate of a breast mass. The Aspiring Minds' Employability Outcomes 2015 dataset aims to predict the salary of an engineering graduate fresh out of college given the candidates' academic background, standardized test performance and personality test scores.

2. Methodology

The Aspiring Minds' Employability Outcomes 2015 dataset is collected from the Aspiring Minds official website. The data is cleaned of missing values and anomalous data points. The attributes are made numeric and categorical for the computation of split points by the decision tree algorithm.

The Breast Cancer Wisconsin (Diagnostic) dataset is collected from the UCI machine learing repository. The dataset is already cleaned, preprocessed and ready for any machine learning tasks.

3. Dataset - AMEO 2015

For every engineer, AMEO dataset provides anonymised bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer: AMEO

Figure 1. 10th Percentage Marks Distribution

2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis & Results

Here, dataset is visualized to infer the patterns and trends. Hypothesis are formed by asking questions that reveal the nature of the data.

Candidates in the dataset can be seen performing worst as they move up the education ladder. The distribution of marks in 10th year was skewed towards the right; most students scored better than their peers. The distribution of marks in 12th year moved to a bell curve; most students performed average as compared to their peers. The distribution of college grades is slightly skewed towards the left; most students were performing worse than their peers. And finally, the distribution of salary is heavily skewed towards the left.

5. Conclusion

Hence, we have discovered some interesting insights into a typical engineering graduate and his/her job prospects. One of the curious observations is the large overlap of data points which will prove to be a challege during the model development phase. The analytics can also suggest graduates methods of improving their job prospects.