DSCI-5240

DATA MINING

**PROJECT REPORT**

ON

**HUMAN RESOURCES ANALYTICS OF IBM**

BY

SRUTHI ANNAMRAJU

11372707

MASTER OF SCIENCE IN DATA SCIENCE

UNIVERSITY OF NORTH TEXAS

**EXECUTIVE SUMMARY**

My objective was to analyze various factors of an employee like employee job level, attrition, performance rating etc. that can lead us to compare and predict monthly income. The end-goal is to build a more complete picture of how and to classify them in a variety of meaningful ways to appropriately advise accurate incomes based on the employee decisions in the future. Through my efforts I have identified a number of actionable items that may require further exploration.

**PROJECT MOTIVATION**

The main motive behind choosing this data set based on Human Resources Analytics is to know how exactly an employee will be evaluated and given appropriate income for his/her work. This project deals with HR Analytics Data provided by the IBM.

**DATA DESCRIPTION**

This is a fictional data set created by IBM data scientists with 1471 data observations.

The given attributes uncover the factors that lead to employee monthly income and explore important question such as ‘compare monthly income by education or attrition or any other variables. The following are the attributes used in the project:

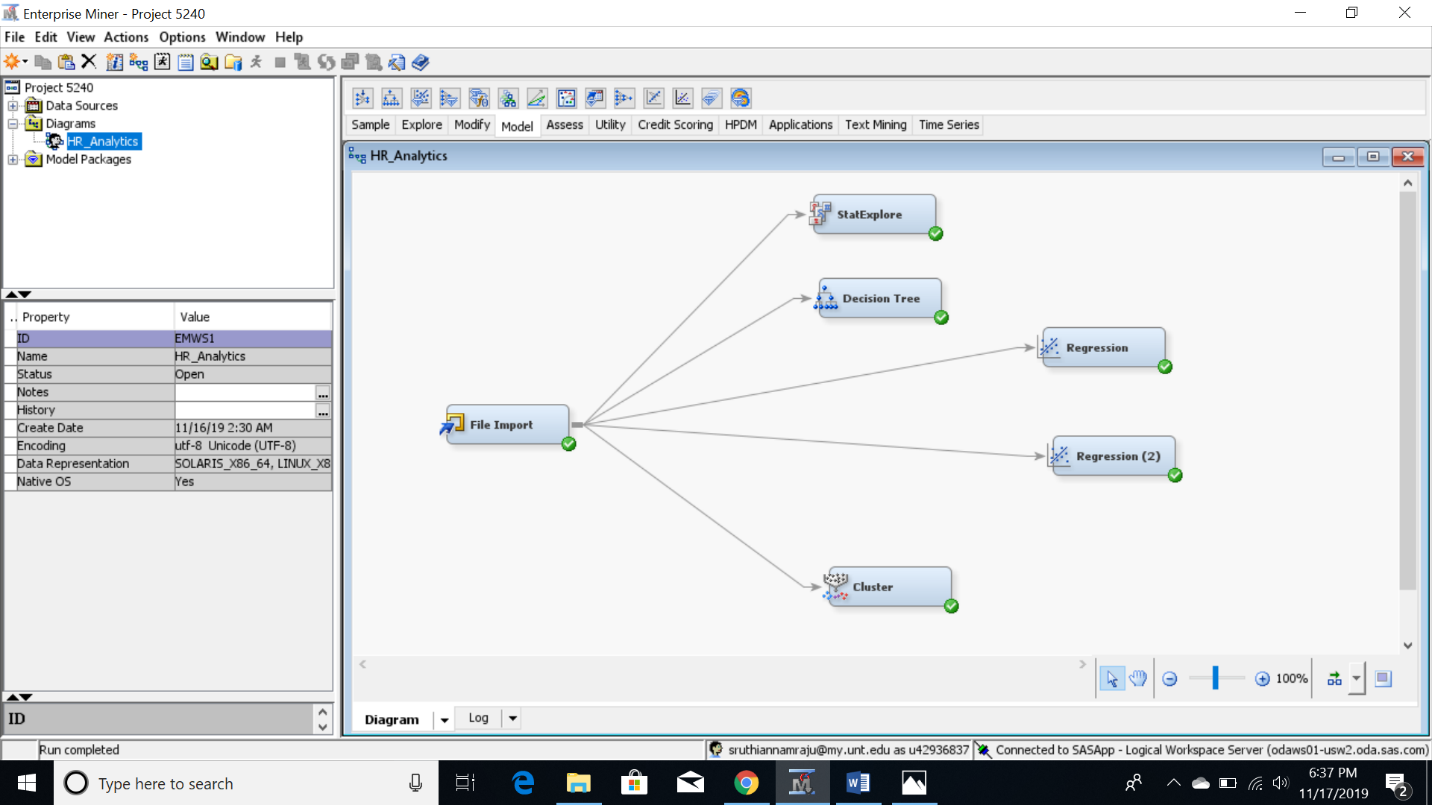
|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLE NAME** | **DESCRIPTION** | **ROLE** | **LEVEL** |
| Attrition | Reduction in number of Employees | INPUT | NOMINAL |
| Department | Department in which employee work | INPUT | NOMINAL |
| Education | Education level of an employee | INPUT | INTERVAL |
| EducationField | major subject in education | INPUT | INTERVAL |
| EnvironmentSatisfaction | employee satisfaction with surroundings | INPUT | INTERVAL |
| HourlyRate | hourly based income | INPUT | INTERVAL |
| JobInvolvement | employee involvement | INPUT | INTERVAL |
| JobLevel | Level of job in which employee is there | INPUT | INTERVAL |
| JobRole | Role description | INPUT | INTERVAL |
| JobSatisfaction | Satisfaction of job | INPUT | INTERVAL |
| MonthlyIncome | Income of employee every month | TARGET | INTERVAL |
| NumCompaniesWorked | Previously worked companies | INPUT | INTERVAL |
| OverTime | time spent over standard hours | INPUT | NOMINAL |
| PercentSalaryHike | percentage increase in salary | INPUT | INTERVAL |
| PerformanceRating | low, good, excellent, outstanding | INPUT | INTERVAL |
| TotalWorkingYears | No. of years associated with the company | INPUT | INTERVAL |
| WorkLifeBalance | Balancing life and work | INPUT | INTERVAL |
| YearsAtCompany | No. of years associated with the company | INPUT | INTERVAL |
| YearsInCurrentRole | No. of years associated with the company with current role | INPUT | INTERVAL |
| YearsSinceLastPromotion | No. of years since promotion | INPUT | INTERVAL |
| YearsWithCurrManager | No. of years associated with current manager | INPUT | INTERVAL |

**DATA PREPARATION**

The initial raw data-set was largely left unaltered as the nature of our investigation is largely based on categorization and not on numerical trends. Data variables such as distance from home, Gender, Relationship status were removed in order to negate their influence on the final results.

**DATA MINING MODELS/DIAGRAMS:**

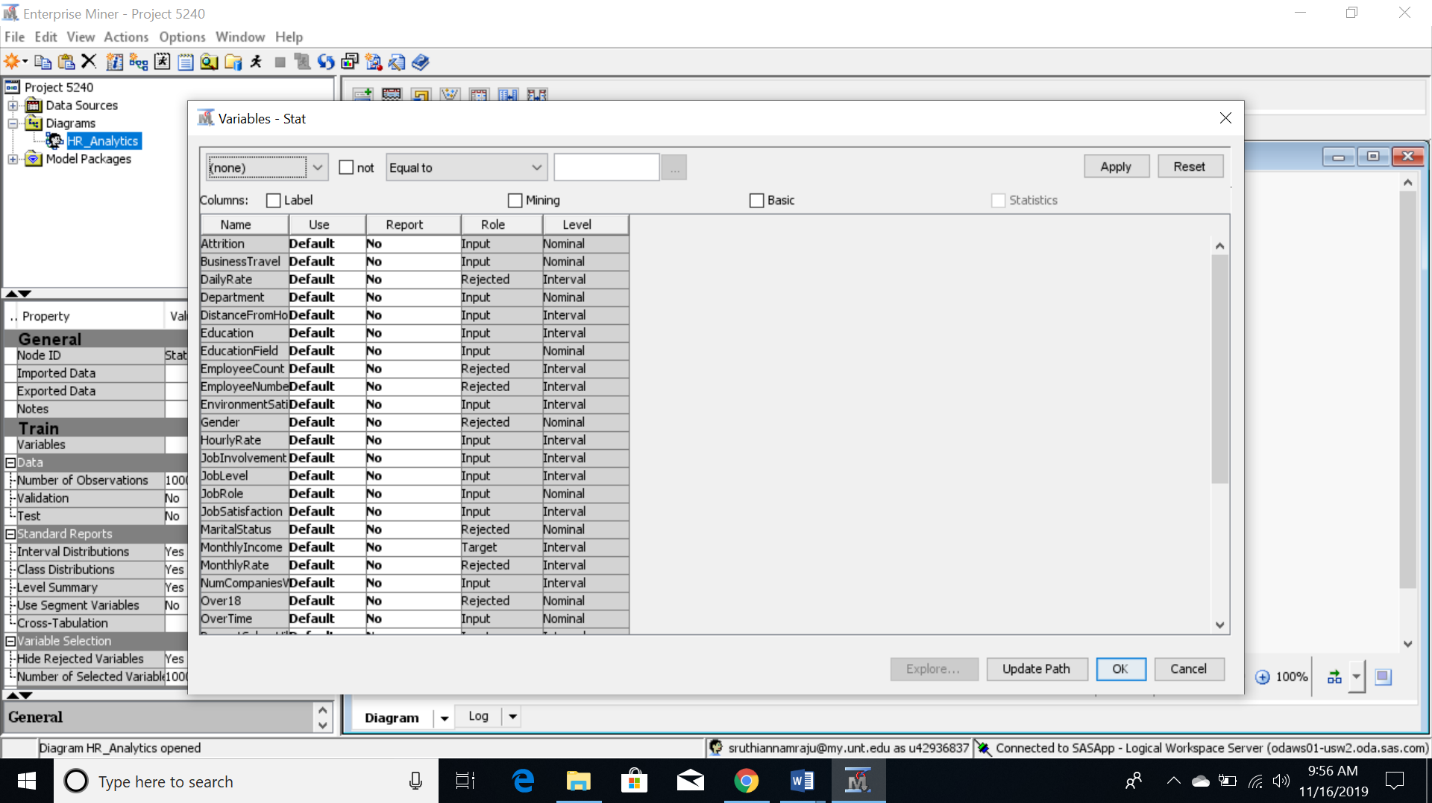
The Final diagram of the project :

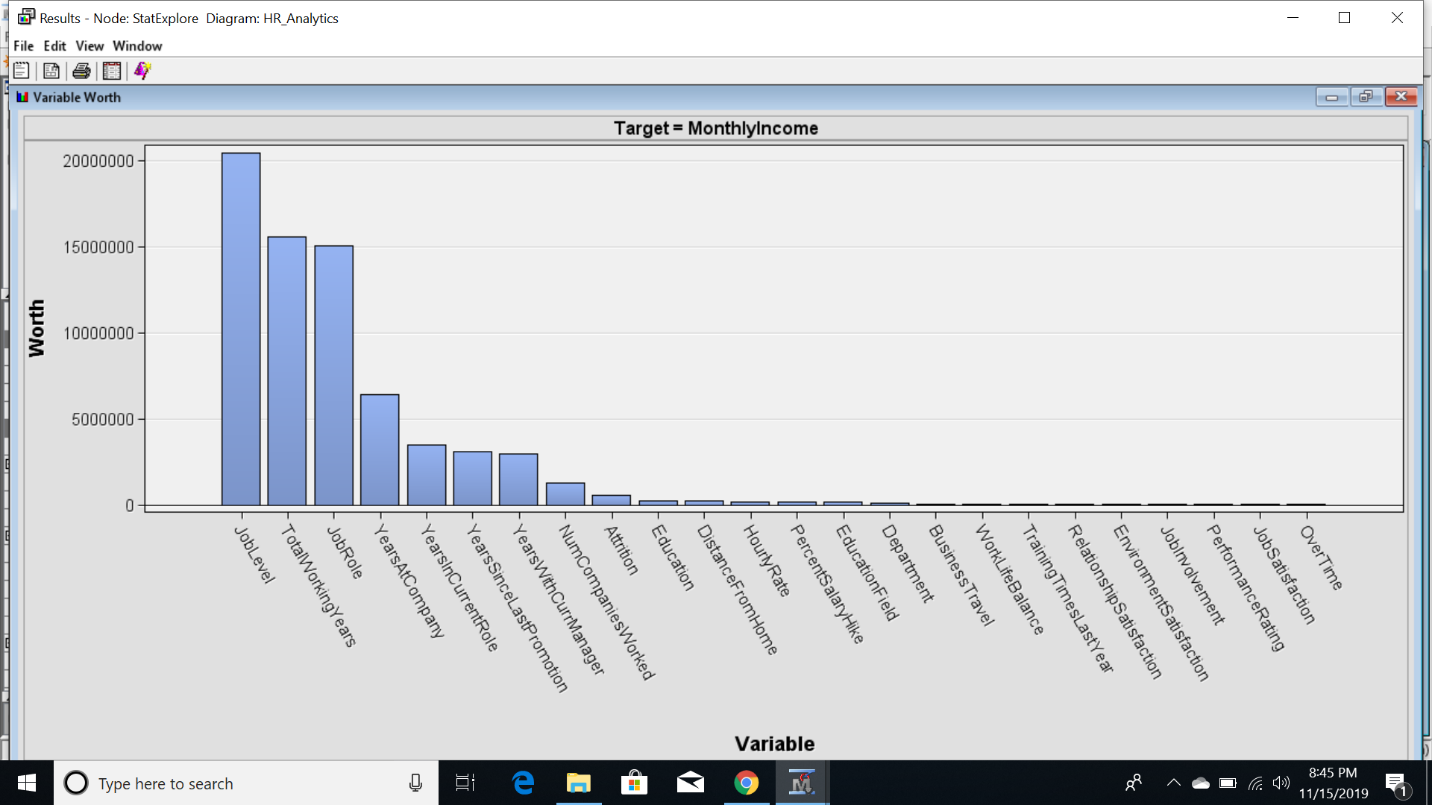


**StatExplore**

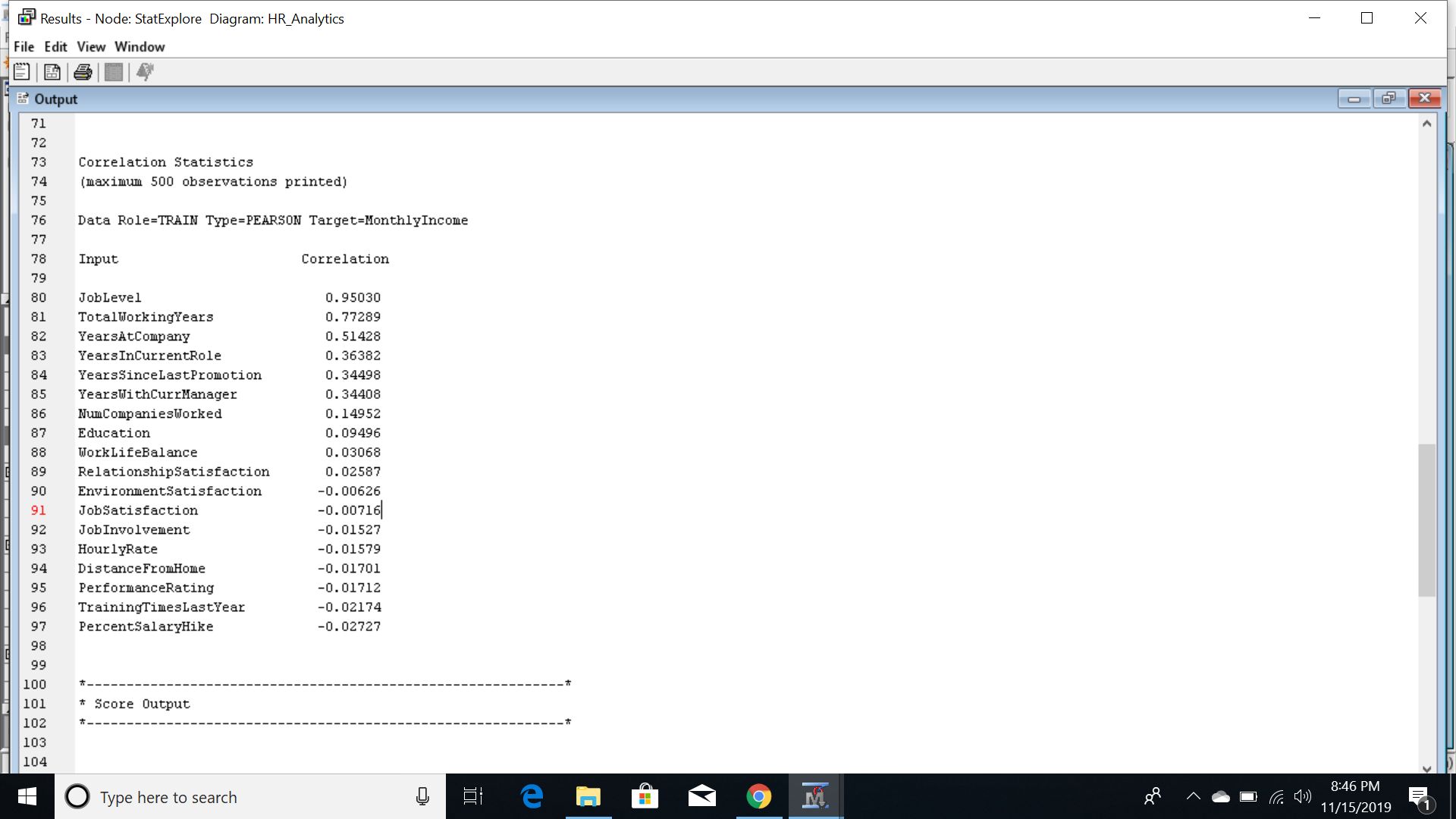
For the imported HR Analytics dataset and this initial investigation, we have listed Monthly income as the target variable. We hope to discover if any useful information can be gleaned regarding the employees represented in the sample and the monthly incomes they have been receiving.

Initially, we have run a Stat-Explore operation to get a statistical overview of the dataset. The variables selected are as follows:





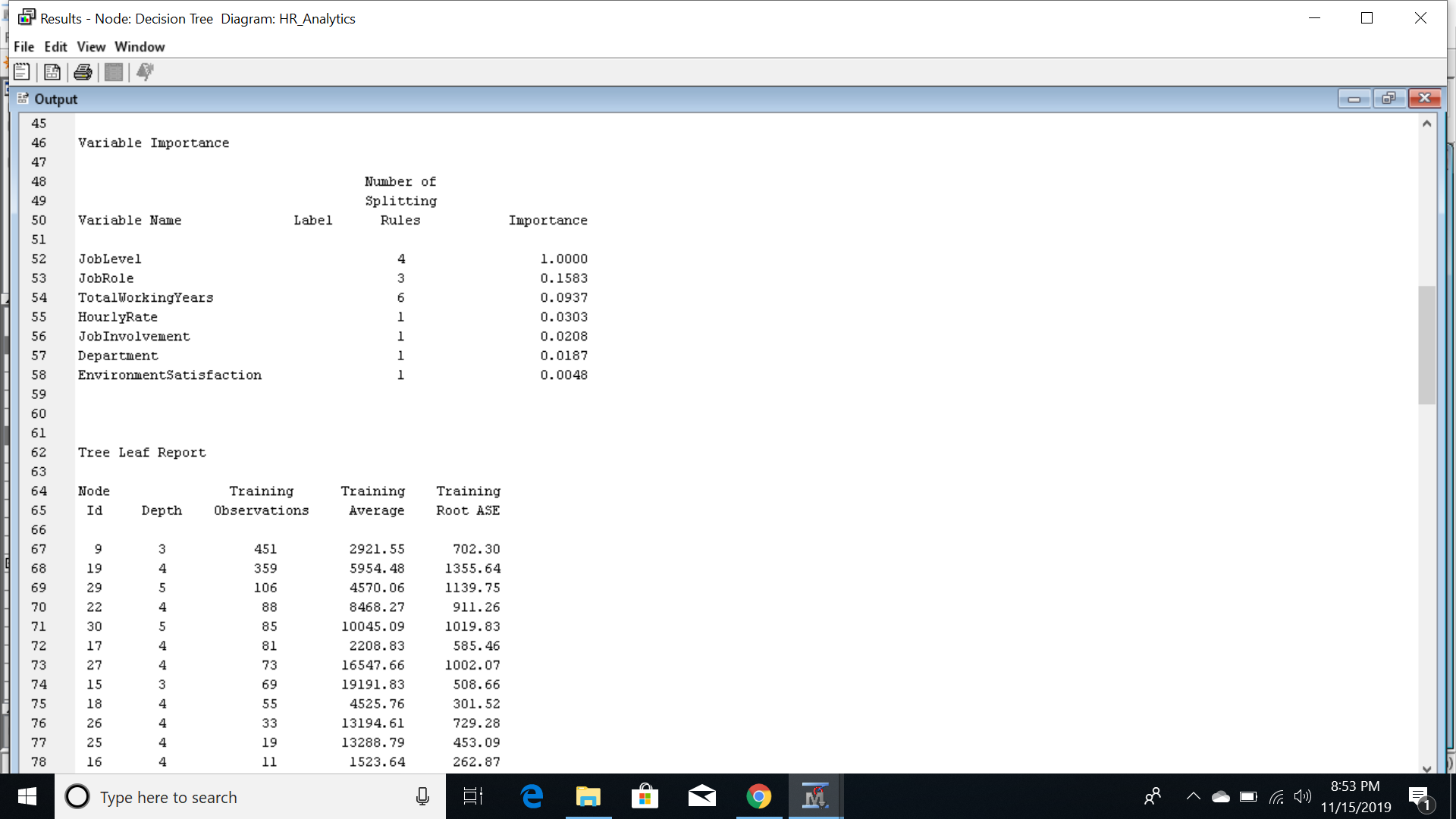
From the above Variable worth graph we can depict the target being Monthly income, Job level,Total working years, Job role has the highest influence on the target variable where as other variables such as OverTime,Job satisfaction contribute least which is very negligible.



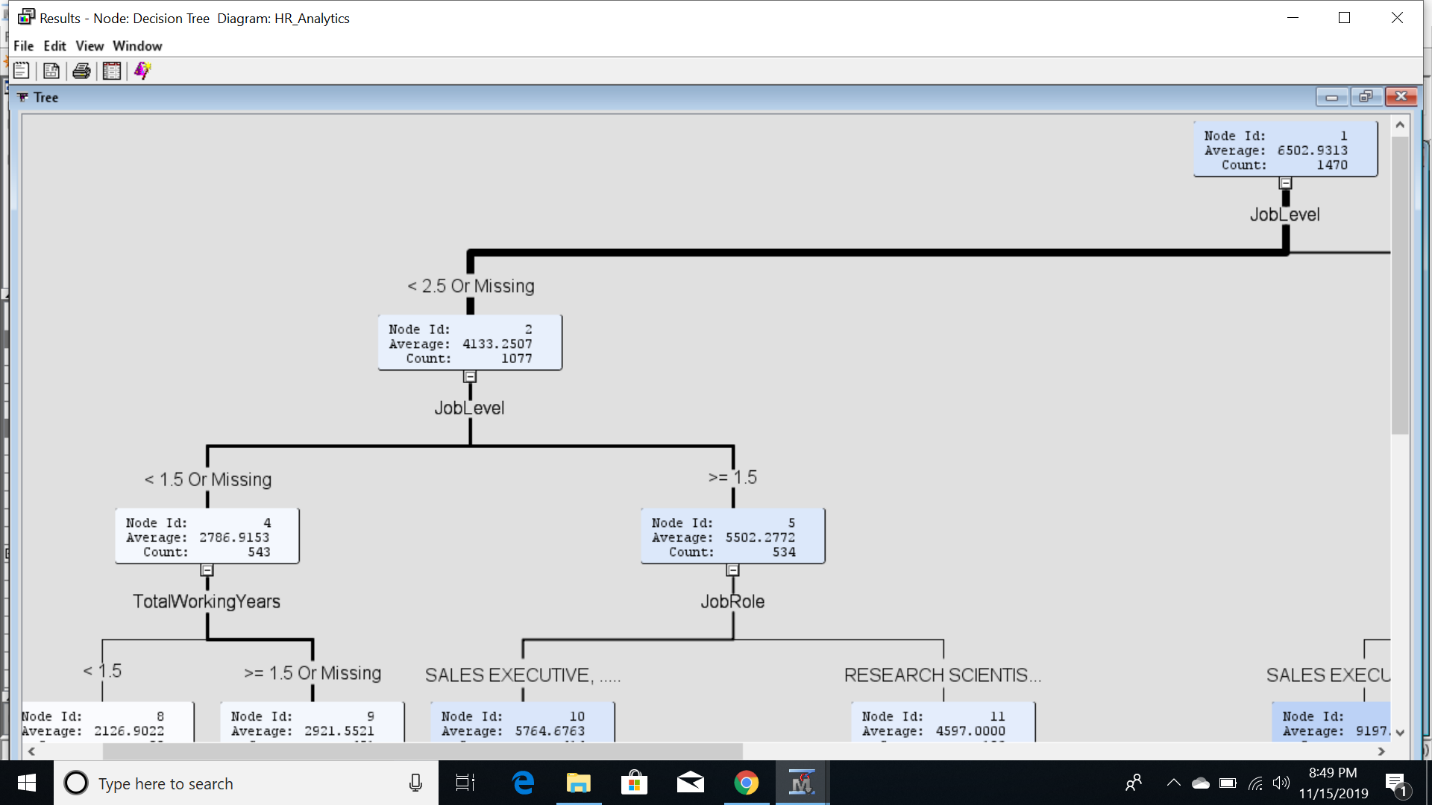
Considering Correlation Statistics, JobLevel has the highest correlation with almost 0.95030 followed by TotalWorkingYears with 0.77289 with the Target variable MonthlyIncome. The other variables like EnvironmentSatisfaction,DistanceFromHome, JobSatisfaction etc. have negative correlation with the target variable.

**Decision Tree**

Next, I ran a decision-tree model to take a closer look at the factors that are affecting the monthly income of an employee at IBM. The Tree diagram has several items of note which we’ll discuss.

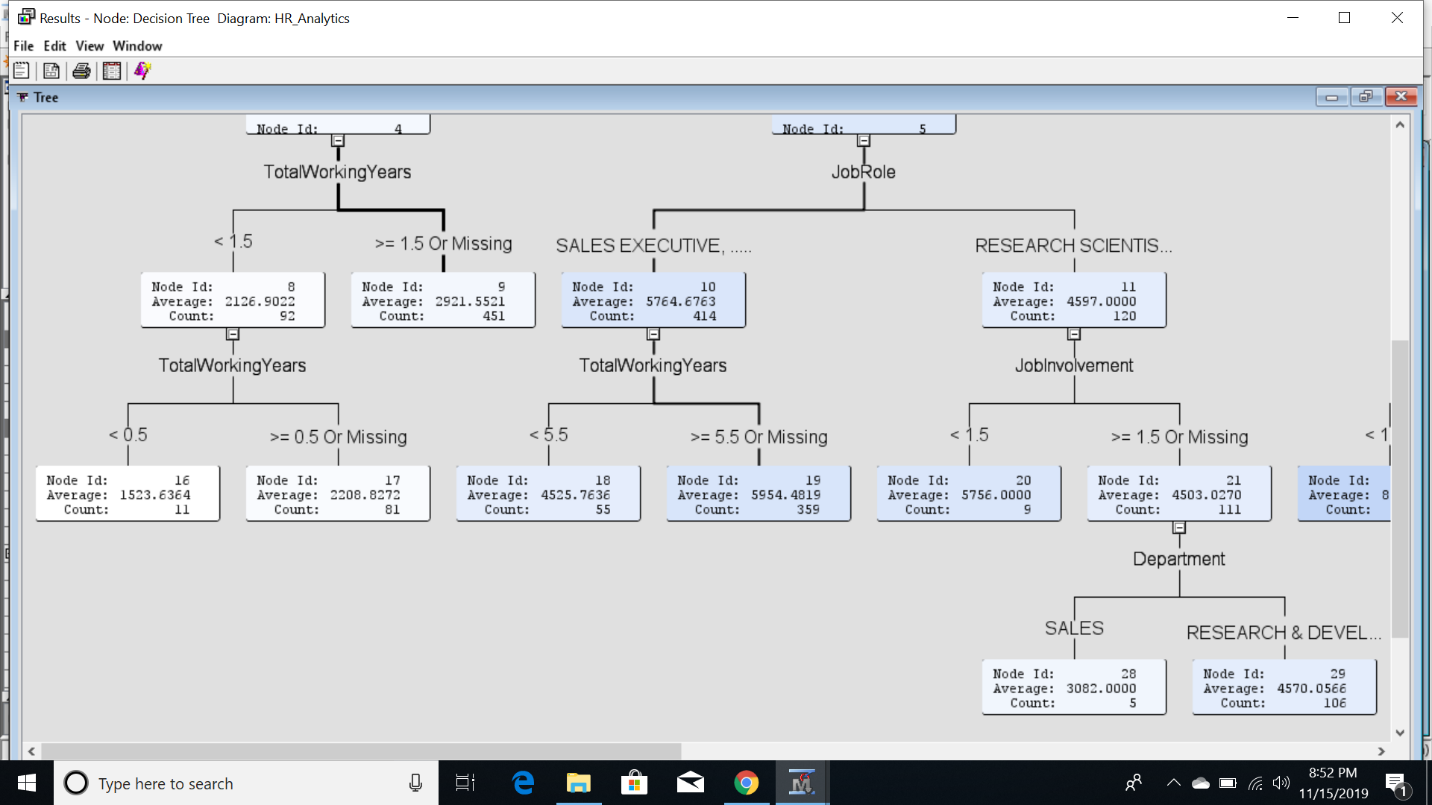


Variables used to split the decision tree are given in the above output. Job Level has been given the highest importance node.



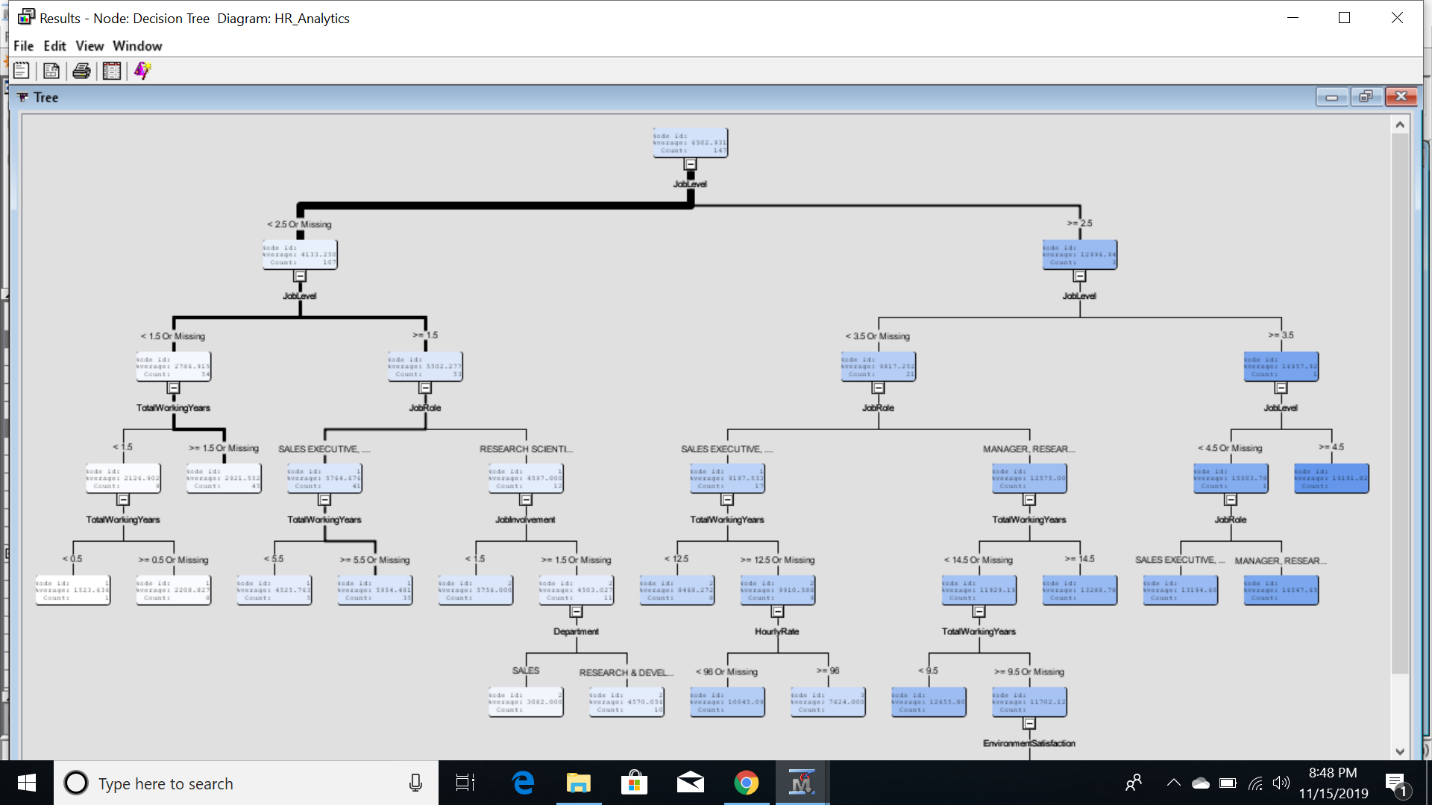
At Node1, JobLevel is splitted into two nodes with almost 73% of employees <2.5 and remaining >2.5.

Further At Node2, JobLevel is splitted into TotalWorkingYears and JobRole in almost equal ratio ie 50%.



At Node5, JobRole is splitted into two sub job roles Sales Executives and Research Scientists which is further splitted by the number of total working years and almost 86.7% of monthly income is decided based on the job roles and their no. of working years.

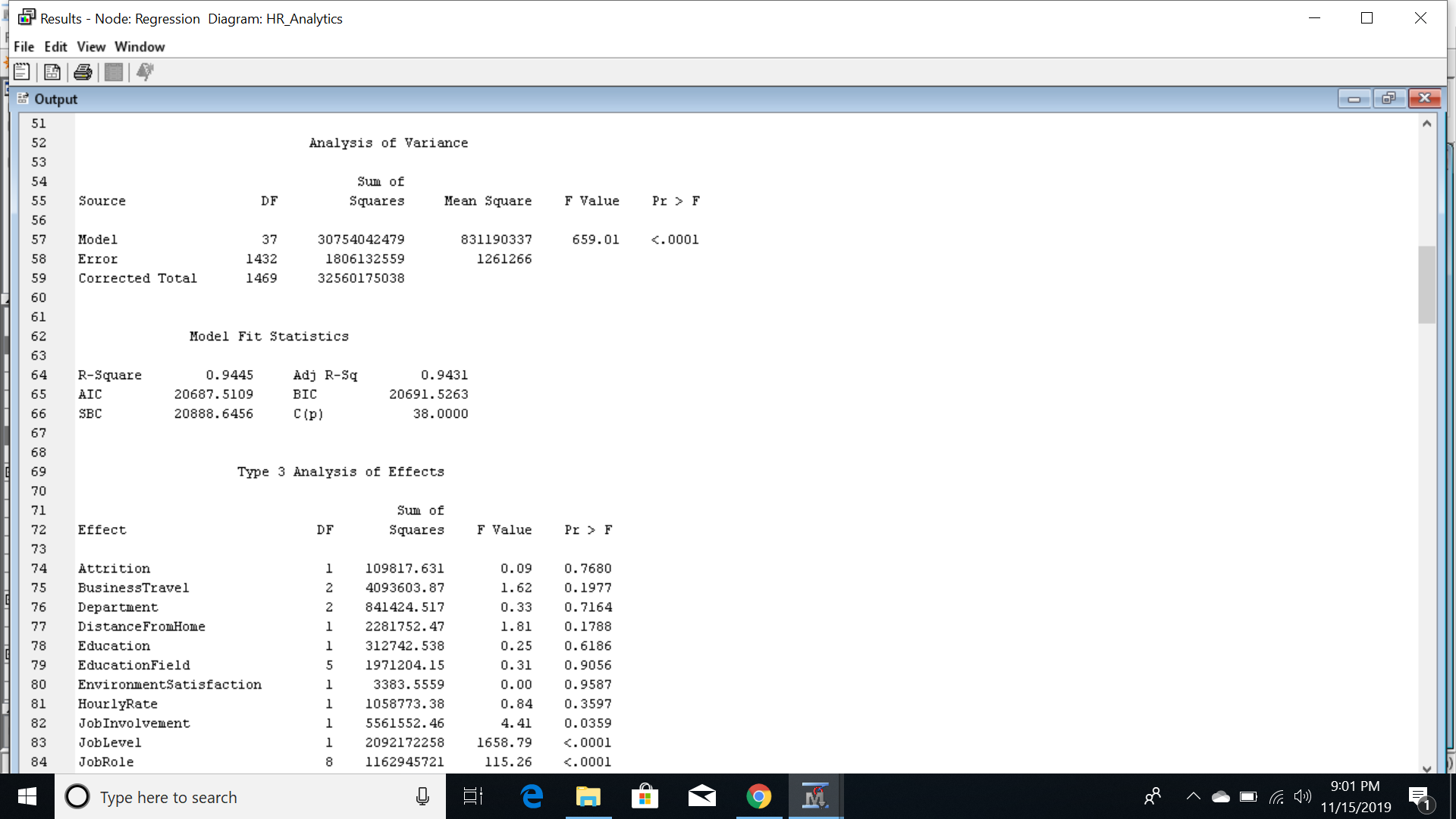
The complete decision tree diagram is as follows:



Further exploring into the decision tree model appears to be warranted. Running additional decision-tree models on each of the independent variables provides us with further insight.

**Regression Analysis**

Further, Regression Analysis is performed to find out whether model is significant or not and about the significant variables affecting the model significance. I performed both Linear Regression with all variables and also only with significant variables.

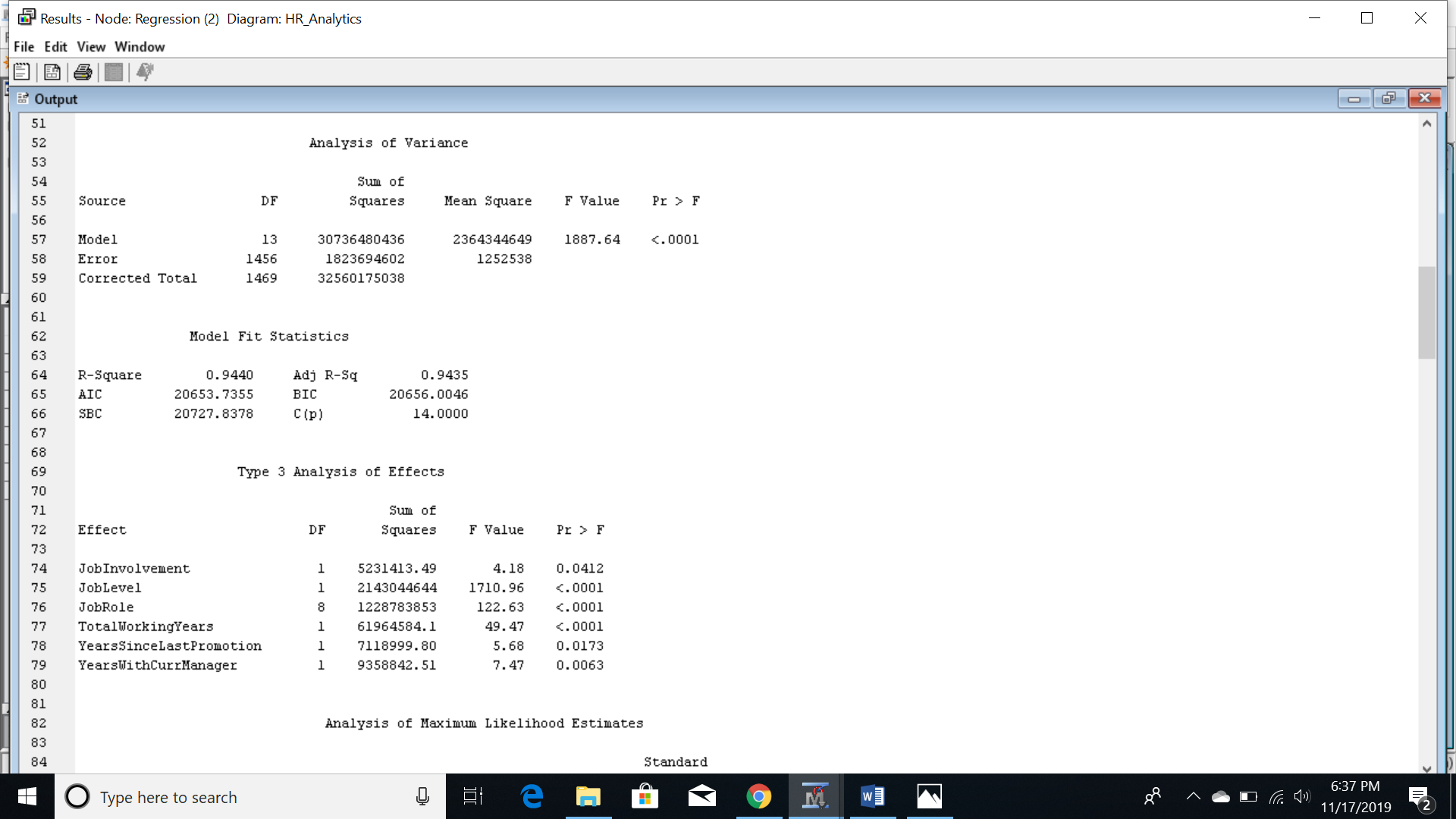
Linear Regression results are as follows,****

The HR Analytics model is significant as Pr value is <.0001 which is <0.05 which is threshold value.

R-square value is 0.9445 and adjusted R-square value is 0.9431 which states the model significance.

There are few significant variables like Job Level, Job Role, Job Involvement, TotalWorkingyears, YearsSinceLastProm, YearsWithCurrManager which plays crucial in model significance.

Regression is performed only using these significant variables and the results are as follows:



Pr value is <.0001 which is <0.05 which is threshold value.

R-square value is 0.9440 and adjusted R-square value is 0.9435.

Comparing both the models, there is not much difference in the variance of R-square.

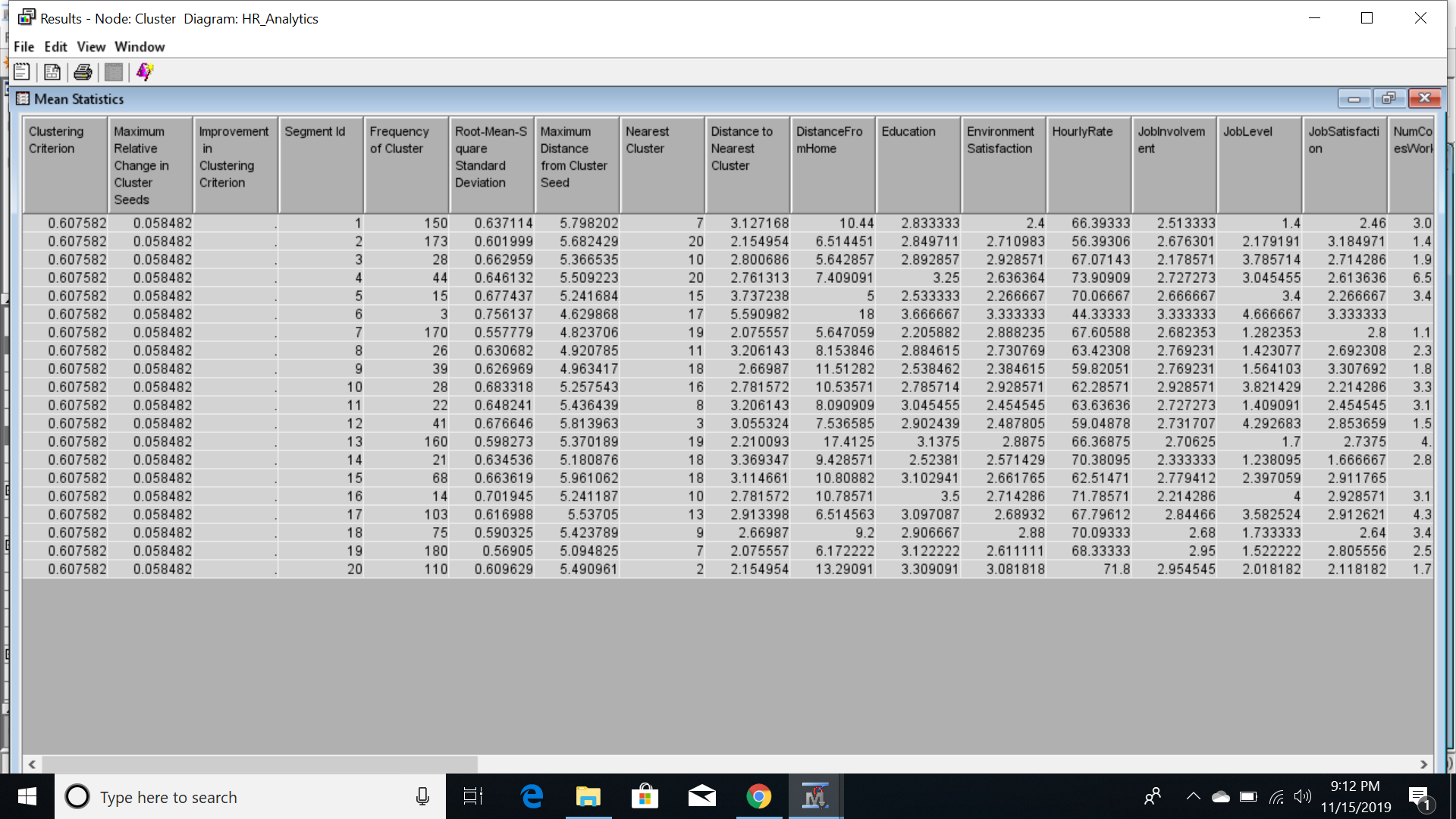
Therefore, around 94.4% of variation is observed related to monthly income.

**Clustering**

Clustering is one of the most flexible modeling approaches. One of the goal of data mining is to identify interesting, potentially useful patterns. One of the most important patterns we look for in data is similarity. When the data suggests that a group of objects are similar, we can begin to build generalized heuristics to help us understand the nature of those objects.

Here, The model is performed clustering so as to identify similar patterns. The results are as follows,

The number of segment IDs are 20 and all the observations are grouped into these particular segments. The frequency of the cluster shows that no obseravtion is left alone and most of them are clustered and have similarity in between them. The highest number of observations are in cluster 19 with 180 observations and least in cluster 8 with only 3.



Clustering can be challenging for many reasons. Cluster analysis often involves a lot of trial and error. The quality of your cluster solution is largely dependent whether we can obtain sensible results.

**CONCLUSION**

After consideration of the above models, we got to conclusion that we can predict monthly income based of various factors provided in the HR Analytics of IBM data set but also requires further exploration. For future exploration, smaller data-sets may be more useful for these kinds of analysis going forward. While the entirety of the data may be useful in finding long term trends, it does not allow for easy control of other concerns such as change in over time, personal or health issues of the employees. In particular, the exploring statistical findings might be more impactful when viewing the overall findings.

**REFERENCES**

Data set and Description: IBM HR\_Analytics : (<https://kaggle.com/>)