***Project 2***

***Social Data Mining Techniques***

***Group 8 : -***

* ***Amul Ghodasara – 200493038***



* ***Amish Dungrani - 200495456***



* ***Lakshmi Sasikumar***



***Topic: -***

* ***Industrial Engineering***

***Dataset***

**Source: -**

<https://www.kaggle.com/datasets/csafrit2/steel-industry-energy-consumption>

***About: -***

# **Steel Industry Energy Consumption**

This company produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown.

Attribute Information:

* **Date** Continuous-time data taken on the first of the month
* **Usage\_kWh** Industry Energy Consumption Continuous kWh
* **Lagging Current** reactive power Continuous kVarh
* **Leading Current** reactive power Continuous kVarh
* **CO2** Continuous ppm
* **NSM** Number of Seconds from midnight Continuous S
* **Week** status Categorical (Weekend (0) or a Weekday(1))
* **Day of week** Categorical Sunday, Monday: Saturday
* **Load Type** Categorical Light Load, Medium Load, Maximum Load

***Decision Tree***

Decision Tree is a learning technique used for both classification and regression models. It breaks down the structure in the form of nodes and leaves. Root is the first node and leaves is terminal node.

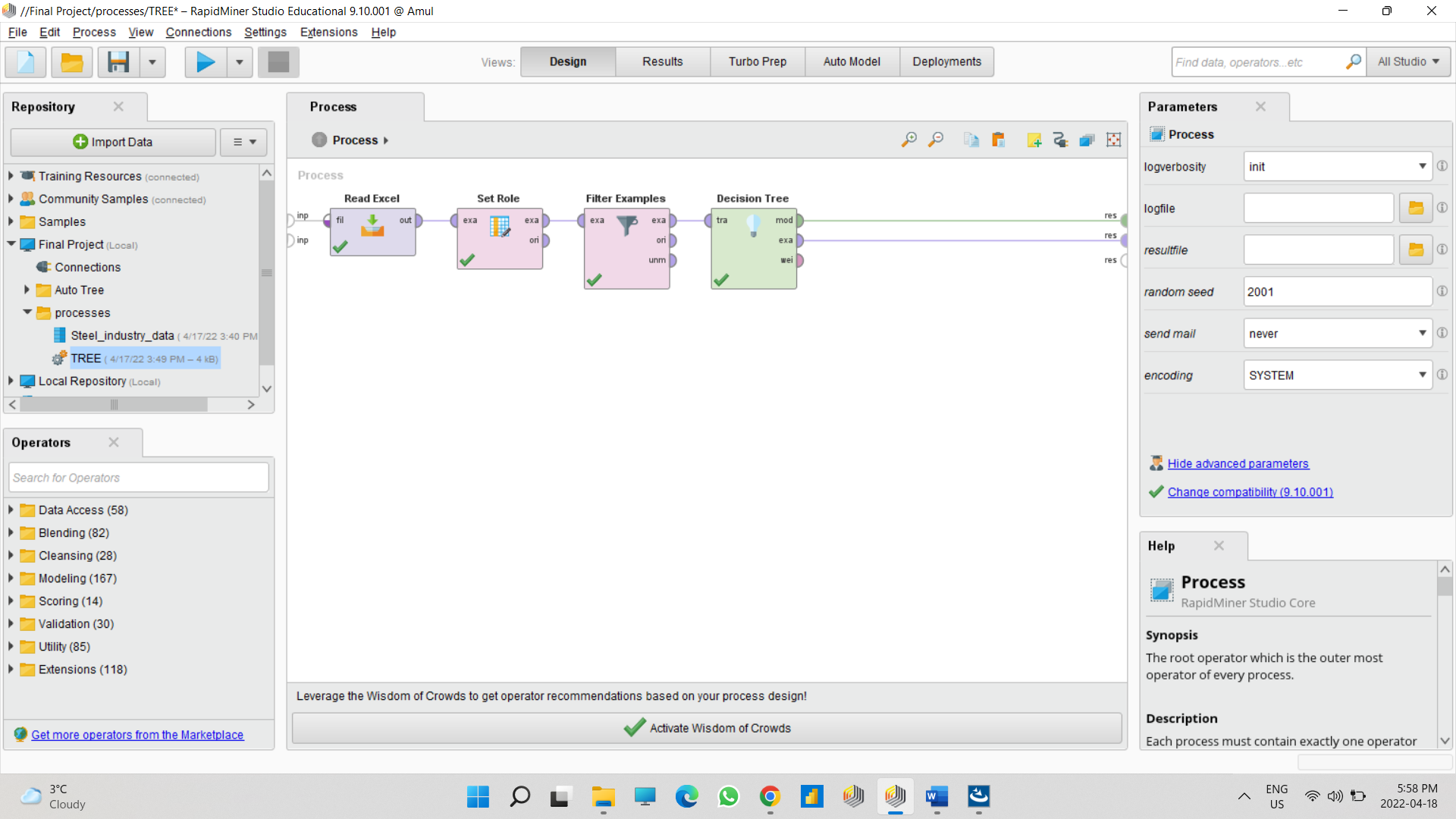
***Manual Technique: -***

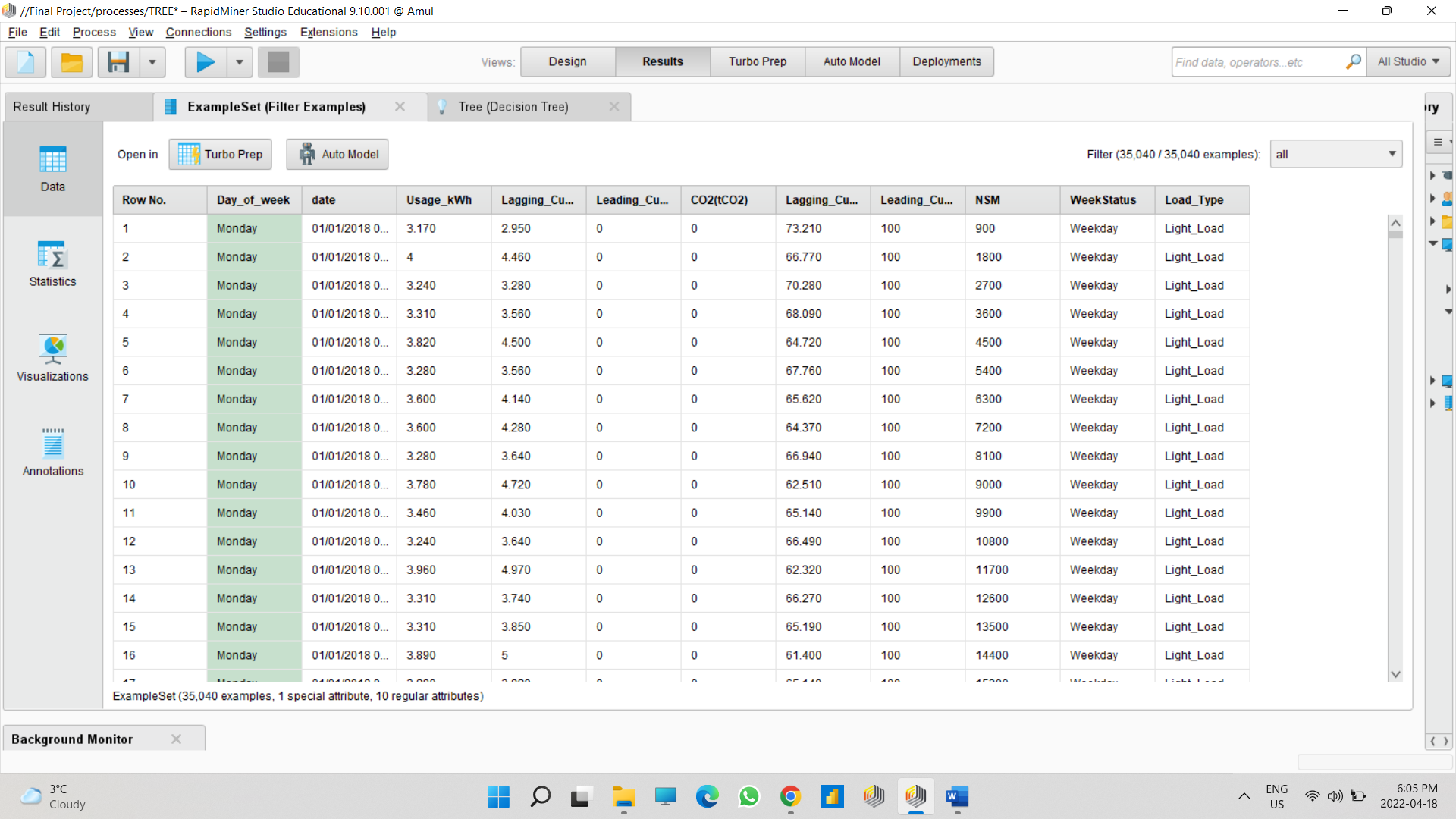
**Read Excel**: It will the load and read the dataset which is Excel format. Here I have loaded the data of steal industry energy consumption

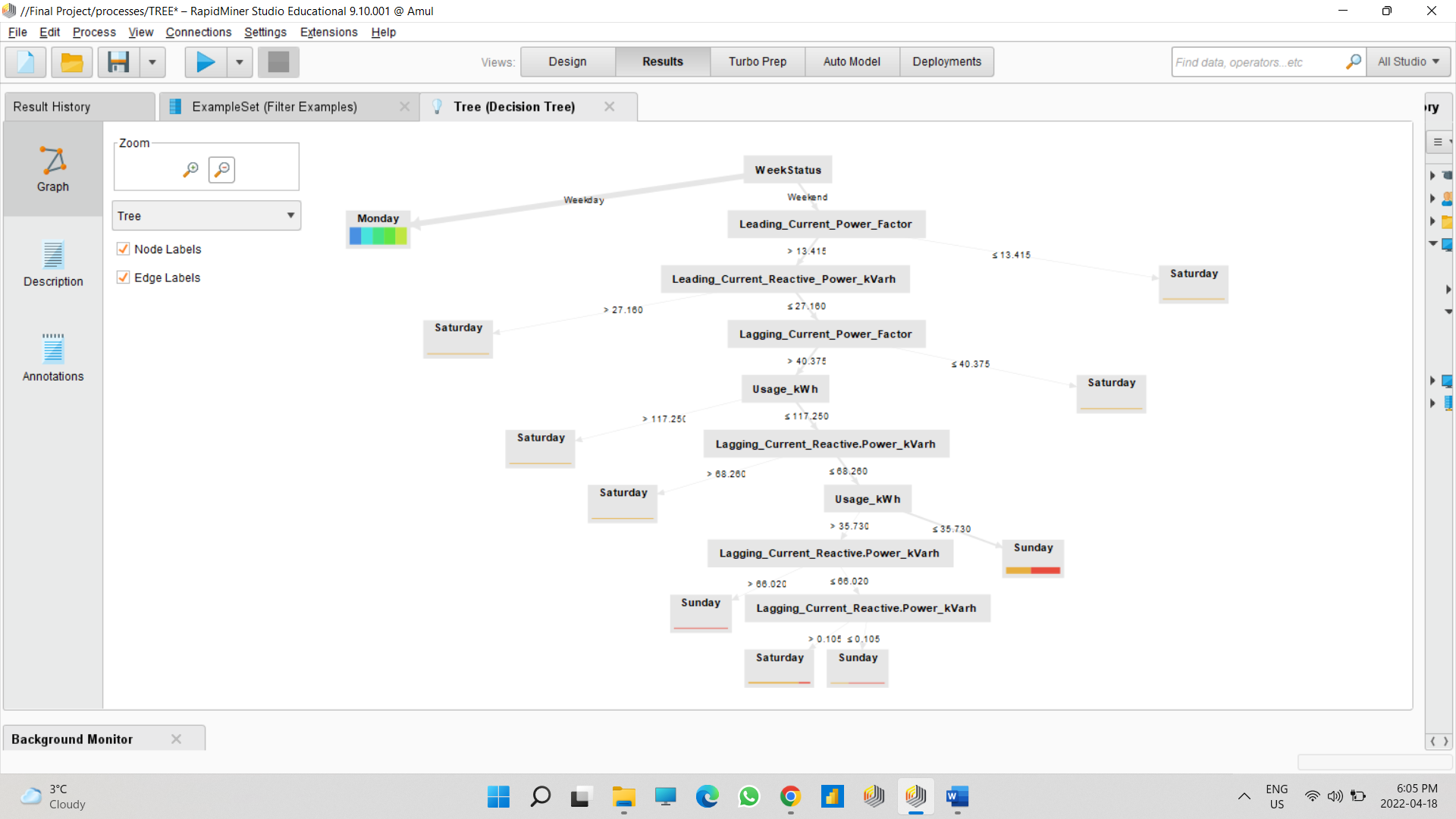
**Set Role**: It is used to set the attribute to target variable. Here day\_of\_week is the targeted variables and assign target role as label. This in decision tree is categorised by gender.

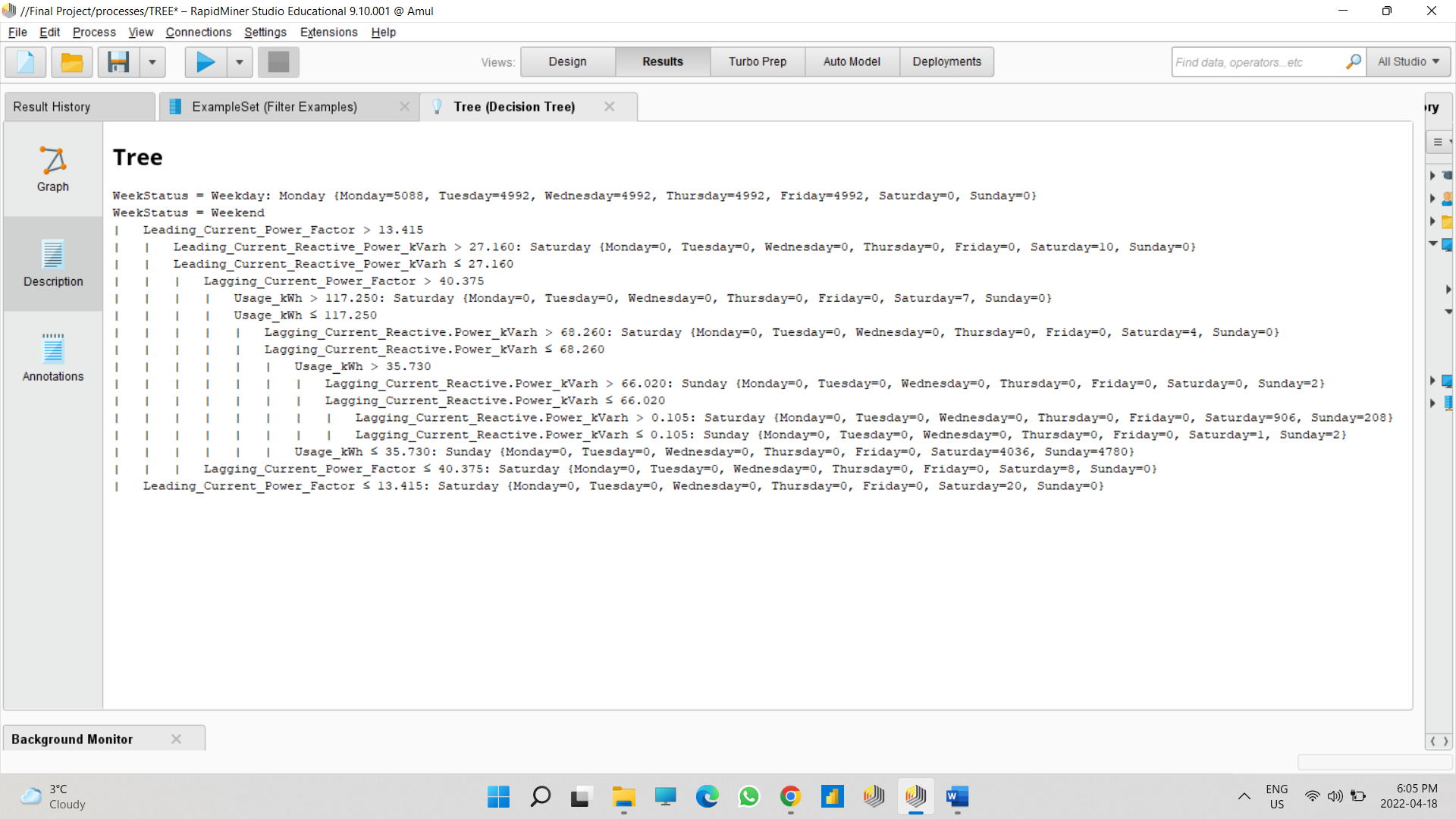
**Filter Examples:** This will filter out all the examples in the dataset. Here for day\_of\_week attribute, I have given the filter for not missing. This will display the data having the gender with no missing value.

**Decision Tree**: This operator build tree like structure from the input given. That is, it will display the decision tree according to Load\_type.

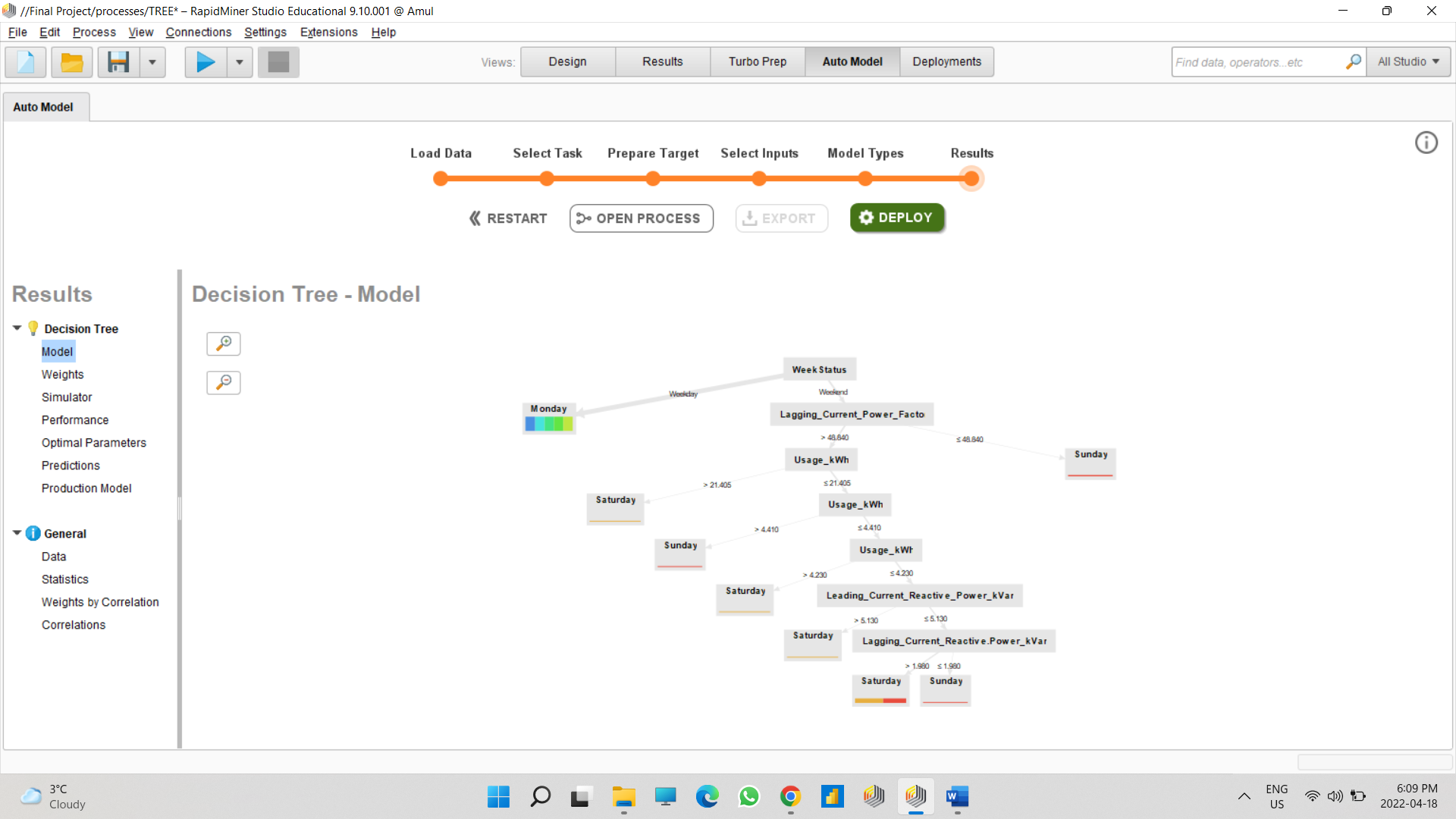


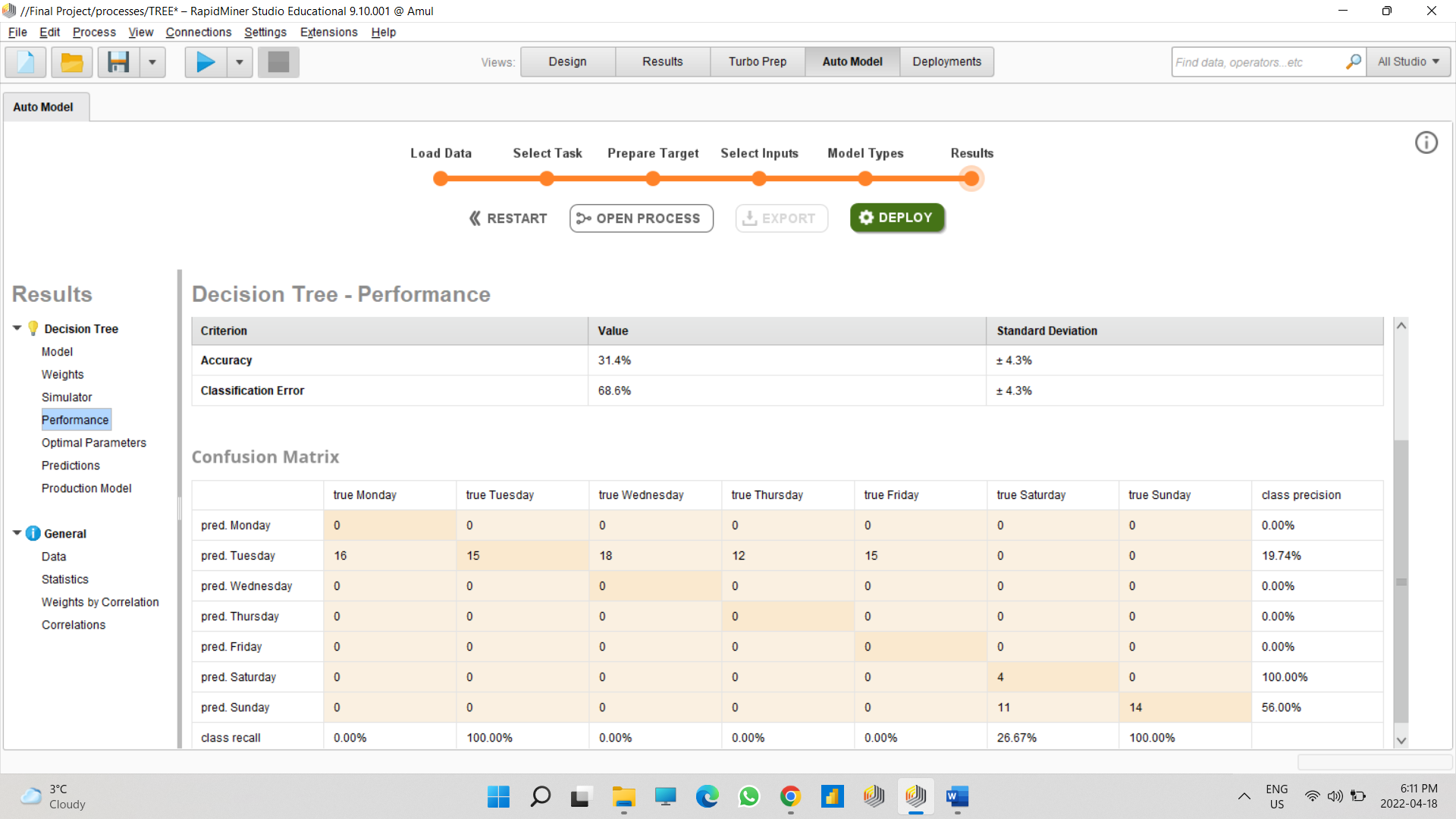


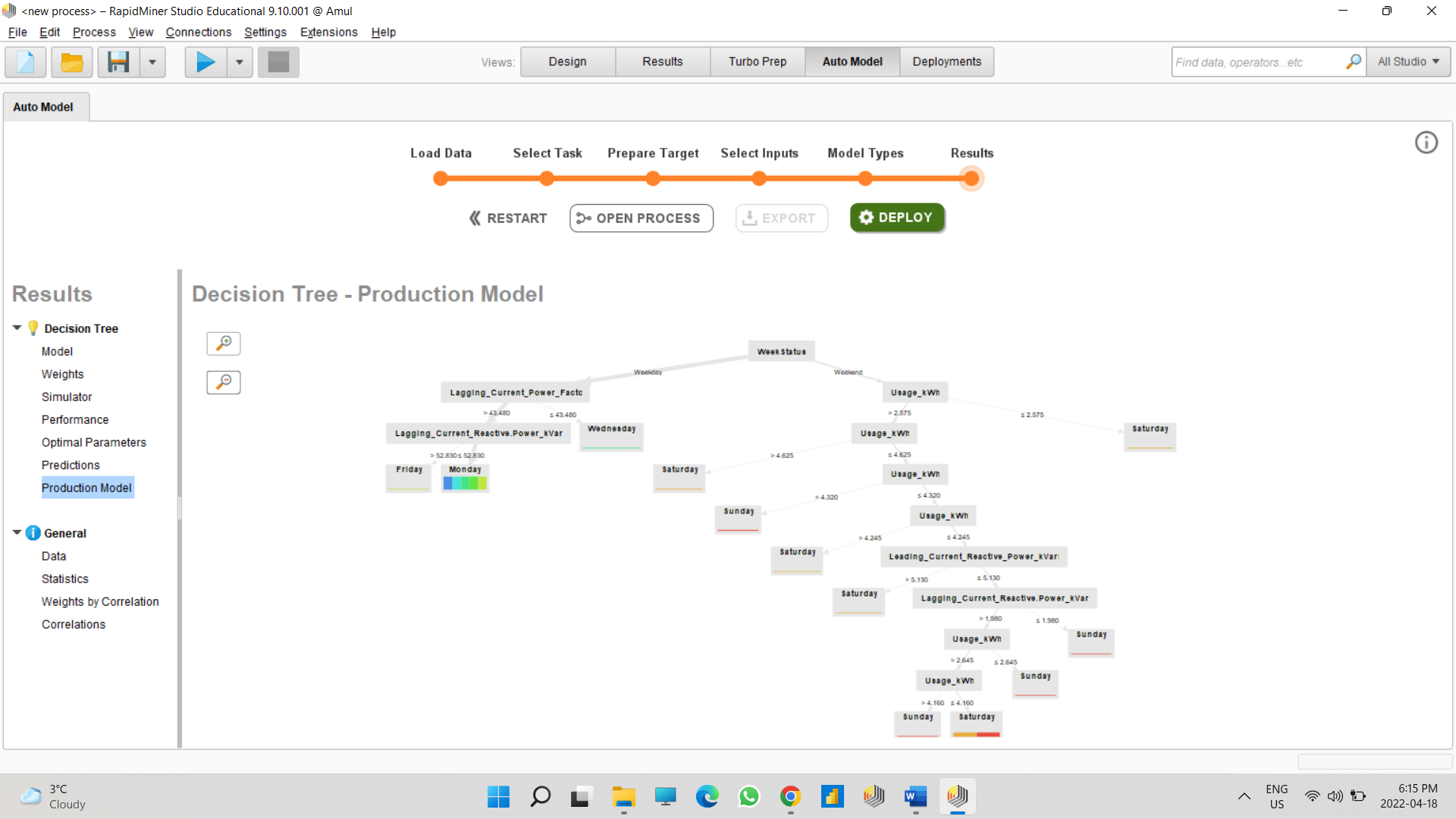




***AutoModel Technique: -***







After Performing Decision tree in manual and auto model technique, results of both techniques different. In auto model, I got 31.4% accuracy and 68.6% classification error.

***Linear Regression***

Here we are trying to predict the pattern and correlation between usage and lagging current reactive power, portioning the subset of data

3 outputs are generated. The first one predicted the usage of the industry based on lagging current. Usage and lagging current ha positive linear association that is 0.96

* Read CSV- It will the read the which is CSV FORMAT
* Select Attributes- This operator allows to select the field user needed from all the attribute displayed to predict linear regression from the file, here the attributes are usage and lagging current. The attribute filter type is subset, this will give multiple selection.
* Set Role- It will change the role of one or more attribute. Here the main attribute is usage, and this will be the dependent attribute with target role as label
* Split Data = It will create a subset of the main example set. Here shuffled sampling is used from different sampling techniques which is random. Here a subset for usage is created and split into two partitions for price into 70% and 30% of the values. 70% partition is giving to Linear Regression which is the core model and 30% to apply model
* Linear Regression – It will feed 70% quantified values and predict the model
* Apply Model – Apply the model as input of second partition and predict the model of core partition of data(output). This will general the label output
* Performance- This output has a performance specific metric like a number. And hence performance is calculated. Here we select one parameter, squared correlation.
* Write excel – Save the output into given location

*Outputs*

Diagram

Description automatically generated

Graphical user interface, text, email

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

*Visualization*

Graphical user interface

Description automatically generated

The visualization I have plotted is average usage and lagging current reactive power. Both are positively correlated and when one value increases the other also increases.

*Auto Model*

Graphical user interface, application, table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

In auto model, the correlation is same. But there is difference in root mean squared error value.

***Multilinear Regression***

Multilinear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. It is a kind of extension of linear regression.

**Manual Technique:**

Write Excel: What this operator will do is whatever the result you get in the tool this write excel operator will save in excel spreadsheet of those result in the tool and will get save at the specified location the user sets in the parameter.

**Select Attribute:** what this operator does is it selects a particular attribute in which users is interested in and add filters to those attributes only like comparing. If I talk about in my analysis, I wanted a particular attribute filter type to be subset and the attribute I wanted to apply it was on price and year from my data this is what select attribute me to help so the tool will select only subset attribute from the excel spreadsheet and will start comparing.

**Set Role:** What this operator does is it changes the role of one or more attributes. We used price, mileage, tax and year over here.

**Split Data:** What this attribute will do is it will create a subset off the core dataset. So to do this split data is have used here is shuffled sampling which is based randomly. We made a random set o the price data over here. Than in that attribute I have done is I have partition the data of price into two different partitions. One has 70% of the values and other has 30% of values. This will produce desired number id subset.

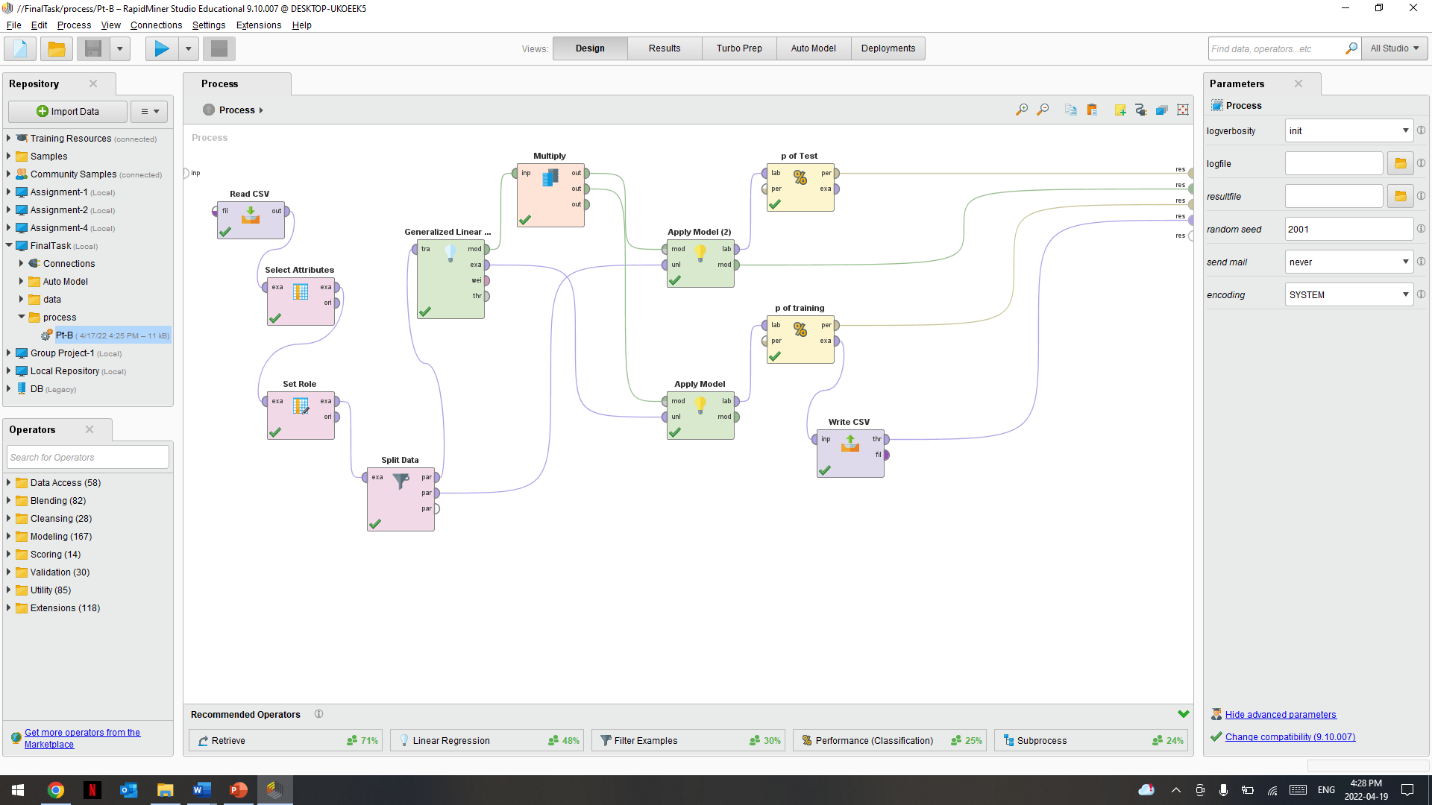
**Apply Model:** What this operator does is it trained on the example set which is left by learning algorithm. This is like a prediction it will predict the values of the unlabeled data and what it will do is it will generate me a label data output and will perform a specific metric like any number.

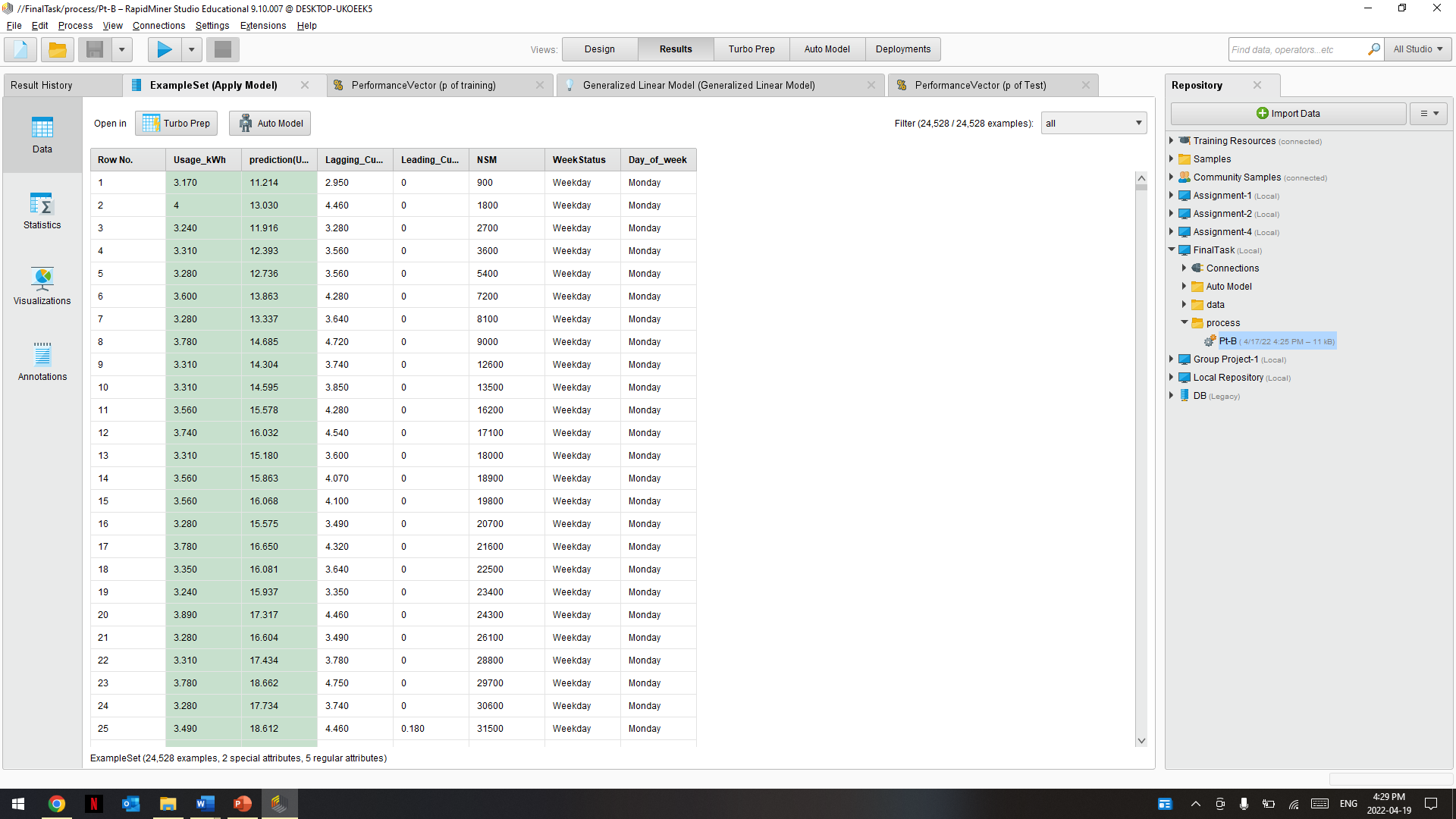
**Generalized Linear Regression:** The partition one from the desired subset will go come here in linear regression. In this linear regression it can only store numerical values. It will calculate a linear regression based on the input. Once the output is generated this regression model which generates can be applied on the dataset.

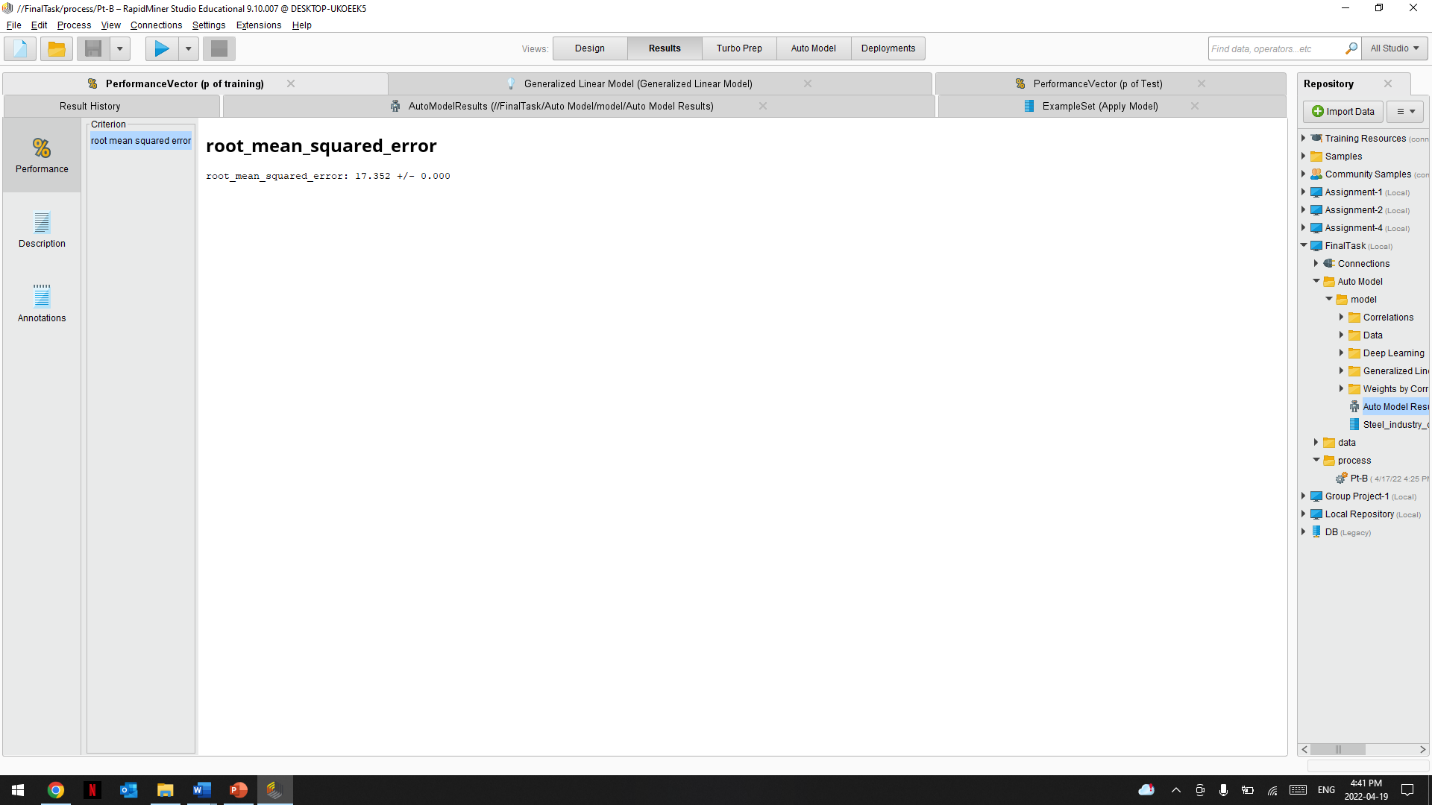
**Multiply:** What this operator will do is it will create a replica copies and send duplicates of whatever the unput comes in from linear regression. And it will contain one output that we can apply to the model and the other output will apply to the second model. We will get multiple duplicate copies.

**P of Test:** what this does is it will evaluate the performance of the regression same with p of testing. It will calculate square correlation.

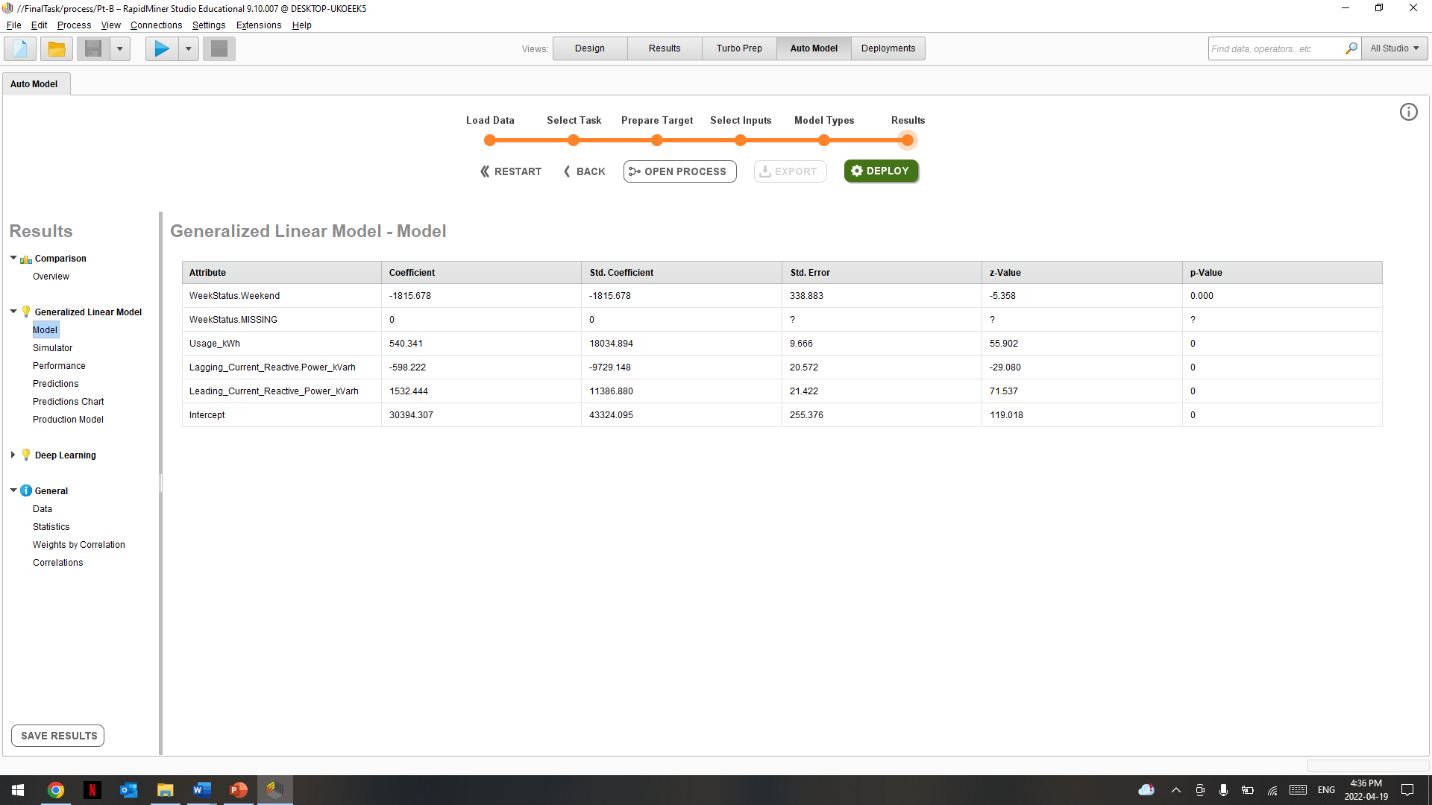
**Read Excel:** This operator is used to load data from the excel spreadsheet which was generated by the write excel in the tool Rapidminer.







**AutoModel Technique:**



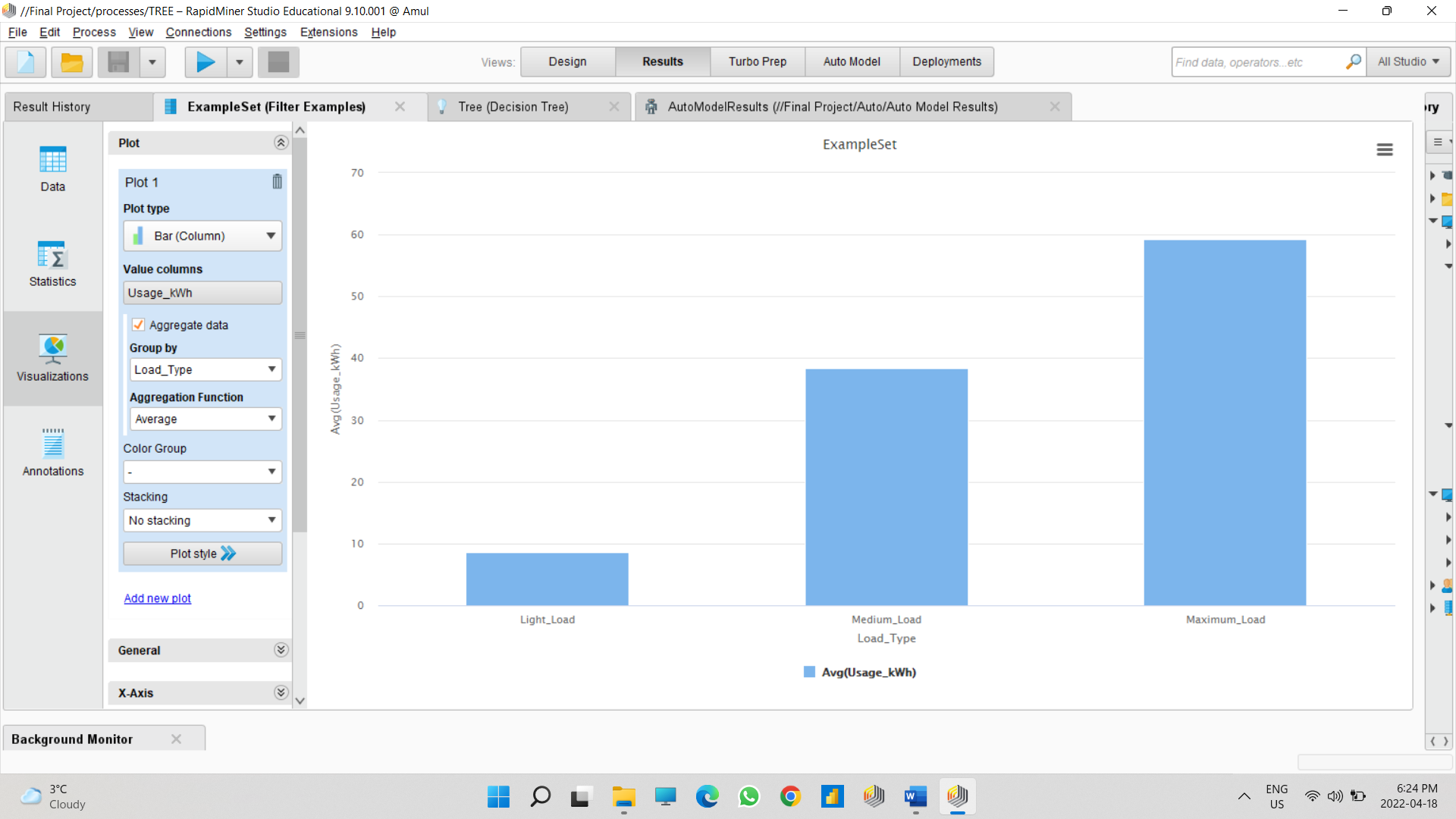




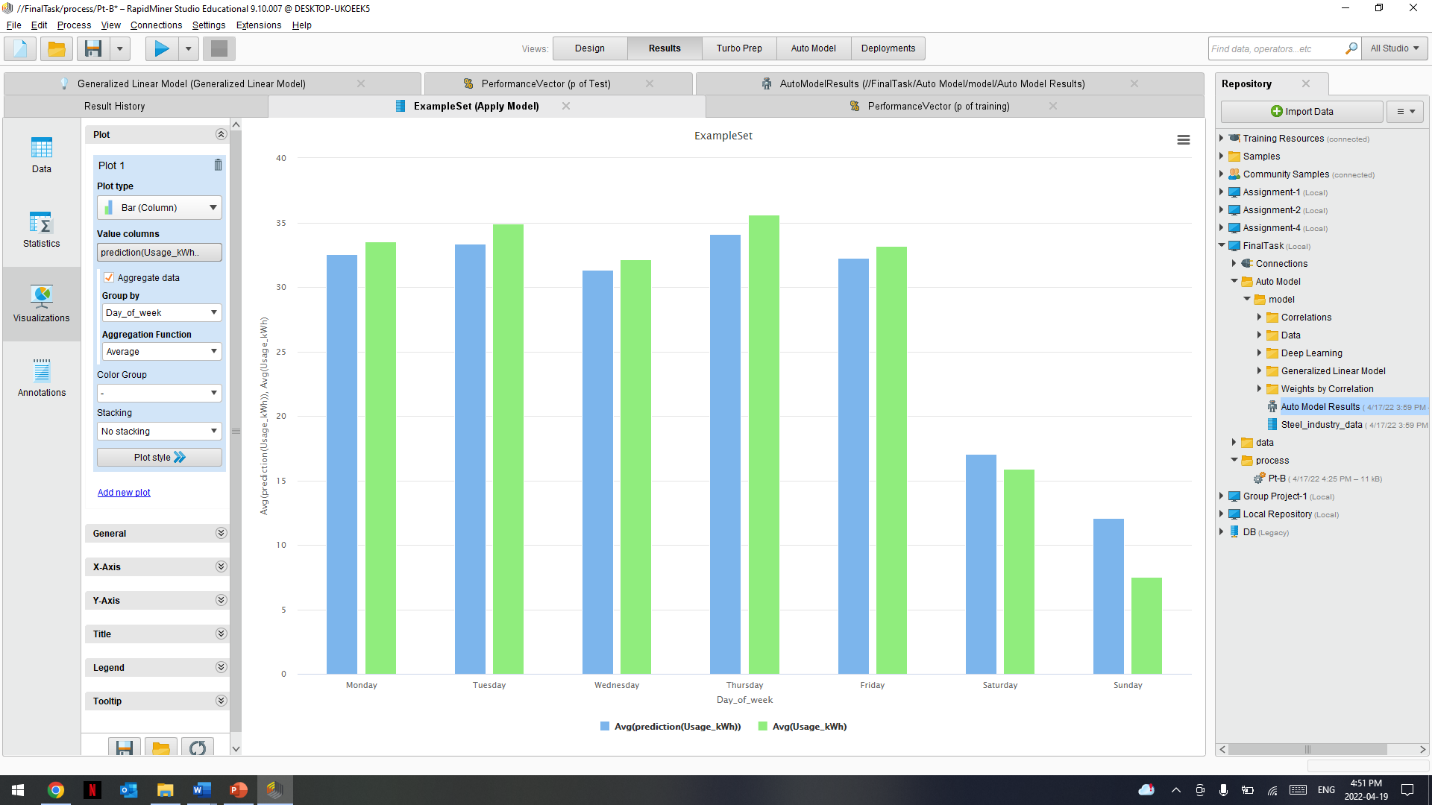
After performing the technique in manual and in auto model and if I try to compare the result of both than in manual the the prediction values of usage kWH is more than the value prediction is auto model.

.

***Visualizations: -***



This Visualization is all about average usage\_kWh by load type. It indicates that which load\_type has highest more average of using usage\_kWh. As we can see here that, Light\_load has less average of Usage\_kWh with 8.62. While Maximum\_load has highest average of usage\_kWh with 59.26.



The Visualization which we have created here is about the usage\_kWH of the prediction and of actual usage\_kWH by the days of week and the is much difference between both the values as you can see in chart the blue bar depicts the avg of the prediction usage\_kWH is less than the actual avg usage\_kWh for each and every day of the week.