UNIVERSITY OF MALAYA

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

WQD7005 DATA MINING

GROUP PROJECT | CROP PRODUCTION STATISTICS

CROP PRODUCTION STATISTICS

1 Introduction

In recent years, crop yield forecasting has become a research hotspot in the field of agricultural science and plays a key role in solving food production problems. Accurate and timely crop yield forecasting is therefore of great importance for the formulation of relevant national food policies, as well as providing a sound basis for agricultural decision-making and an important basis for crop improvement measures.

In this project, a dataset containing comprehensive data (disaggregated by state and region) on crop production statistics in India was used. The dataset covers annual production and yield information for crops grown in different regions of India from 1997 to 2023. By analyzing this dataset, factors affecting crop yields and production can be identified and crop yields can be projected for different regions of the country.

The use of forecasting technology to scientifically predict and evaluate crop yields can effectively prevent and improve irrational problems in the process of crop production, and is of great practical importance for the scientific deployment of agricultural production and the formulation and adjustment of agricultural policies, as well as for the country to promote food production, enhance food production capacity and ensure food security and stability.

2 Dataset

The dataset contains detailed information on crop production statistics in India, categorized by state and district, covering four major crop seasons from 1997 to 2023. This dataset is valuable for researchers, policymakers, and farmers who wish to understand crop production patterns in different regions of India. Analyzing the data can help identify factors that affect crop yields and production, enabling informed decisions on improving agricultural productivity. Policymakers can use this information to create and implement sustainable farming policies that improve food security. Farmers can also benefit by making informed decisions on the best crops to grow and how to manage them. Furthermore, the dataset can be used to train machine learning models for predicting crop yields and production in different regions, which is useful for agricultural organizations. Overall, this dataset offers a comprehensive view of crop production statistics in India, critical for understanding the agricultural landscape and developing strategies for sustainable agriculture.

3 Business Understanding

3.1 Analysis Goal

The analysis of the study is to identify factors affecting crop yields and production and forecast crop yields for different regions of the country. Provide an important basis for the designation of relevant food policies and agricultural-related measures.

3.2 Analysis Data

This dataset contains detailed information on agricultural production statistics in India, sourced from the Government of India's Area Production Statistics (APS) database, maintained by the Ministry of Agriculture and Farmers' Welfare, which provides detailed data on crop production, yield and area under cultivation for each state and district in India. The below objectives are set in this project:

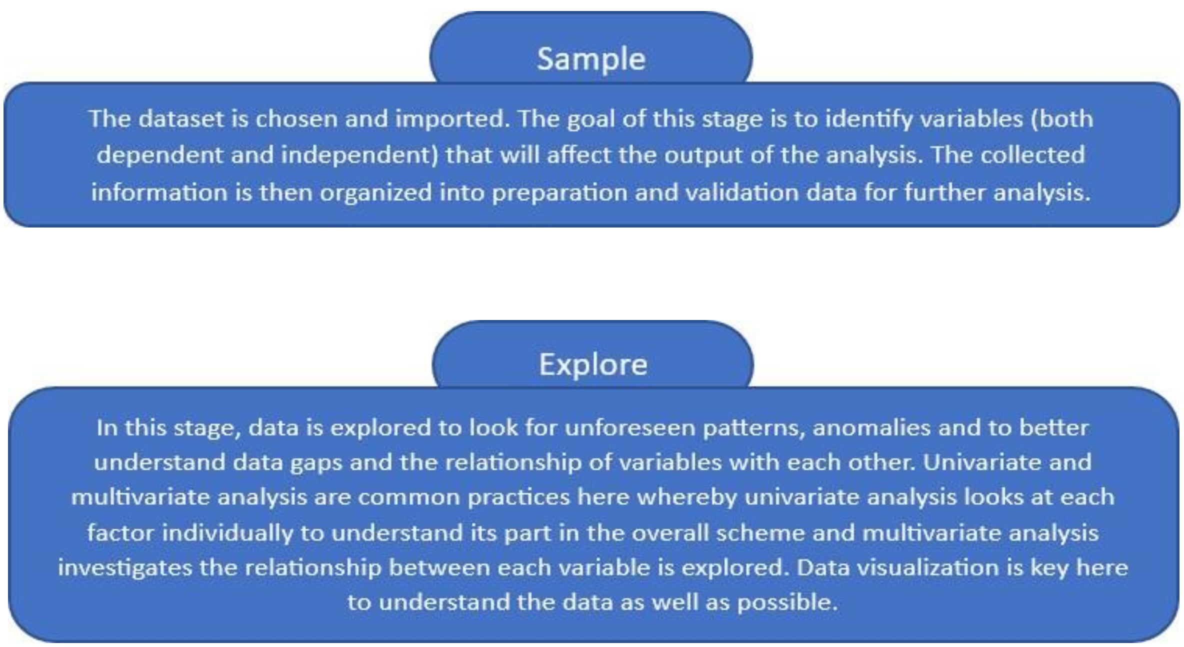
A. To analyze trends and patterns in our dataset.

B. Determine the factors that have the greatest influence on crop yield.

C. Executing and evaluating decision tree models to predict future outcomes.

4 Methodology

Data-driven organizations incorporate the SEMMA methodology into their data analytics to gain competitive advantage, enhance performance and provide useful services to clients. SEMMA is created as a Data Science methodology to assist practitioners in transforming raw data into fruitful insights. SEMMA is leveraged as an efficient toolset or is claimed by such as SAS to be associated with SAS Enterprise Miner software. The main 5 steps in the SEMMA process are **Sample**, **Explore**, **Modify**, **Model** and **Assess**. In this project of analyzing crop production statistics in India, the first 2 steps, **Sample and Explore**, are implemented to initiate the analysis. Below is the in-depth explanation of all 5 stages involved in SEMMA procedure.



Modify

In this phase, data is parsed and cleaned, before passed onto the modelling stage, and explored if the data requires further refinement and transformation. Data is altered by creating, selecting, and transforming variables to centre the model selection and any additional information or variables can be added to make information output more significant.

Model

Various modelling or data mining techniques are applied to the pre-processed data to benchmark their performance against desired outcomes. This stage is essential to produce a projected model of how this data achieves the final, desired outcome of the process.

Assess

In this final SEMMA stage, the model is evaluated for how useful and reliable it is for the studied topic. Evaluation and interpretation of data are performed to compare the model outcome with the actual outcome and further analysis of model limitation can be done to overcome it.

5 Results

5.1 Sample

This dataset possesses extensive information on agricultural production statistics in India, sourced from the Indian government's database of Area Production Statistics (APS). This database provides detailed data on crop production, yield, and area under cultivation across different states and districts and is maintained by the Ministry of Agriculture and Farmers Welfare in India. The dataset has 345337 rows of observations (including headers) and 8 variables in which one of the variables is the target output.

Variables are as below:

i. State - Name of the State
ii. District - Name of the District
iii. Crop - Variety of Crops
iv. Crop Year - Year in which crop was produced
v. Season - Periods of the year marked by marked by particular weather patterns and daylight hours
vi. Area - Area in Hectares
vii. Production - Production in Tonnes
viii. Yield - Yield (Tonnes/Hectare)

5.1.1 Metadata

The Figure 5.1 below shows the column metadata of the dataset imported into the diagram.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Type	Format	Informat	Length
Area	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Crop	Input	Nominal	No		No	.	.	Character	$19.	$19.	19
Crop_Year	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
District	Input	Nominal	No		No	.	.	Character	$24.	$24.	24
Production	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Season	Input	Nominal	No		No	.	.	Character	$10.	$10.	10
State	Input	Nominal	No		No	.	.	Character	$26.	$26.	26
Yield	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8

Figure 5.1 Column metadata (Basic setting)

The type of each dataset is mentioned for each column variable such as numeric and character. From the imported dataset, it is observed that there are 4 interval variables and 4 class variables. By default SAS identifies any numeric data as interval variables and character data as class variables. Hence, manual adjustment is required as some variables require changes in the role and level assigned.

5.1.2 Adjustments of role and level of variables

Adjustments were performed on variables that required role and level change. The Figure 5.2 below shows the changes in role and level before and after manual adjustments were performed.

Name	Role	Level
Area	Input	Interval
Crop	Input	Nominal
Crop_Year	Input	Interval
District	Input	Nominal
Production	Input	Interval
Season	Input	Nominal
State	Input	Nominal
Yield	Input	Interval

Name	Role	Level
Area	Input	Interval
Crop	Input	Nominal
Crop_Year	Input	Nominal
District	Input	Nominal
Production	Input	Interval
Season	Input	Nominal
State	Input	Nominal
Yield	Target	Interval

Figure 5.2 Comparison between role and measurement level between advance settings and manual reclassification

The 'Crop_Year' variable consists of numeric values in which SAS identified as an interval variable by default. In this project, this variable would be used as a category (class) variable to identify which year produced the highest yields for each crop type. The 'Yield' variable is a target variable which will be used to evaluate the amount of crop produced based on respective categorical variables.

5.1.3 Handling of missing values

The Figure 5.3 below shows the missing values in each class and interval variables respectively.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Crop	INPUT	60	318	Rice	6.25	Maize	5.93
TRAIN	Crop_Year	INPUT	25	618	2019	5.55	2018	5.27
TRAIN	District	INPUT	513	0	BILASPUR	0.53	BELAGAVI	0.51
TRAIN	Season	INPUT	249	309	Kharif	40.04	Rabi	29.11
TRAIN	State	INPUT	38	0	Uttar Pradesh	12.97	Madhya Pradesh	8.66

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Area	INPUT	11766.46	48880.76	345028	309	0.004	532	8580100	43.73859	4999.735
Production	INPUT	959187.6	21540404	340080	5257	0	712	1.5978E9	35.85735	1582.846
Yield	TARGET	79.16214	915.2795	344719	618	0	1	43958.33	14.80243	261.9655

Figure 5.3 Missing values observations

It is observed that most variables have missing values. The count of rows with missing values is insignificant when compared with the total size of the dataset. Hence, data reduction is performed in the modify stage to eliminate rows with missing values.

5.2 Explore

The goal of this phase is to get deeper into the data, identify patterns and trends, and discover insights that can inform decisions. By exploring the data, analysts can better understand the characteristics and quality of the data and identify potential outliers or errors. At the same time, through the exploration of the data, the relationship between different variables in the data can be identified, distribution and correlation can be explored, and visualization can be created to make full preparation for the next step of data analysis.

5.2.1 Summary Statistics

Summary statistics in SAS is a useful feature that can help generate a clear overview of the data pattern such as minimum and maximum value, mean, and missing values. Figure 5.4 showed detailed information about the interval variable summary statistics and Figure 5.4 showed the summary statistics for class variables.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Crop	INPUT	60	318	Rice	6.25	Maize	5.93
TRAIN	Crop_Year	INPUT	25	618	2019	5.55	2018	5.27
TRAIN	District	INPUT	513	0	BILASPUR	0.53	BELAGAVI	0.51
TRAIN	Season	INPUT	249	309	Kharif	40.04	Rabi	29.11
TRAIN	State	INPUT	38	0	Uttar Pradesh	12.97	Madhya Pradesh	8.66

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Area	INPUT	11766.46	48880.76	345028	309	0.004	532	8580100	43.73859	4999.735
Production	INPUT	959187.6	21540404	340080	5257	0	712	1.5978E9	35.85735	1582.846
Yield	TARGET	79.16214	915.2795	344719	618	0	1	43958.33	14.80243	261.9655

Figure 5.4 detailed information about the summary statistics

5.2.2 Variable analysis

The Figure 5.5 pie chart below shows the count of states that contribute to yield regardless of other factors. It is observed that the state of Uttar Pradesh contributes to the highest yield compared to other states, and its percentage is 13.14%. On the contrary, the state of Gujarat contributes the lowest yield, and its percentage is 4.11%. On the other hand, production in other regions is more average.

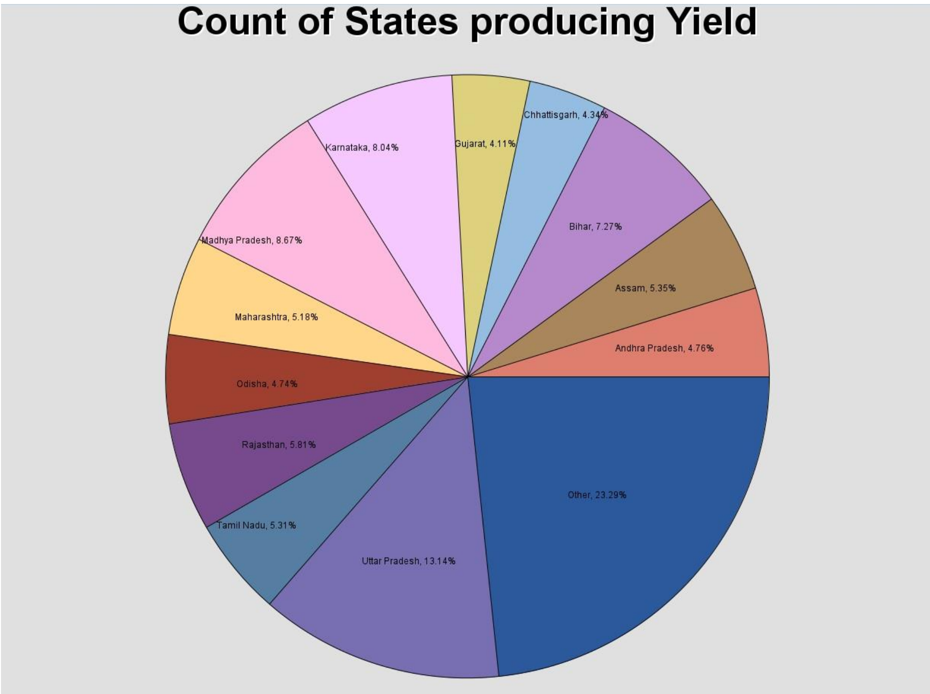
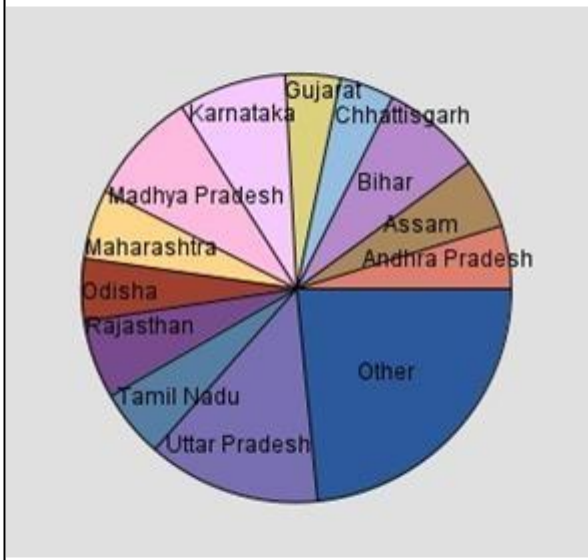


Figure 5.5 Count of States Producing Yield

For nominal variables, pie charts are used to present the proportional breakdown of data in nominal variables in a visual way, especially when there are only a few categories. The following Table 5.1 shows the overview of the variables in pie charts and findings.

No.	Count of some variables producing yield	Findings
1	**State**  	1. The state of Uttar Pradesh contributes to the highest yield compare to other states 2. The state of Gujarat contributes the lowest yield 3. Production in other regions is more average.

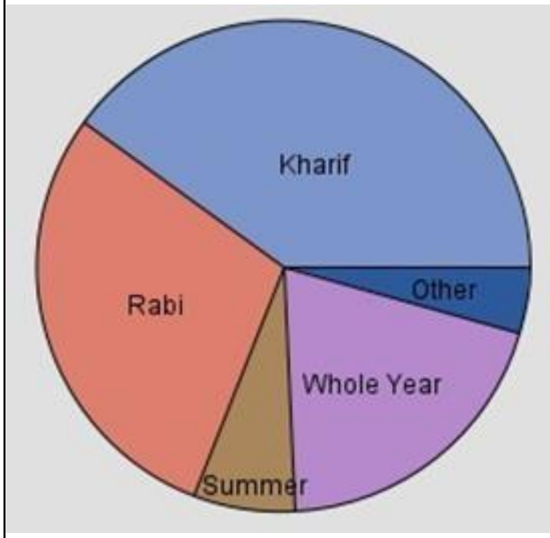
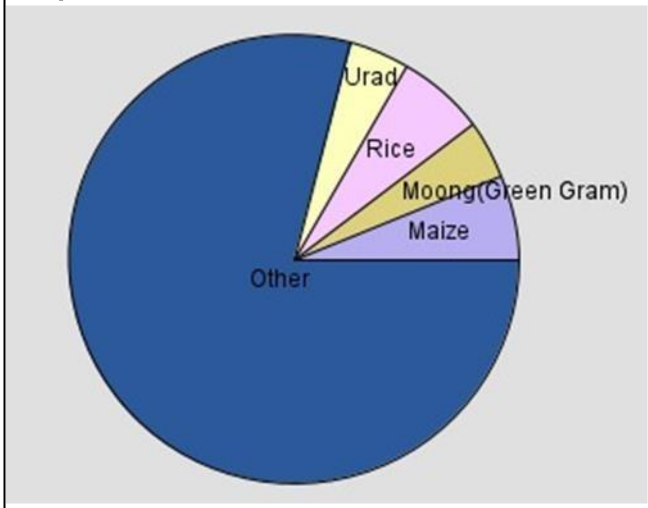
2	**Season** 	1. Crop yields are lowest in the summer. 2. Crop yields are highest in the Kharif.
3	**Crop** 	In the metadata the variable Crop has more than four types Urad, Rice, Maize and Moong, there are other classifications but SAS Enterprise Miner misclassified the remaining classifications as Other.

Table 5.1 Pie charts of variables

The Figure 5.6 bar graph below shows information on which year produced the highest yield. As observed, the highest amount of yield was produced in 2011. And it is observed that there is lowest production in 1987. Besides,they had the same crop production in 2009 and 2011.The trend of overall production is increasing year by year.

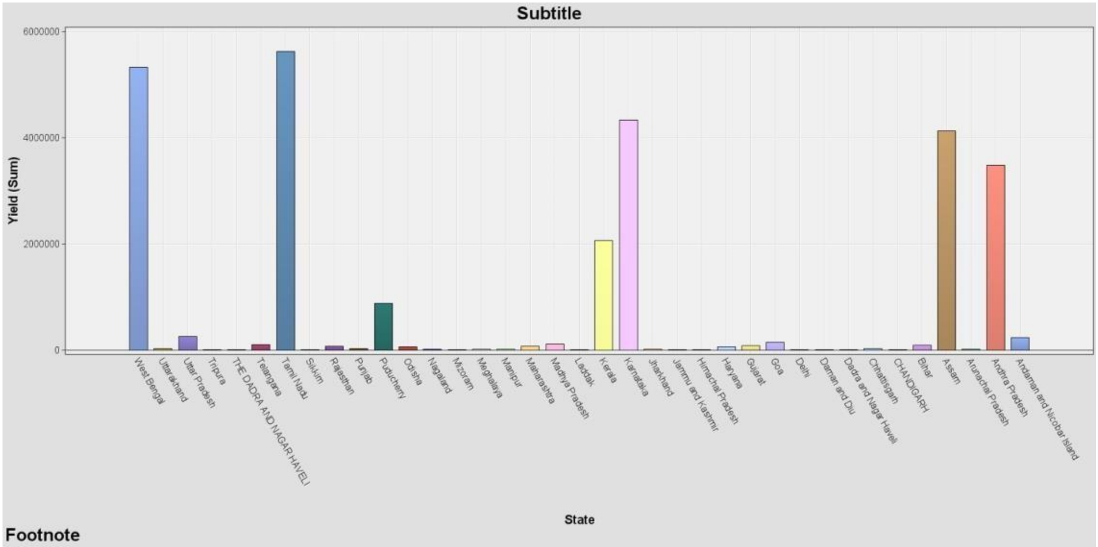


Figure 5.6 Yield （Sum） vs State

The Figure 5.7 bar graph below shows information on which year produced the highest yield. As observed, the highest amount of yield was produced in 2011. And it is observed that there is lowest production in 1997.The trend of overall production is increasing year by year.

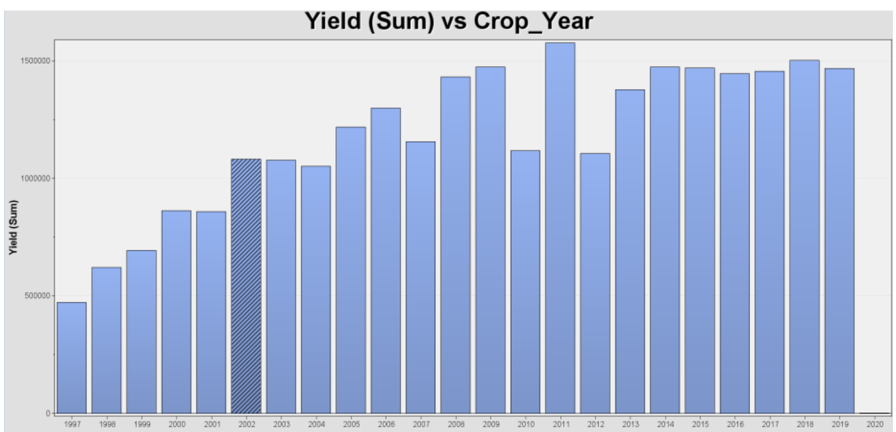


Figure 5.7 Yield (Sum) vs Crop_Year

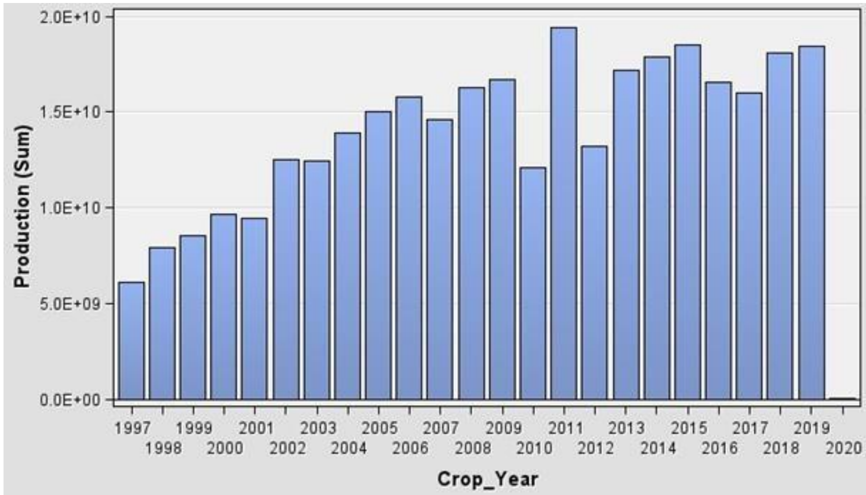
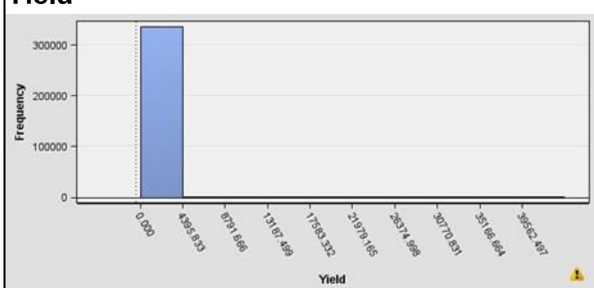
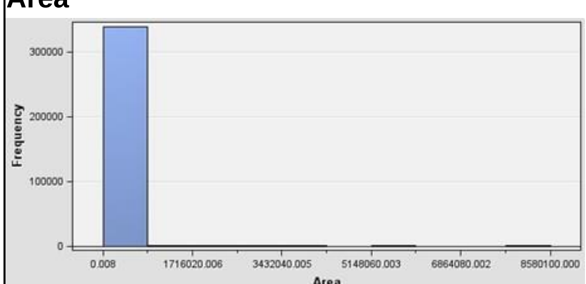


Figure 5.8 Production（Sum）vs Crop-Year

Figure 5.8 above shows us that it is obvious that production peaked in 2011, in contrast to the lowest in 1997. Between 2009 and 2011, production fluctuated the most. Besides, the overall output showed an upward trend. The following Table 5.2 shows more findings.

No.	Variable	Findings
1	**Yield** 	1. The crop yield is mainly between 0 to 4395.833.
2	**Area** 	1. It is obvious that the area in Hectares is mainly between 0.008 to 1608010.003.

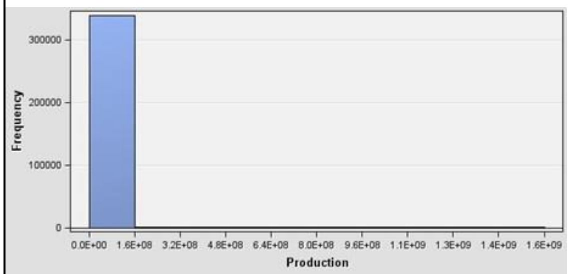
3	**Production** 	1. The production of crops is mainly concentrated in a range between 0 and 1.6x10^8

Table 5.2 Histogram’s findings

R-square (R^2) is a statistic that measures how well the model fits the data. It represents the proportion of the variation in the dependent variable that can be explained by the independent variable in the model. R-square ranges from 0 to 1. The closer the value is to 1, the better the model fits the data. The diagram above shows the model constructed at this point for two variables: Crop and G_Crop.

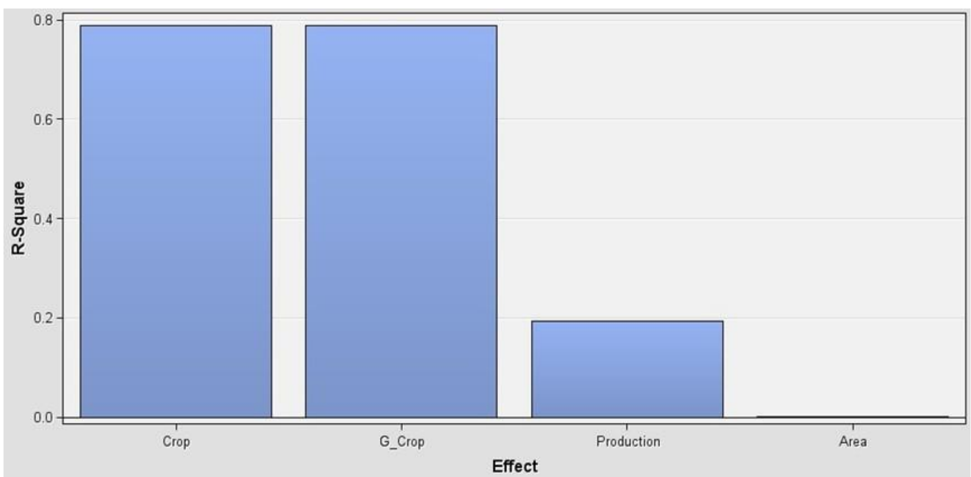


Figure 5.9 R-Square Effect

Sequential R-square is a performance indicator in the multiple linear regression model, which measures the improvement of the explanatory power of the model with the addition of a new argument. It is calculated by comparing the R-square of two models, one containing the arguments X1 and the other containing the arguments X1 and X2. The larger the Sequential R-square, the more explanatory power the newly added argument X2 has on the model. It is observed that the variable G_Crop (Crop variables grouped) has explanatory power on the model.

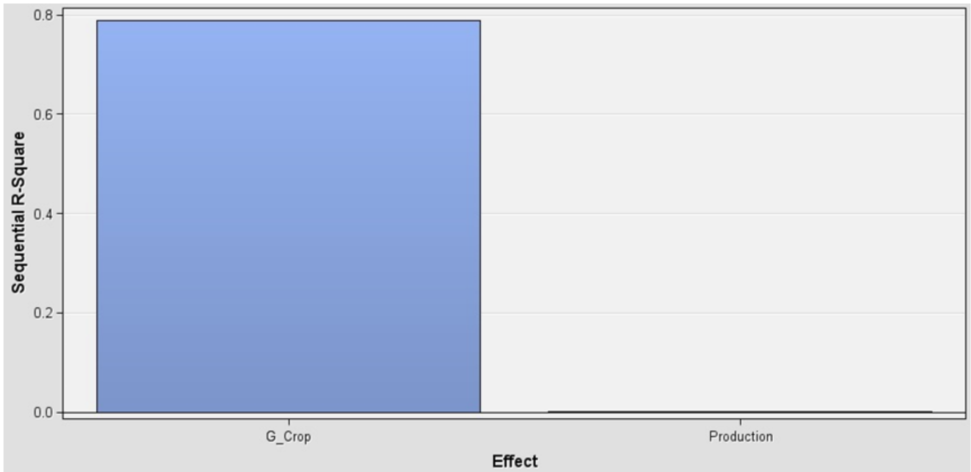


Figure 5.10 Sequential R-Square Effect

Box plots are considered to be one of the best tools for visualizing outliers. Box plots were created for three interval variables to verify the presence of outliers, as shown in the figure 5.11.

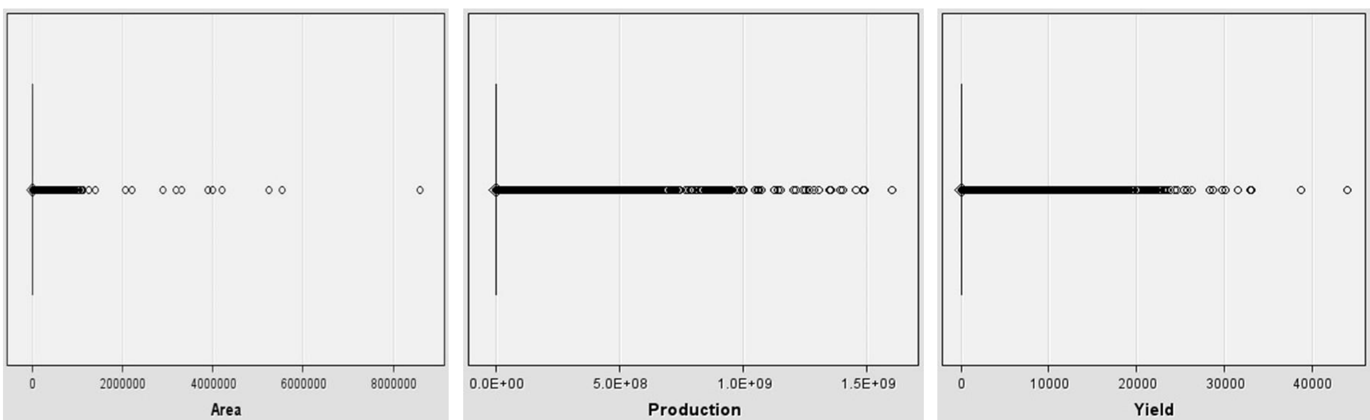


Figure 5.11 Box plots of interval variables

The Figure 5.11 shows the outliers for all the variables detected. The nature of the outliers for the remaining variables needs to be further examined to determine if they are caused by errors or natural behavior before excluding them from the project.

Besides, bivariate analysis refers to the analysis of the relationship between two variables. In bivariate analysis, two variables are said to be related if the value of one affects the value of the other. Charts such as scatter plots, bar plots and box plots can be extremely helpful in finding simple insights. Table 5.3 displayed the findings between variables in the dataset.

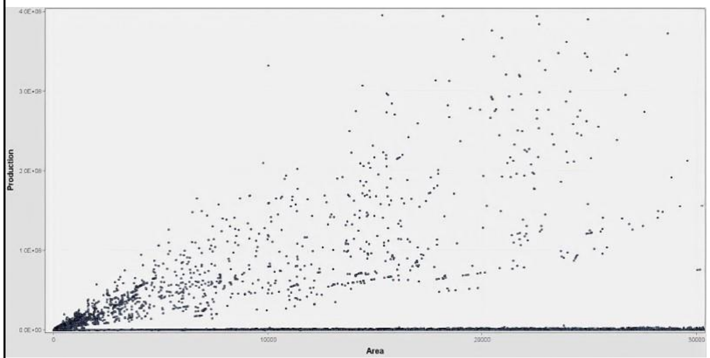
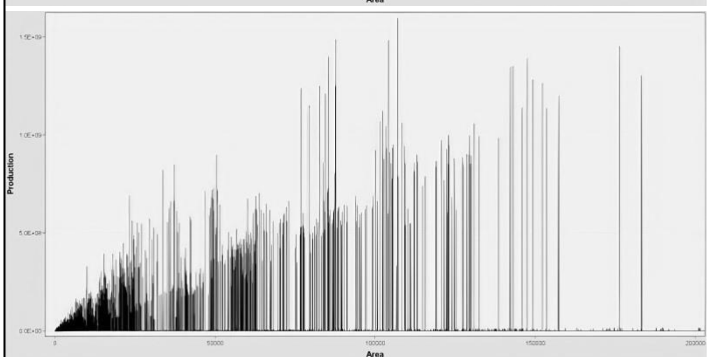
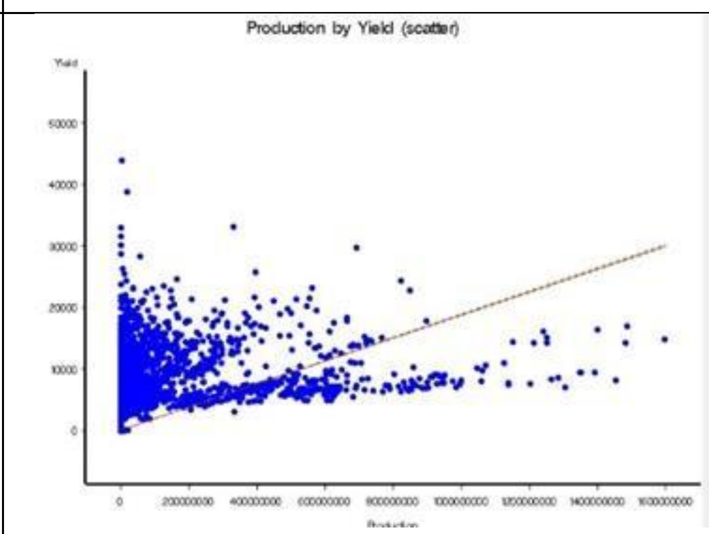
No.	Variable	Findings
1	**Area vs Production**   	The positive correlation between Area and Production in a scatter plot. That is, when the variable Area increases, the variable Production increases as well.
2	 	There's a tendency to be linear between the two variables Production and Yield. According to the available data in the future, the two will show a more obvious linear relationship. Outliers are to be removed for better visualization of the data.

Table 5.3 Bivariate Analysis findings

5.3 Modify

At this stage, we will reduce and transform the data after we explore it. This stage is crucial for further modeling of the data and directly affects the accuracy of the predictive model.

5.3.1 Data Reduction

The Figure 5.12 below shows the amount of data that was excluded as part of the data reduction process to remove rows with missing values.

Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	339771	5566	345337

Figure 5.12 Remove rows with missing values

We removed the missing values. As you can see in Figure 5.13, the missing value is now shown as 0, which proves that the problem has been solved. 5566 rows of observations were removed from 345337 rows of total observations. 98% of data is retained as a result of the data reduction process performed here. The resultant data is used to perform subsequent steps on modeling.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Crop	INPUT	53	0	Rice	6.75	Maize	5.79
TRAIN	Crop_Year	INPUT	23	0	2019	5.08	2016	5.01
TRAIN	District	INPUT	224	0	KADAPA	0.99	VISAKHAPATANAM	0.98
TRAIN	Season	INPUT	6	0	Kharif	38.06	Rabi	30.16
TRAIN	State	INPUT	13	0	Bihar	24.69	Assam	18.17

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Area	INPUT	9030.217	31516.01	100000	0	0.1	519	877029	8.314311	111.621
Production	INPUT	358044.1	10868918	100000	0	0	703	8.9806E8	54.55759	3362.924
Yield	TARGET	82.58753	934.7775	100000	0	0	1.07	43958.33	15.30915	300.3269

Figure 5.13 After performing data reduction

5.3.2 Binned these values

For the next step in our project, the dataset is modified to provide accurate classification results. Transform variable node is used to group our yield values into bins based on quantile. This is because the distribution of the yield is right-skewed. The values were binned using the quantile method to ensure all values are equally distributed. This method is also known as equal height binning.

As shown in Figure 5.14, the dataset will be divided into 10 classes (from lowest to highest) according to yield.

Formatted Value
01:low-0.32
02:0.32-0.5
03:0.5-0.66
04:0.66-0.84
05:0.84-1.02
06:1.02-1.4
07:1.4-2
08:2-3.27
09:3.27-10.76
10:10.76-high

Figure 5.14 The quantified dataset

For classification purposes, we added a yield-based attribute called 'Transformed Yield', labeling the yield low-0.32 as 01 and the yield 10.76-high as 10. The ‘Yield’ column was transformed using binning (quantile approach) into 10 groups as can be seen below.

Obs #	State	District	Crop	Crop_Year	Season	Area	Production	Yield	Transformed Yield
1	Andaman and Nicobar Island	NICOBARS	Arecanut	2007	Kharif	2439.6	3415	1.4	07:1.4-2
2	Andaman and Nicobar Island	NICOBARS	Arecanut	2007	Rabi	1626.4	2277	1.4	07:1.4-2
3	Andaman and Nicobar Island	NICOBARS	Arecanut	2008	Autumn	4147	3060	0.74	04:0.66-0.84
4	Andaman and Nicobar Island	NICOBARS	Arecanut	2008	Summer...	4147	2660	0.64	03:0.5-0.66
5	Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Autumn	4153	3120	0.75	04:0.66-0.84
6	Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Summer...	4153	2080	0.5	03:0.5-0.66
7	Andaman and Nicobar Island	NICOBARS	Arecanut	2000	Kharif	1254	2000	1.59	07:1.4-2
8	Andaman and Nicobar Island	NICOBARS	Arecanut	2001	Kharif	1254	2061	1.64	07:1.4-2
9	Andaman and Nicobar Island	NICOBARS	Arecanut	2002	Whole Y...	1258	2083	1.66	07:1.4-2
10	Andaman and Nicobar Island	NICOBARS	Arecanut	2003	Whole Y...	1261	1525	1.21	06:1.02-1.4
11	Andaman and Nicobar Island	NICOBARS	Arecanut	2004	Whole Y...	1264.7	806	0.64	03:0.5-0.66
12	Andaman and Nicobar Island	NICOBARS	Arecanut	2006	Whole Y...	896	478	0.53	03:0.5-0.66
13	Andaman and Nicobar Island	NICOBARS	Arecanut	2010	Rabi	944	1610	1.71	07:1.4-2
14	Andaman and Nicobar Island	NICOBARS	Arecanut	2011	Rabi	957	1090	1.14	06:1.02-1.4
15	Andaman and Nicobar Island	NICOBARS	Arecanut	2012	Rabi	959	1362	1.42	07:1.4-2
16	Andaman and Nicobar Island	NICOBARS	Arecanut	2013	Rabi	890.5	846	0.95	05:0.84-1.02
17	Andaman and Nicobar Island	NICOBARS	Arecanut	2014	Rabi	876.5	639	0.73	04:0.66-0.84
18	Andaman and Nicobar Island	NICOBARS	Arecanut	2015	Rabi	888.5	83	0.09	01:low-0.32
19	Andaman and Nicobar Island	NICOBARS	Arecanut	2016	Rabi	888.5	99	0.11	01:low-0.32
20	Andaman and Nicobar Island	NICOBARS	Arecanut	2017	Rabi	534.1	125	0.23	01:low-0.32
21	Andaman and Nicobar Island	NICOBARS	Arecanut	2018	Rabi	558	85	0.15	01:low-0.32
22	Andaman and Nicobar Island	NICOBARS	Arecanut	2019	Rabi	612.5	175	0.29	01:low-0.32
23	Andaman and Nicobar Island	NORTH AND MIDDLE A...	Arecanut	2000	Kharif	3100	5200	1.68	07:1.4-2
24	Andaman and Nicobar Island	NORTH AND MIDDLE A...	Arecanut	2001	Kharif	3100	5239	1.69	07:1.4-2
25	Andaman and Nicobar Island	NORTH AND MIDDLE A...	Arecanut	2006	Whole Y...	1160	3012	2.6	08:2-3.27

Figure 5.15 Create ‘Transformed Yield’ variable

As can be seen in Figure 5.16, the 'Transformed Yield' variable has a missing value of 0. And with this bar chart, it is possible to visualize the distribution of the rating rates for different yield values. Yield values classified as 02: 0.32-0.5 appear with the highest rating rate, while yield values classified as 05: 0.84-1.02 appear with the lowest frequency.

5.3.3 Creating Training and Validation Data

By using the Data Partition node, we are able to split the dataset to create Training and Validation sets. The dataset was split into training and validation sets using a 60:40 ratio, where 60% of the data was allocated to the training set, and the remaining 40% was assigned to the validation set. This division ensured that a substantial portion of the data was utilized for training the model, while still reserving a separate portion for evaluating its performance on unseen data.

The split ratio for this project is 60:40 on Training and Validation as shown on the figure below:

General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes

Figure 5.16 Data Set Allocations

Below Figure 5.17 is the diagram showing the frequency of data split into training and validation sets before proceeding into the classification stage.

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
PCTL_Yield	.	01:low-0.32	32929	9.6915	Transformed Yield
PCTL_Yield	.	02:0.32-0.5	33706	9.9202	Transformed Yield
PCTL_Yield	.	03:0.5-0.66	35038	10.3122	Transformed Yield
PCTL_Yield	.	04:0.66-0.84	32876	9.6759	Transformed Yield
PCTL_Yield	.	05:0.84-1.02	34555	10.1701	Transformed Yield
PCTL_Yield	.	06:1.02-1.4	34308	10.0974	Transformed Yield
PCTL_Yield	.	07:1.4-2	32805	9.6550	Transformed Yield
PCTL_Yield	.	08:2-3.27	35493	10.4462	Transformed Yield
PCTL_Yield	.	09:3.27-10.76	34079	10.0300	Transformed Yield
PCTL_Yield	.	10:10.76-high	33982	10.0014	Transformed Yield

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
PCTL_Yield	.	01:low-0.32	19757	9.6916	Transformed Yield
PCTL_Yield	.	02:0.32-0.5	20224	9.9207	Transformed Yield
PCTL_Yield	.	03:0.5-0.66	21022	10.3121	Transformed Yield
PCTL_Yield	.	04:0.66-0.84	19725	9.6759	Transformed Yield
PCTL_Yield	.	05:0.84-1.02	20733	10.1704	Transformed Yield
PCTL_Yield	.	06:1.02-1.4	20585	10.0978	Transformed Yield
PCTL_Yield	.	07:1.4-2	19682	9.6548	Transformed Yield
PCTL_Yield	.	08:2-3.27	21295	10.4460	Transformed Yield
PCTL_Yield	.	09:3.27-10.76	20446	10.0296	Transformed Yield
PCTL_Yield	.	10:10.76-high	20388	10.0011	Transformed Yield

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
PCTL_Yield	.	01:low-0.32	13172	9.6914	Transformed Yield
PCTL_Yield	.	02:0.32-0.5	13482	9.9195	Transformed Yield
PCTL_Yield	.	03:0.5-0.66	14016	10.3124	Transformed Yield
PCTL_Yield	.	04:0.66-0.84	13151	9.6760	Transformed Yield
PCTL_Yield	.	05:0.84-1.02	13822	10.1697	Transformed Yield
PCTL_Yield	.	06:1.02-1.4	13723	10.0968	Transformed Yield
PCTL_Yield	.	07:1.4-2	13123	9.6554	Transformed Yield
PCTL_Yield	.	08:2-3.27	14198	10.4463	Transformed Yield
PCTL_Yield	.	09:3.27-10.76	13633	10.0306	Transformed Yield
PCTL_Yield	.	10:10.76-high	13594	10.0019	Transformed Yield

Figure 5.17 Report Output for Data Partition node

5.4 Model

After the data was divided into a training and validation set in a 60:40 ratio, the training data was used to build several models.

Once the data is split, 3 different decision tree models are implemented on the data. A decision tree is a supervised model based on a tree structure consisting of a root node, internal nodes, branches and leaf nodes. In each node, decision rules are generated to classify targets in order to solve the classification task.

Each decision tree model had 2, 3 and 4 branches configured respectively. Below description shows information such as True Positive & Negatives and False Positives & Negatives values extracted from the decision tree models executed.

5.4.1 Constructing a Decision Tree Model had 2 branches

Event Classification Table

Data Role=TRAIN Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
4524	181730	1739	15864

Data Role=VALIDATE Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
3060	121124	1196	10534

Figure 5.18 Decision Tree 1 (2 branches)

- Accuracy: The model achieved an accuracy of **96.86%**, demonstrating a high level of overall correctness in its predictions.
- Precision: The model achieved a precision of **89.80%**, suggesting that 89.80% of the samples predicted as positive were actually positive.
- Recall: The model achieved a recall of **77.49%**, showing its ability to identify a significant proportion of the actual positive samples.
- F1 Score: The model achieved an F1 score of **0.83**, indicating a balanced performance between precision and recall.

5.4.2 Constructing a Decision Tree Model had 3 branches

Event Classification Table

Data Role=TRAIN Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
4068	182067	1402	16320

Data Role=VALIDATE Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
2864	121361	959	10730

Figure 5.19 Decision Tree 2 (3 branches)

- Accuracy: The model achieved an accuracy of **97.18%**, demonstrating a high level of overall correctness in its predictions.
- Precision: The model achieved a precision of **91.79%**, suggesting that 91.79% of the samples predicted as positive were actually positive.
- Recall: The model achieved a recall of **78.93%**, showing its ability to identify a significant proportion of the actual positive samples.
- F1 Score: The model achieved an F1 score of **0.84**, indicating a balanced performance between precision and recall.

5.4.3 Constructing a Decision Tree Model had 4 branches

Event Classification Table

Data Role=TRAIN Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
3340	181837	1632	17048

Data Role=VALIDATE Target=PCTL_Yield Target Label=Transformed Yield

False Negative	True Negative	False Positive	True Positive
2392	121216	1104	11202

Figure 5.20 Decision Tree 3 (4 branches)

- Accuracy: The model achieved an accuracy of **97.42%**, demonstrating a high level of overall correctness in its predictions.
- Precision: The model achieved a precision of **91.02%**, suggesting that 91.02% of the samples predicted as positive were actually positive.

- Recall: The model achieved a recall of **82.40%**, showing its ability to identify a significant proportion of the actual positive samples.
- F1 Score: The model achieved an F1 score of **0.86**, indicating a balanced performance between precision and recall.

After the calculation of decision tree models, we can find that Decision Tree 1 and Decision Tree 2 demonstrate good overall accuracy and precision. However, their recall is slightly lower at 80%, indicating that they may miss some actual positive samples. On the other hand, Decision Tree 3 achieves a higher accuracy of 97.42% and shows better recall at 82.40%. Besides, its precision is 91.02%, and it maintains a good balance between precision and recall with an F1 score of 0.86.

Considering the trade-offs between precision and recall , Decision Tree 3 appears to be the best model. Its higher recall suggests a stronger ability to identify positive samples, which could be crucial in applications where false negatives are undesirable. Therefore, based on the performance metrics and the trade-offs involved, Decision Tree 3 is the best model to be implemented in our project due to its higher accuracy, recall, and reasonable precision make it the best-performing model for the given problem.

The variable importance for all decision trees are acquired as below. It is found that the two variables that have highest importance are Crop and State. Besides, according to the table, the ratio is close to one, this suggests that variable's predictive power remains stable and can be relied upon when applying the model to new data.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Crop		15	1.0000	1.0000	1.0000
State		14	0.5787	0.5742	0.9923
Production		9	0.1747	0.1665	0.9528
Crop_Year		2	0.1137	0.1175	1.0329
Area		1	0.0804	0.0703	0.8745

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Crop		33	1.0000	1.0000	1.0000
State		44	0.6882	0.6801	0.9883
Production		48	0.3456	0.3356	0.9710
Area		33	0.2294	0.2093	0.9123
Crop_Year		14	0.2046	0.1924	0.9404
Season		9	0.1528	0.1432	0.9373

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Crop		58	1.0000	1.0000	1.0000
State		74	0.6833	0.6786	0.9932
Production		126	0.4604	0.4299	0.9338
Area		117	0.4076	0.3739	0.9173
Crop_Year		65	0.3351	0.3071	0.9165
Season		21	0.1482	0.1371	0.9247

Figure 5.21 Decision Tree Variable Importance across all decision tree mode

5.5 Assess

Finally, all 3 decision models were compared using the model comparison node to evaluate the performance of each decision tree model.

Low misclassification rates in Decision Tree 3 compared to other models show that decision trees with 4 branches configuration performed the best to show more accurate predictions compared to using 2 and 3 branches.

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree3	Decision Tree 3	0.48331	0.057030	0.46969	0.058676
	Tree	Decision Tree 2	0.53677	0.063425	0.53060	0.064120
	Tree2	Decision Tree 1	0.61140	0.070268	0.60742	0.070581

Figure 5.22 Model Comparison

6 Conclusion

Initially, the dataset contained 4 interval variables and 4 nominal variables. After manually modifying the metadata, the output is shown in Table 6.1.

Role	Type of Variable	Count
Input	Interval	2
Input	Nominal	5
Target	Interval	1

Table 6.1 Reversed Metadata

It was observed that all variables in this dataset had missing values and the count of rows with missing values was insignificant compared to the total size of the dataset. After the data reduction and transformation process, 98% of the data were retained and utilized for further steps in modeling.

Through the modeling and assessment phase of SEMMA, we built three different decision tree models, each configured with 2, 3 and 4 branches respectively. Our subsequent results found that the decision tree model with 4 branches performed best and showed more accurate predictions.

Below is an overview of the SEMMA process implemented in our project.

SEMMA	Summarize
Sample	At this stage, we selected a representative dataset on agricultural production statistics in India, manually adjusted the roles and variables in the dataset and found that there was noisy data such as missing values that needed further processing.
Explore	In the exploration phase, we use visualization tools such as pie charts, bar charts and box plots to explore the data set. We found that: 1. Overall crop yields showed an increasing trend from year to year.

	2. There is a positive correlation between area and yield.
Modify	At this stage, we performed data reduction, removed all missing values, and binned the dataset. These operations ensure the accuracy and consistency of the data.
Model	After dividing the dataset into training and validation sets in the ratio of 60:40, we built three decision tree models with 2,3 and 4 branches respectively. Crop is selected as the root node as it has a high level of information gain.
Assess	We evaluated the performance of each decision tree model and found that the decision tree model configured with four branches performed best and had the lowest misclassification rate.

Table 6.2 Summarize

Objective	Key Findings
A. To analyze trends and patterns in our dataset.	Tamil Nadu and West Bengal had highest crop yields Production and yield were the highest in the year 2011 Crop variable showed high predicting power in variable analysis High occurrences of variables in dataset (State: Uttar Pradesh, Season: Kharif, Crop: Rice) As expected, positive correlation was observed between production and yield

	variable
B. Determine the factors that have the greatest influence on crop yield	**Crop** and **State** are identified as the primary factors exerting significant influence on crop yield based on the machine learning models constructed.
C. To detect which model performed the best	Considering the performance metrics and the inherent trade-offs, the implementation of **Decision Tree 3** emerges as the optimal model for our project. This determination is substantiated by its superior accuracy, recall, and commendable precision, establishing it as the most proficient model for addressing the specific problem at hand.

Table 6.3 Key Findings for Objective A,B,C

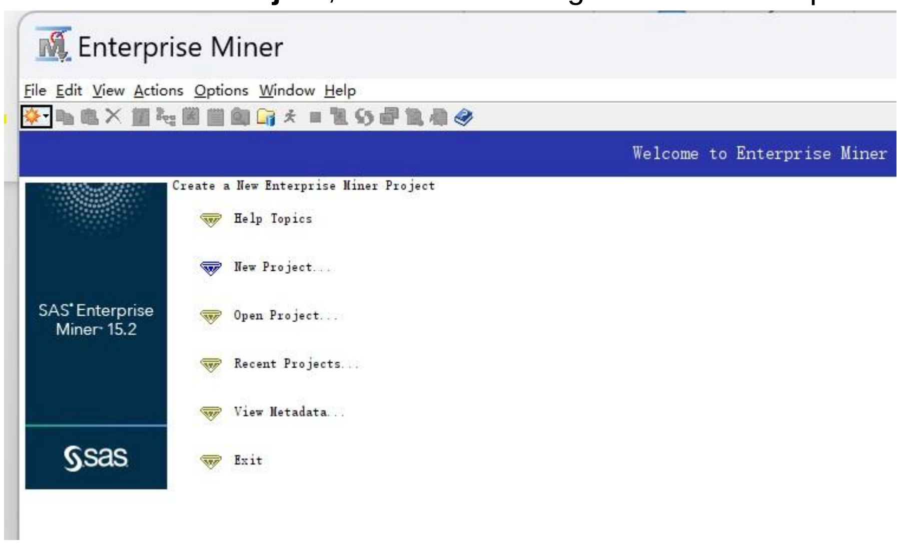
7 Appendix

Attached below are the procedures for performing the all steps of SEMMA in SAS Enterprise Miner - Sample, Explore,Modify, Model and Assess.

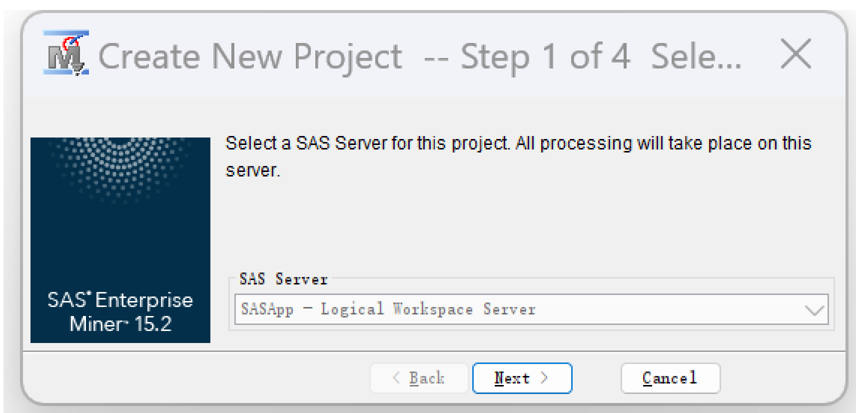
A.1 SEMMA first process - Sample

A.1.1 Create a new Enterprise Miner Project

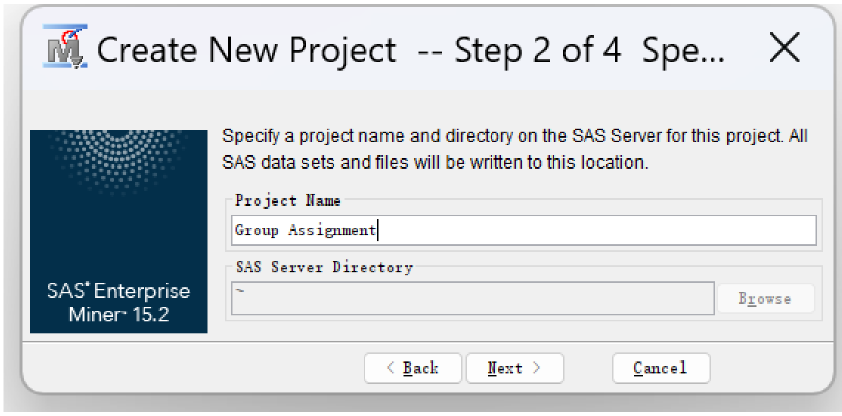
- Click create **New Project**, and the following window show up.



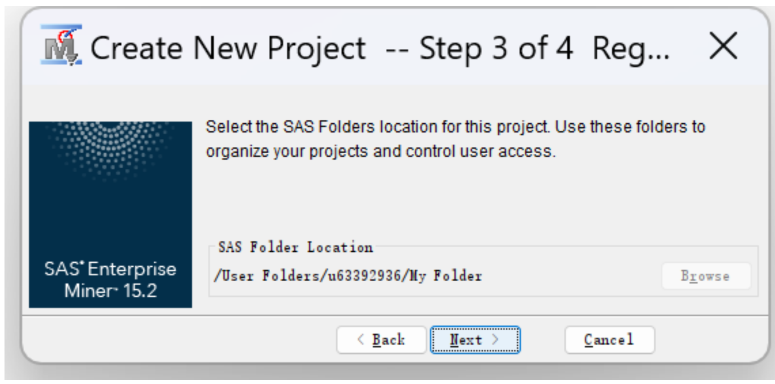
- Click **Next**.



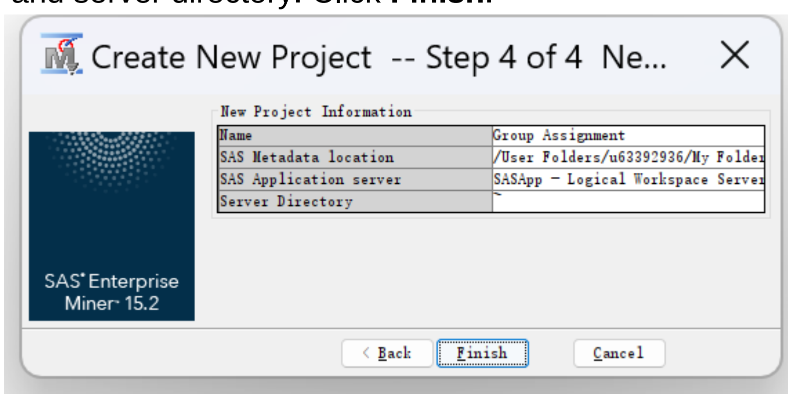
- Enter desired project name. Click **Next**.



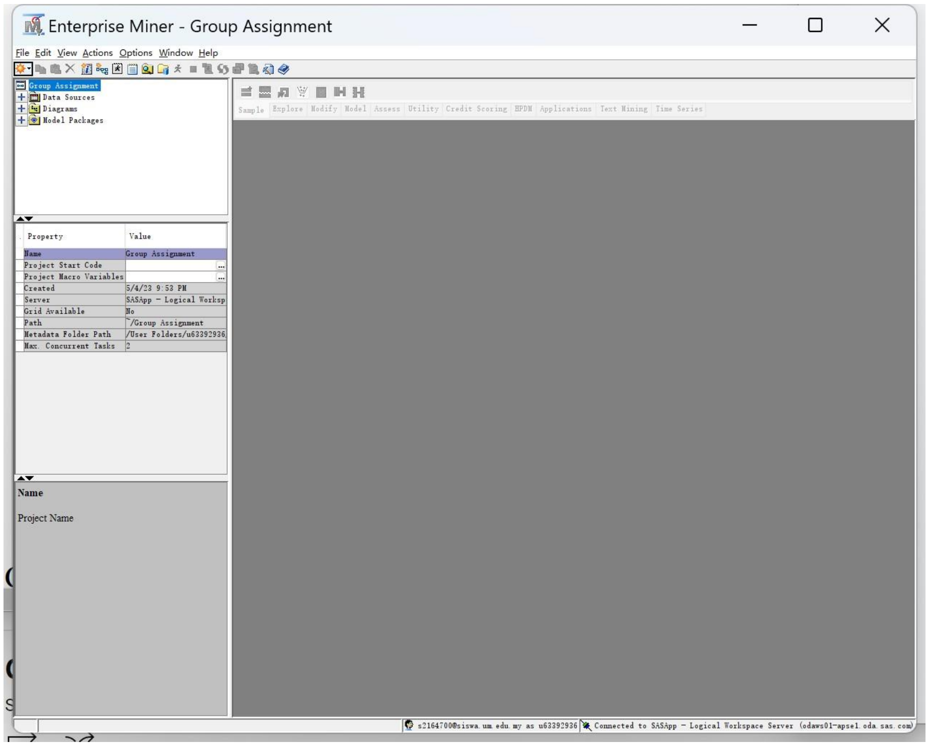
- Click **Next**.



- Now we can see that the New Project Information includes the project name and server directory. Click **Finish**.

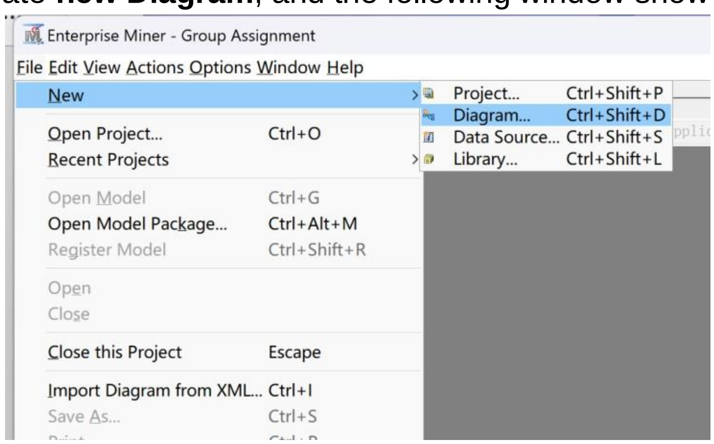


- Now completed the creation of new Enterprise Miner Project

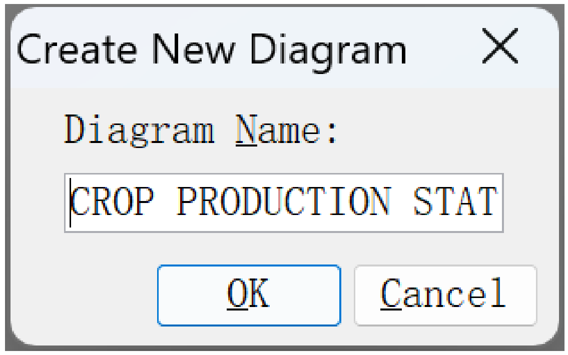


A.1.2 Create a diagram

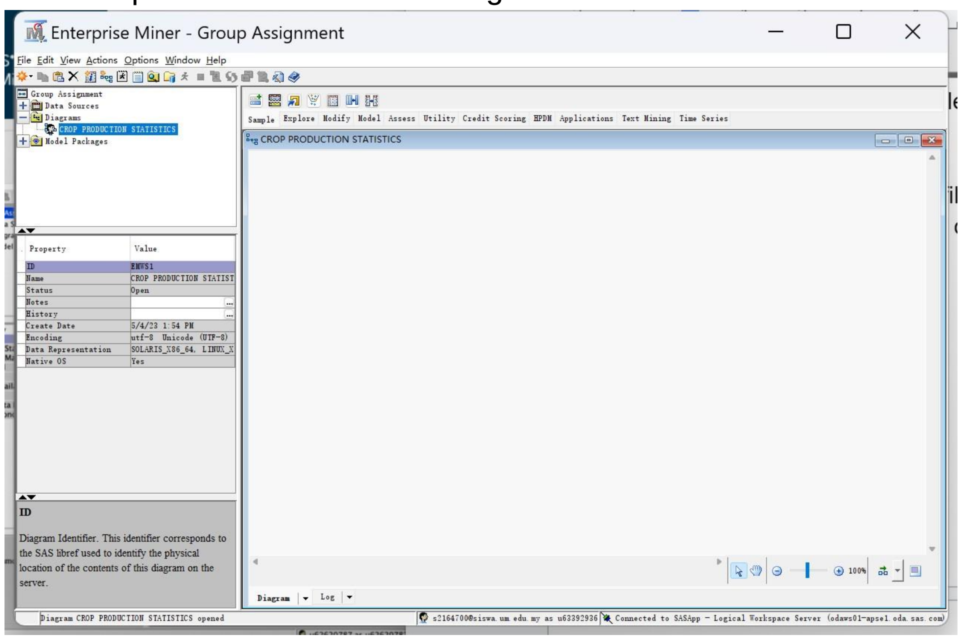
- Click create **new Diagram**, and the following window show up.



- Input the Diagram Name. Click **OK**.

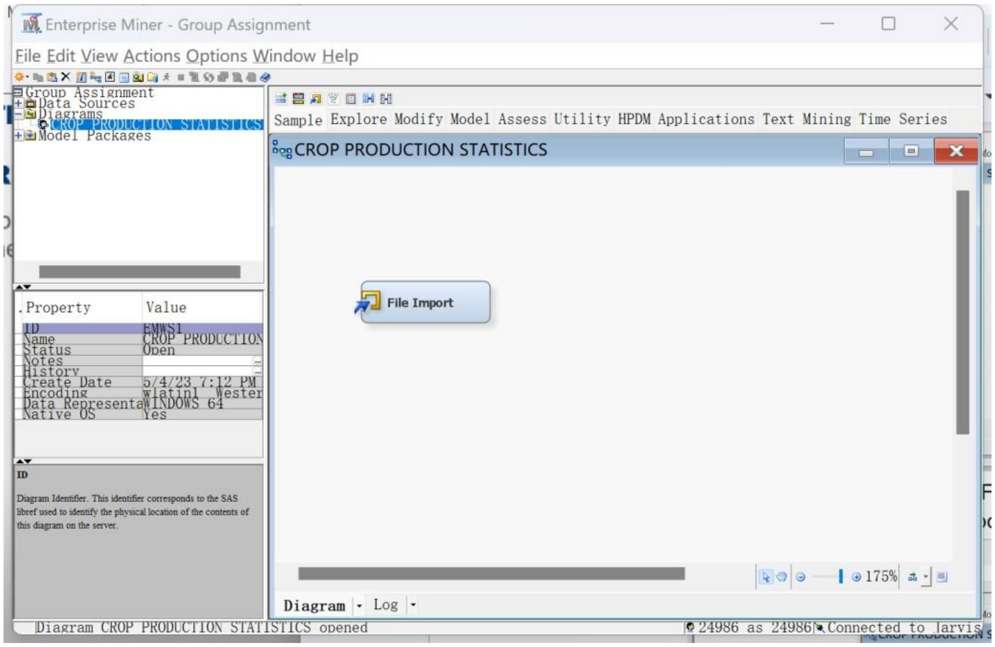


- Now completed the creation of Diagram.

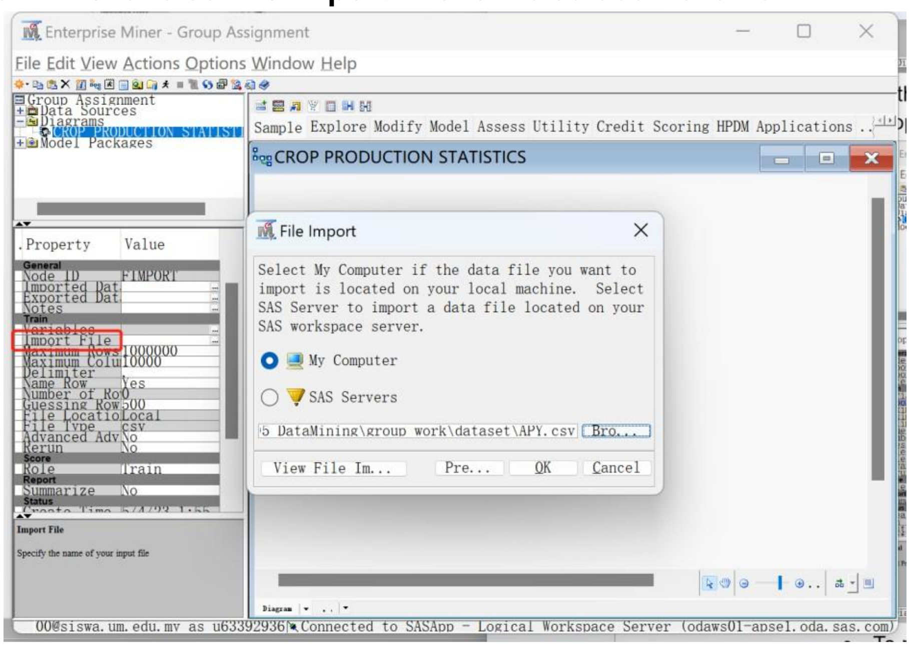


A.1.3 Import data as a file and save as SAS file

- Select and drag the **File Import** node onto the diagram workspace.



- In the Properties Panel for the **File Import** node, and then use the drop-down menu to set the **Import File** to the dataset. Click OK.



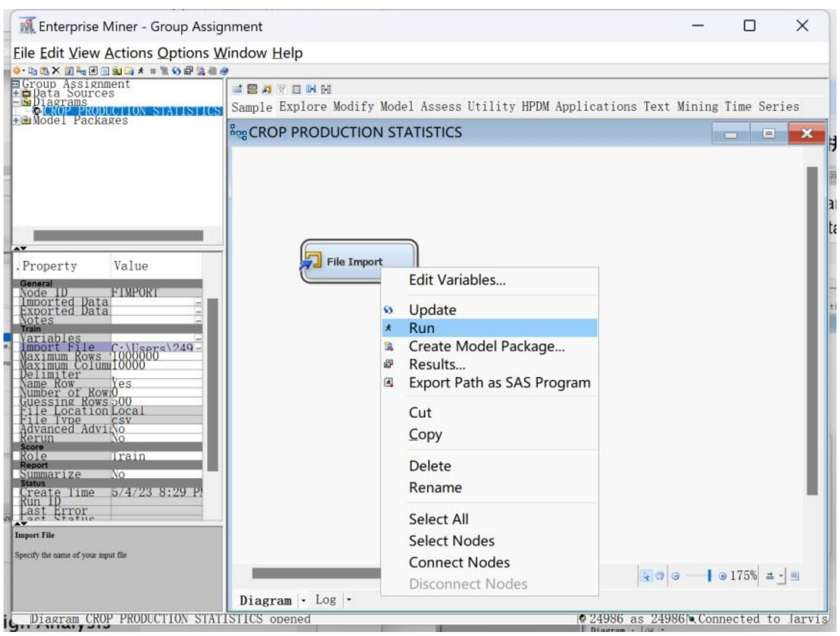
- The number of rows are inserted into the dataset in the “Maximum Rows to Import” option. The File import node is run

General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	C:\Users\Asus ...
Maximum Rows to Import	345337
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train

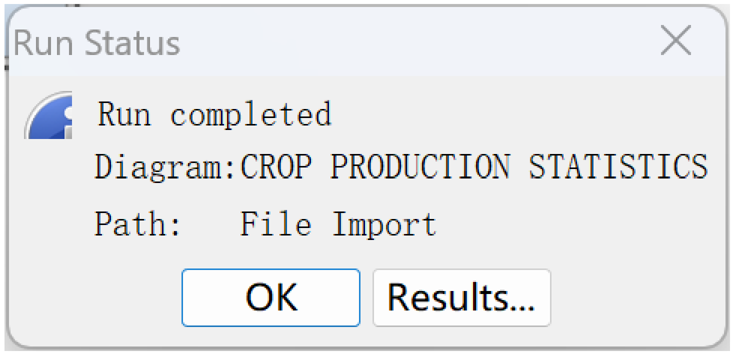
- The variables levels are predefined appropriately. "Crop Year" data level is changed to nominal and "Yield" data role is changed to target.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Area	Input	Interval	No		No	.	.
Crop	Input	Nominal	No		No	.	.
Crop_Year	Input	Nominal	No		No	.	.
District	Input	Nominal	No		No	.	.
Production	Input	Interval	No		No	.	.
Season	Input	Nominal	No		No	.	.
State	Input	Nominal	No		No	.	.
Yield	Target	Interval	No		No	.	.

- To run the **File Import** node, right-click it in the diagram workspace, and click **Run** from the menu.



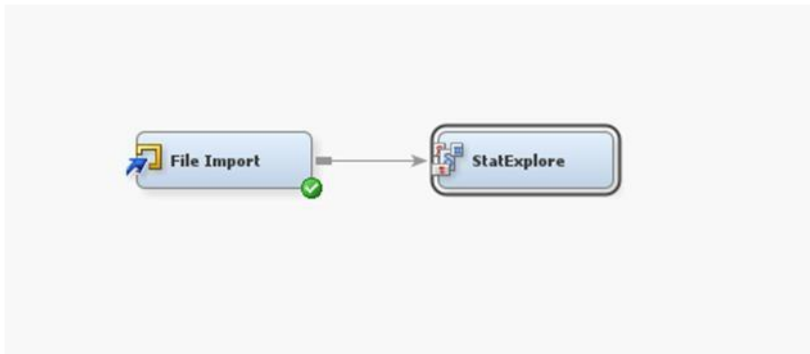
- Run completed. Click **OK** or results.



A.2 SEMMA second process - Explore

A.2.1 Set a sample size to be used for creating graphs

- Add **StatExplore** node to check for missing values

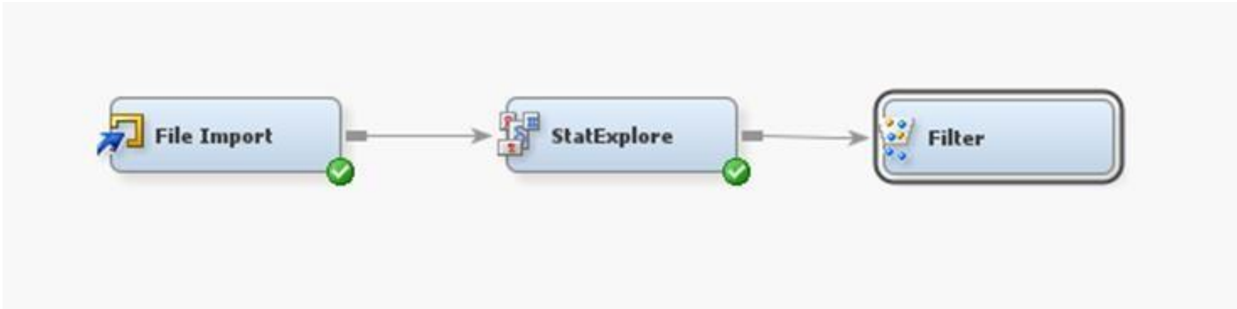


- Configure **StatExplore** node settings as below and run the node

General	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Data	
Number of Observations	ALL
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	...
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	10000

Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	10000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	Yes
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	5/5/23 2:18 PM
Run ID	cc2a3152-7f3d-414a-bff7-3b
Last Error	
Last Status	Complete
Last Run Time	5/7/23 2:33 PM
Run Duration	0 Hr. 1 Min. 10.59 Sec.
Grid Host	
User-Added Node	No

- Next, to remove rows with missing values, the filter node is added into the diagram and pipeline is connected

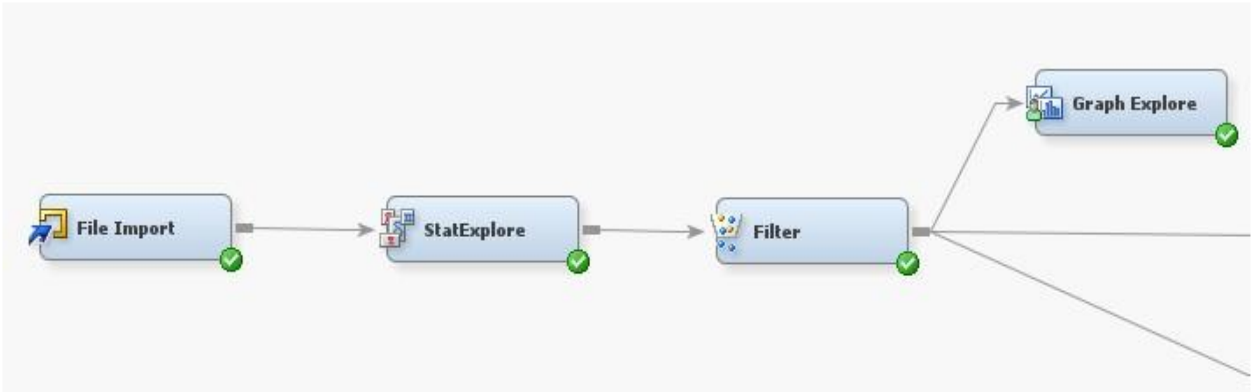


- The below options are configured for class and interval variables. Once configured, the node is run to check results.

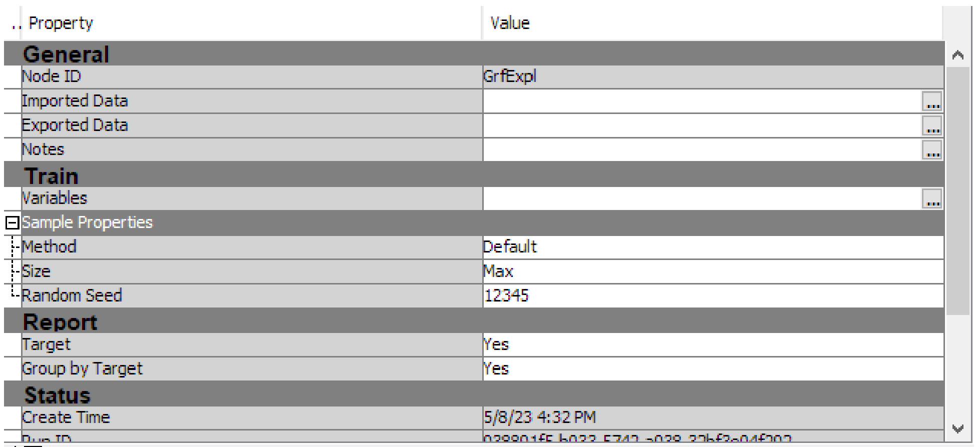
Property	Value
Train	
Export Table	Filtered
Tables to Filter	All Data Sets
Distribution Data Sets	No
Class Variables	
Class Variables	...
Default Filtering Method	None
Keep Missing Values	No
Normalized Values	No
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cut	25
Interval Variables	
Interval Variables	...
Default Filtering Method	None
Keep Missing Values	No
Tuning Parameters	...
Score	
Create Score Code	Yes

A.2.2 To create graph

- Add **GraphExplore** node to analyze the data based on the graph

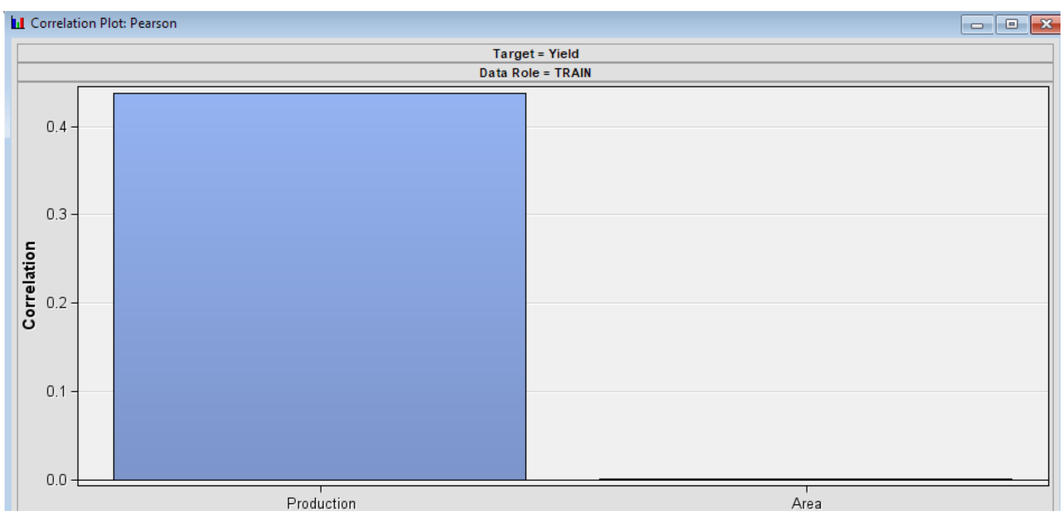


- Configure **Graph Explore** node settings as below and run the node



Property	Value
General	
Node ID	GrfExpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Sample Properties	
Method	Default
Size	Max
Random Seed	12345
Report	
Target	Yes
Group by Target	Yes
Status	
Create Time	5/8/23 4:32 PM
Run ID	[illegible]

- Then all the graphs are being done using plot in the result
- For Correlation Analysis, **StatExplore** is added
- Select Run > Results.. > View > Plot > Correlation Plot :Pearson



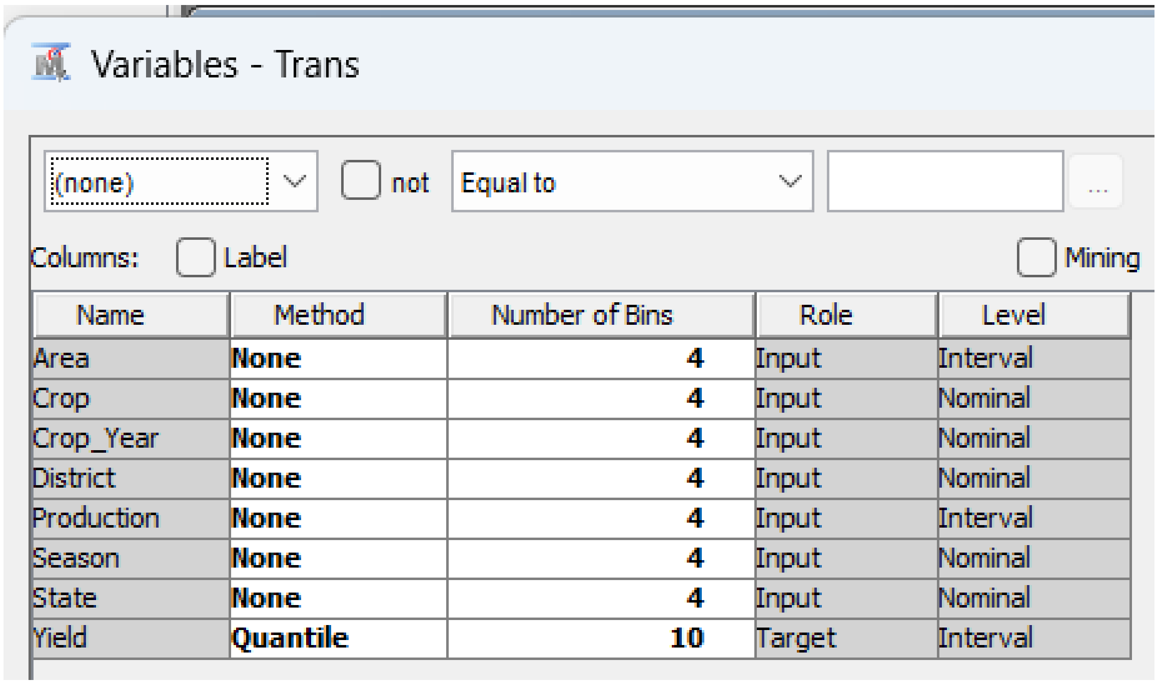
- For Variable Worth, **Variable Clustering** is added
- Select Run > Results.. > View > Plot > Correlation Plot :Pearson

A.3 SEMMA third process - Modify

A.3.1 Add the Transform variables

- Transform variables node is added

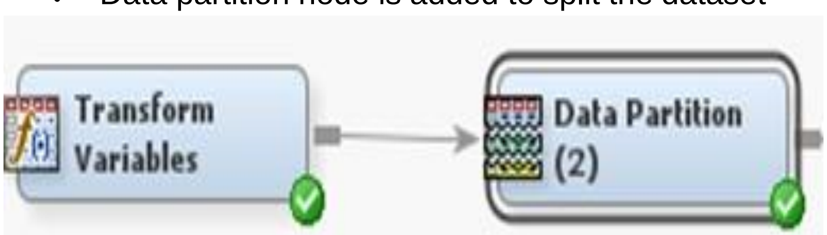
- The yield column values are to be grouped into bins of 10 for classification. Remaining columns are not changed. The node is run.



Name	Method	Number of Bins	Role	Level
Area	None	4	Input	Interval
Crop	None	4	Input	Nominal
Crop_Year	None	4	Input	Nominal
District	None	4	Input	Nominal
Production	None	4	Input	Interval
Season	None	4	Input	Nominal
State	None	4	Input	Nominal
Yield	Quantile	10	Target	Interval

A.3.2 Data Partition

- Data partition node is added to split the dataset



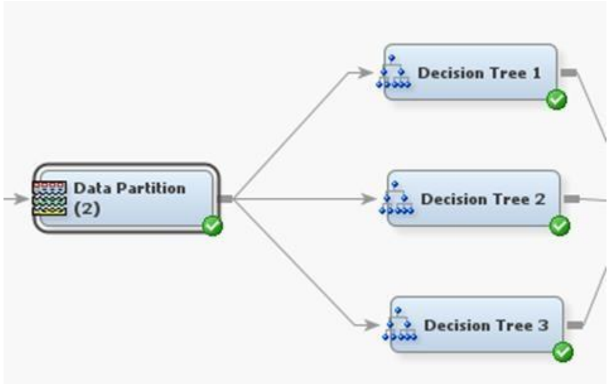
- The dataset is split into 60% training and 40% validation. The node is run.

General	
Node ID	Part2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
-Training	60.0
-Validation	40.0
-Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	6/15/23 2:25 PM
Run ID	c574842e-cf80-a34d-a9ea-85
Last Error	
Last Status	Complete
Last Run Time	6/17/23 3:25 PM
Run Duration	0 Hr. 4 Min. 17.07 Sec.
Grid Host	
User-Added Node	No

A.4 SEMMA fourth process - <u>Model</u>

A.4.1 Decision Tree Model

- After the dataset is split, 3 different decision tree nodes are added



- Each decision tree model is updated with branch values of 2, 3 and 4 respectively. The decision tree nodes are run.

General	
Node ID	Tree2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

General	
Node ID	Tree3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	4
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

A.5 SEMMA fifth process - Assess

A.5.1 Compare between Decision Tree1, Decision Tree2, Decision Tree 3

- Model comparison node is added and run to evaluate performance comparison between the 3 decision tree models.