Data analysis using KNIME Analytics for Brazil dataset

By Shruthi Rajeshwari G S

1. Introduction

This report examines the factors influencing customer satisfaction in Brazil using a dataset rich in demographic characteristics, income levels, consumer preferences, and awareness of technological advancements. By analyzing these variables, this study aims to uncover key insights into what drives customer satisfaction among different population segments.

The primary objective of this analysis is to predict customer satisfaction, providing a framework for identifying which factors most significantly impact satisfaction levels. The KNIME Analytics Platform was chosen for its robust capabilities in data preprocessing, visualization, and machine learning. The analysis will employ decision trees, random forests, and gradient boosting to identify patterns and determine the most effective predictive model for customer satisfaction.

Findings from this report are expected to shed light on critical drivers of customer satisfaction, enabling stakeholders to develop targeted strategies that align with consumer preferences and demographic trends. By comparing the performance of various predictive models, this report also highlights the strengths of each approach and suggests best practices for future applications in customer satisfaction prediction.

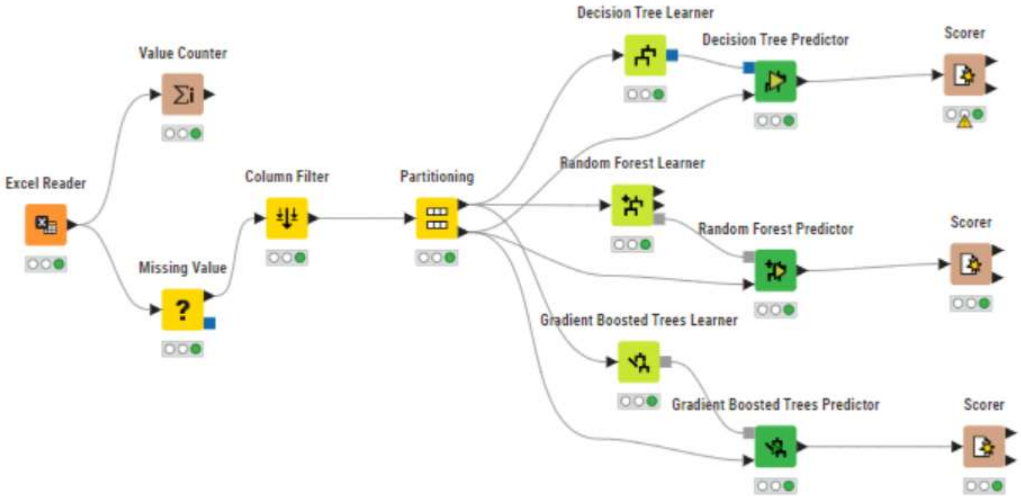


Fig 1.1 KNIME Analytics workspace

2. Project Goal and Objectives

Goal:

To analyze and predict customer satisfaction in Brazil based on demographic and socioeconomic factors, offering insights to help enhance consumer experiences.

Objectives:

1. **Data Preparation**
   - Explore and preprocess the dataset to ensure data quality and relevance.
2. **Feature Analysis**
   - Identify key patterns and relationships between customer satisfaction and other variables.
3. **Predictive Modeling**
   - Develop and compare machine learning models (Decision Tree, Random Forest, Gradient Boosting) to predict customer satisfaction.
4. **Model Evaluation and Insights**
   - Evaluate models to identify the best predictor of satisfaction and understand influential factors.
5. **Recommendations**

- Provide actionable insights and suggest future analysis directions for deeper understanding of customer satisfaction.

3. Data Understanding

The Brazil dataset provides a snapshot of various demographic and socioeconomic attributes, allowing for an analysis of factors influencing customer satisfaction. Key variables in the dataset include:

1. **Demographic Characteristics**
   - Age: Age groups or exact ages of respondents.
   - Gender: Categorical data showing the gender distribution.
   - Education Level: Represents the respondents' highest level of education, which may correlate with consumer preferences and satisfaction.
2. **Income Levels**
   - Income Brackets: Income data segmented into categories, providing insights into financial status and its effect on satisfaction levels.

3. **Consumer Preferences**
    - Product and Service Preferences: Categories reflecting consumer interest in certain products or services, potentially linked to satisfaction.
4. **Technological Awareness**
    - Familiarity with Technology: Measures the respondents' awareness or usage of technology, which could influence satisfaction in tech-driven services.
5. **Customer Satisfaction (Target Variable)**
    - Satisfaction Rating: The outcome variable used for prediction, likely represented on a scale (e.g., from dissatisfied to very satisfied), allowing for analysis of the factors that contribute to higher satisfaction scores.

This data provides the foundation for univariate and bivariate analyses, where patterns and relationships between customer satisfaction and these variables can be identified, guiding model development and predictive insights.

4. Correlation Coefficient Analysis

- To understand relationships between various features in the dataset, we performed a linear correlation analysis. The correlation coefficient values are displayed in the correlation matrix, where the intensity of color indicates the strength and direction of relationships between variables. Key points from the analysis are:
- **Strong Positive Correlations**: There are clusters of variables showing strong positive correlations (dark blue squares), particularly among features related to user satisfaction and technological awareness (e.g., ease_of_use_s, Trust_s, online_satisfaction, and loyalty_s). This suggests that satisfaction levels may be closely linked with factors like trust and ease of use.
- **Weak or No Correlation**: Variables like gender, age, education, and certain consumer preferences (e.g., most_f_Grocery, most_f_HealthCare) show weak or no correlation with other features, indicated by the lack of color in those matrix sections. This implies that these demographic and preference variables may not have a significant linear relationship with satisfaction outcomes.
- **Noteworthy Patterns**: The matrix shows clusters related to technology awareness tools (aware_tools_* features), which could indicate a relationship between familiarity with these technologies and satisfaction levels, though specific correlation strengths may vary.
- **Interpretation Limitations**: Linear correlation coefficients only capture linear relationships between variables. Thus, while this matrix highlights potential relationships, further analysis (e.g., logistic regression or machine learning

models) is needed to fully understand the predictive impact of these variables on customer satisfaction.

This correlation analysis serves as an initial exploratory step to identify patterns and relationships in the data, guiding further steps in predictive modeling and deeper investigation.

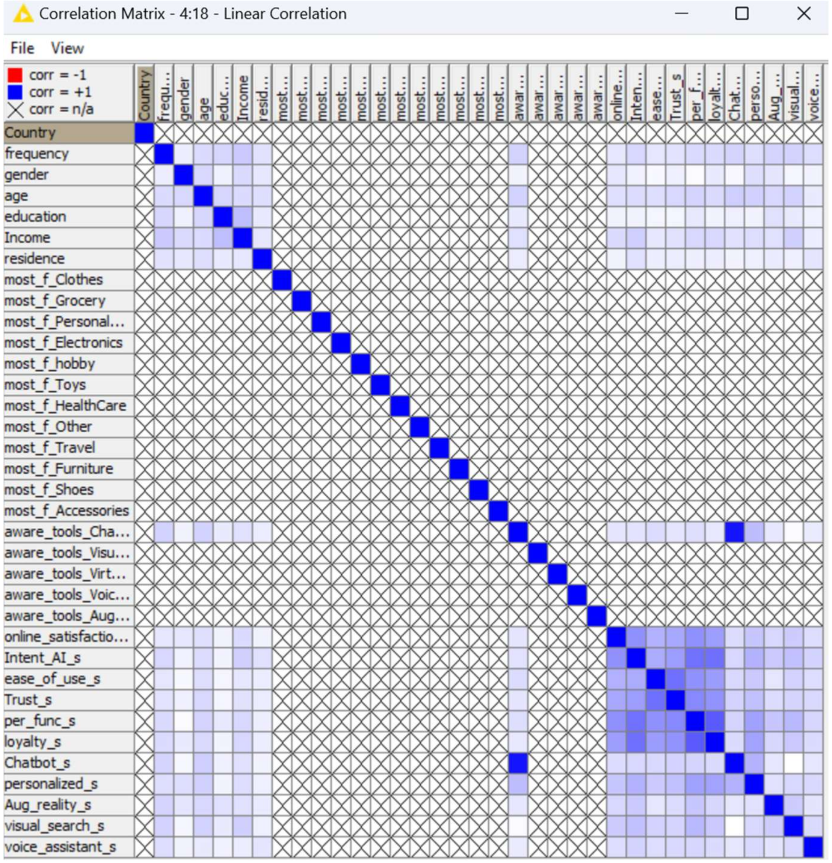


Fig 4.2 Correlation Matrix

5. Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean, structured, and ready for analysis. Below is a detailed description of the preprocessing steps applied to the survey data collected from respondents in Brazil.

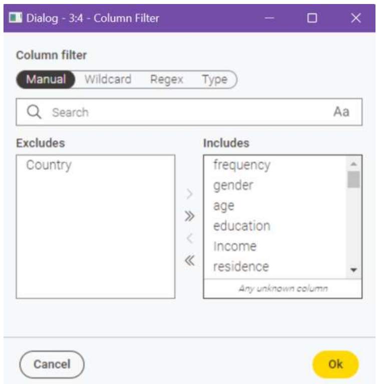
5.1. Data Cleaning

- **Handling Missing Values**: The dataset contained **no missing values**, so no imputation or deletion of rows was necessary.

- **Duplicate Entries**: The dataset was checked for any duplicate rows or repeated respondent entries. No duplicates were found, ensuring the data was unique and accurate.

5.2. Data Filtering

- **Country Filtering**: The dataset was filtered to include only **Brazilian respondents**, ensuring the analysis focused on data relevant to the study.



Dialog - 3:4 - Column Filter

Column filter

Manual

Excludes: Country

Includes: frequency, gender, age, education, Income, residence

Any unknown column

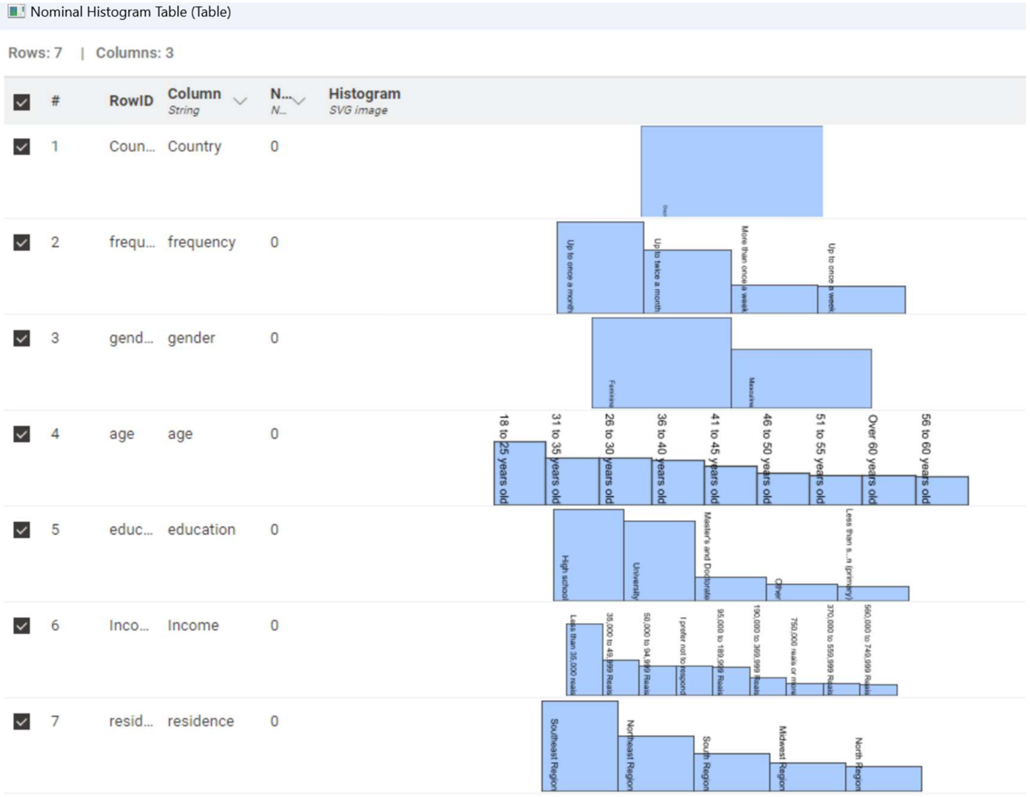
Fig 5.3 Column filter (Country removed)

6. Data Exploration and Statistics

To get an initial understanding of the dataset, exploratory analysis was conducted using the following tools:

- **Data Explorer Node:** This node provided an overview of the dataset, displaying basic summary statistics (mean, median, and standard deviation) and data types for each variable. It allowed us to assess the structure and distribution of the data.

- **Statistics Node:** The Statistics node was used to generate detailed statistical summaries, including histograms and key metrics like the mean, median, mode, and standard deviation for each numerical variable.



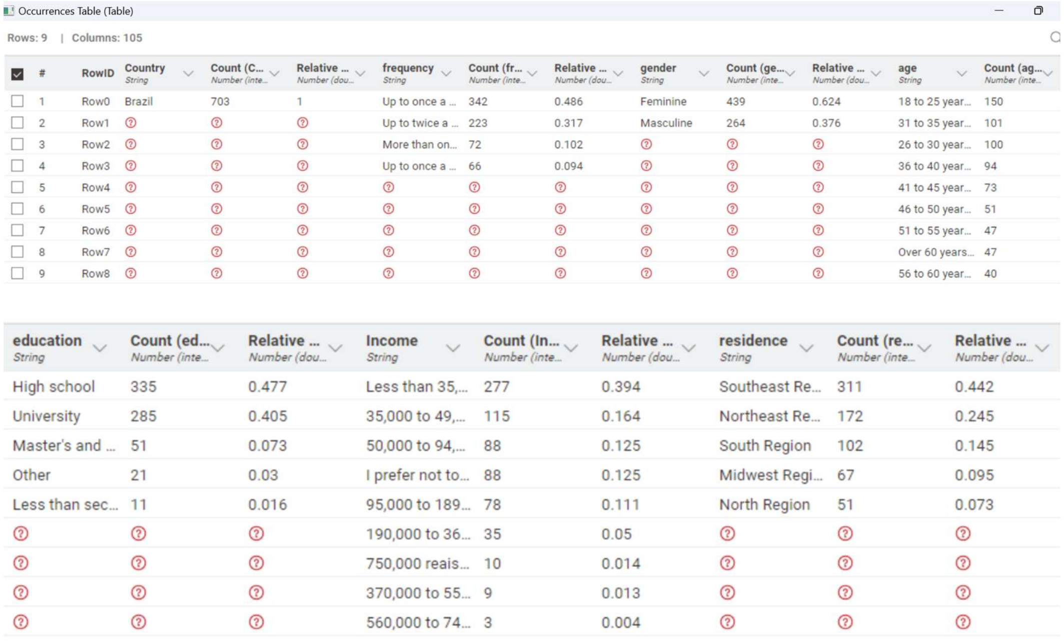
Nominal Histogram Table (Table)

Rows: 7 | Columns: 3

☑	#	RowID	Column String	N... N...	Histogram SVG image
☑	1	Coun...	Country	0	
☑	2	frequ...	frequency	0	
☑	3	gend...	gender	0	
☑	4	age	age	0	
☑	5	educ...	education	0	
☑	6	Inco...	Income	0	
☑	7	resid...	residence	0	

Fig 6.4 Nominal histogram table

- **Occurrence Table:** The occurrence table helped to visualize the count of unique entries in categorical variables, such as gender, education level, and residence location. This gave a clearer view of the distribution within these categories.



Occurrences Table (Table)

Rows: 9 | Columns: 105

☑	#	RowID	Country String	Count (C... Number (inte...	Relative ... Number (dou...	frequency String	Count (fr... Number (inte...	Relative ... Number (dou...	gender String	Count (ge... Number (inte...	Relative ... Number (dou...	age String	Count (ag... Number (inte...
□	1	Row0	Brazil	703	1	Up to once a ...	342	0.486	Feminine	439	0.624	18 to 25 year...	150
□	2	Row1	?	?	?	Up to twice a ...	223	0.317	Masculine	264	0.376	31 to 35 year...	101
□	3	Row2	?	?	?	More than on...	72	0.102	?	?	?	26 to 30 year...	100
□	4	Row3	?	?	?	Up to once a ...	66	0.094	?	?	?	36 to 40 year...	94
□	5	Row4	?	?	?	?	?	?	?	?	?	41 to 45 year...	73
□	6	Row5	?	?	?	?	?	?	?	?	?	46 to 50 year...	51
□	7	Row6	?	?	?	?	?	?	?	?	?	51 to 55 year...	47
□	8	Row7	?	?	?	?	?	?	?	?	?	Over 60 years...	47
□	9	Row8	?	?	?	?	?	?	?	?	?	56 to 60 year...	40

education String	Count (ed... Number (inte...	Relative ... Number (dou...	Income String	Count (In... Number (inte...	Relative ... Number (dou...	residence String	Count (re... Number (inte...	Relative ... Number (dou...
High school	335	0.477	Less than 35,...	277	0.394	Southeast Re...	311	0.442
University	285	0.405	35,000 to 49,...	115	0.164	Northeast Re...	172	0.245
Master's and ...	51	0.073	50,000 to 94,...	88	0.125	South Region	102	0.145
Other	21	0.03	I prefer not to...	88	0.125	Midwest Regi...	67	0.095
Less than sec...	11	0.016	95,000 to 189...	78	0.111	North Region	51	0.073
?	?	?	190,000 to 36...	35	0.05	?	?	?
?	?	?	750,000 reais...	10	0.014	?	?	?
?	?	?	370,000 to 55...	9	0.013	?	?	?
?	?	?	560,000 to 74...	3	0.004	?	?	?

Fig 6.2 Occurence Table

7. Visualizations and Analysis

The Bar Chart node for the quick and easy visualization of categorical data by representing the frequency of each category in a bar format. By connecting the node to the dataset, selecting the appropriate x axis (category) and y axis (count or value), and customizing the labels, these charts help in visually summarizing and interpreting the data distribution.

7.1. Gender Distribution

A bar chart was created to visualize the gender distribution of the respondents. The data revealed a fairly balanced distribution between Feminine and Masculine respondents, allowing for an unbiased analysis of gender based differences in technology awareness and preferences.

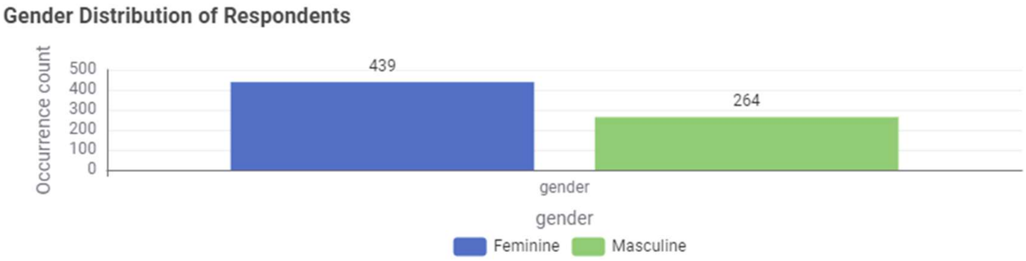


Fig 7.5 Gender distribution of respondents

- **Feminine Respondents:** Show higher levels of awareness and engagement with technology.
- **Masculine Respondents**: Tend to have slightly lower levels of awareness and engagement across certain technologies.

7.2. Frequency Distribution

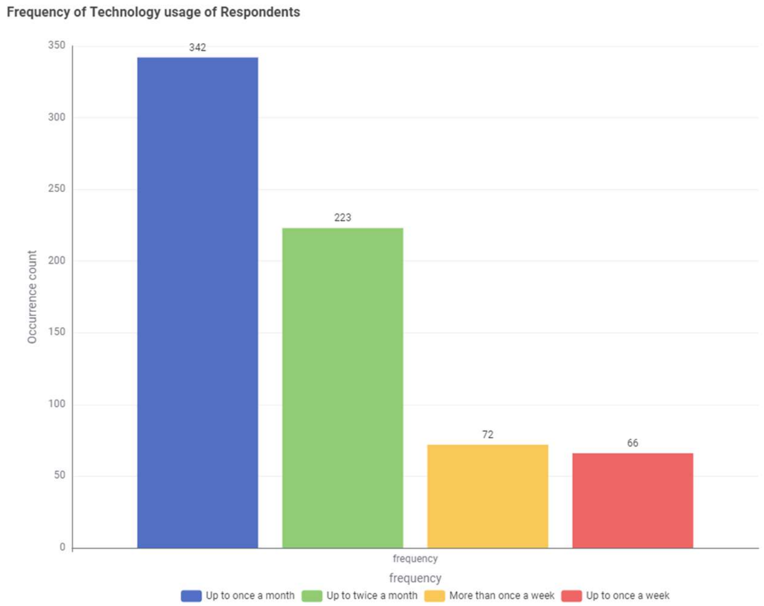


Fig 7.2 Frequency of technology usage of respondents

The frequency distribution chart highlights how often respondents participate in specific activities. The majority of respondents (342) engage in the activity once a month, while fewer respondents do so more frequently, such as twice a month (223), more than once a week (72), or once a week (66). This suggests that the activity has a predominantly monthly participation pattern.

7.3. Education Distribution

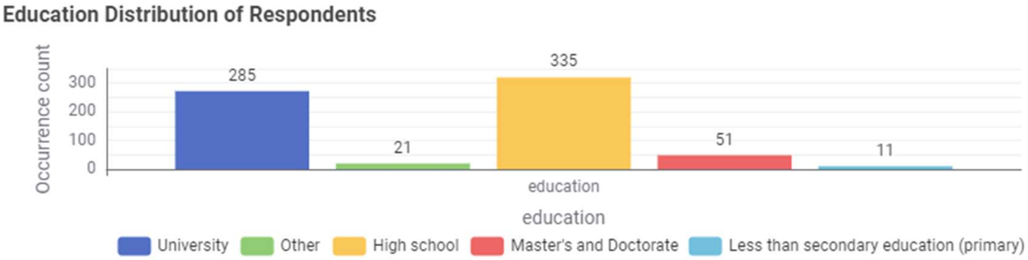


Fig 7.3 Education distribution of respondents

The education distribution chart shows that most respondents have completed secondary or higher education, with high school and university levels being the most common. There is a smaller percentage of respondents with advanced degrees or less than secondary education. This indicates a well educated sample population.

7.4. Residence Distribution

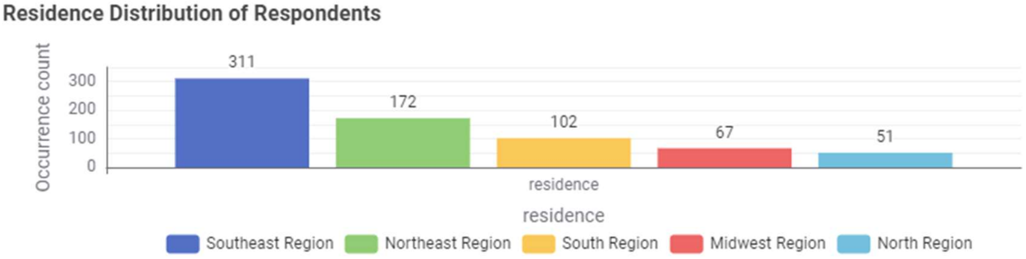


Fig 7.4 Residence distribution of respondents

The residence distribution chart shows that respondents are primarily concentrated in urbanized and economically developed regions of Brazil, particularly the Southeast and Northeast. The lower representation from the North and Midwest regions could be due to their smaller populations and less developed infrastructure, which may limit participation in such surveys.

7.5 Age Distribution

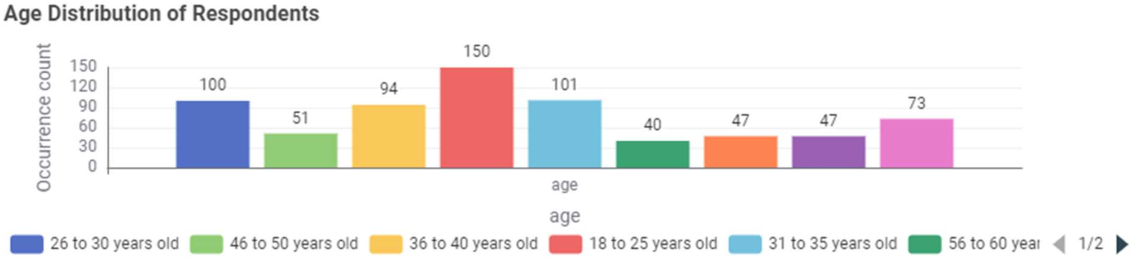


Fig 7.5 Age distribution of respondents

The age distribution chart reveals a strong skew towards younger respondents, particularly those in the 1825 age group, which accounts for the highest number of respondents (150). Participation declines as age increases, with significantly fewer respondents in the 5660 and 60+ age groups, indicating lower participation among older individuals.

- **Younger Age Groups**: High representation, indicating their active engagement in surveys and interest in the topics covered.
- **Middle Aged and Older Adults:** Fewer respondents from these groups, suggesting possible barriers to participation, such as access to technology or survey reach.

7.6. Income Distribution

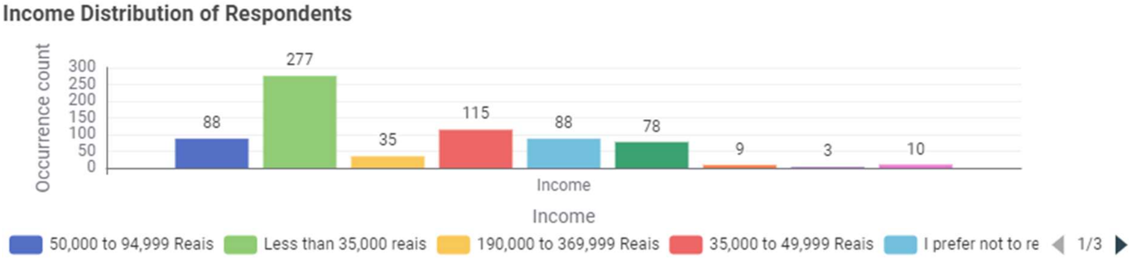
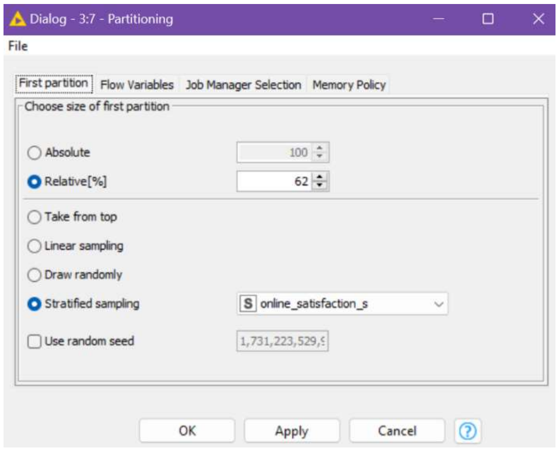


Fig 7.6 Income distribution of respondents

The income distribution chart shows a strong skew towards lower income levels, with a noticeable decline in the number of respondents as income levels increase. Additionally, a significant portion of respondents chose not to disclose their income, which might reflect sensitivity or discomfort around income related questions.

8. Data Partitioning

- **Stratified Sampling**: **Stratified sampling** was used to ensure that both the **training** and **test** datasets had the same distribution of the satisfaction variable. This approach addressed the imbalance in the target variable.
- **Training and Test Data Split**: The data was split into **training** and **test** datasets, ensuring that the training data represented the overall distribution of satisfaction levels.



Choose size of first partition

○ Absolute 100

● Relative[%] 62

○ Take from top

○ Linear sampling

○ Draw randomly

● Stratified sampling S online_satisfaction_s

□ Use random seed 1,731,223,529,

Fig 8.6 Partitioning node

The preprocessing steps outlined above ensured that the dataset was clean, balanced, and ready for analysis. The dataset was filtered to include only Brazilian respondents, and the necessary categorical variables were encoded. The final dataset was split using stratified sampling to address class imbalance, ensuring that both the training and test sets were representative of the overall satisfaction levels. The data is now prepared for modeling and predictive analysis.

9. Modelling

Modeling is the next step after preprocessing, where various techniques are applied to predict the target variable (customer satisfaction) based on the features in the dataset. different machine learning algorithms were applied to predict **customer satisfaction** (binary outcome: **Satisfied** or **Unsatisfied**) based on demographic characteristics, consumer preferences, and awareness of technological advancements. The models considered here include **Decision Tree**, **Random Forest**, and **Gradient Boosting** methods.

9.1. Model Selection

For binary classification, three machine learning algorithms were selected due to their effectiveness in handling non-linear relationships and their ability to model complex interactions:

- **Decision Tree**: A simple, interpretable model that splits the data into subsets based on feature values to make predictions. Each decision node represents a feature, and each leaf node represents a class label.
- **Random Forest**: An ensemble of decision trees that works by averaging the predictions of multiple trees to improve accuracy and reduce overfitting.
- **Gradient Boosting**: A more advanced ensemble method that builds trees sequentially, where each tree corrects the errors made by the previous one. It is known for its high accuracy and handling of complex datasets.

9.2. Decision tree Model

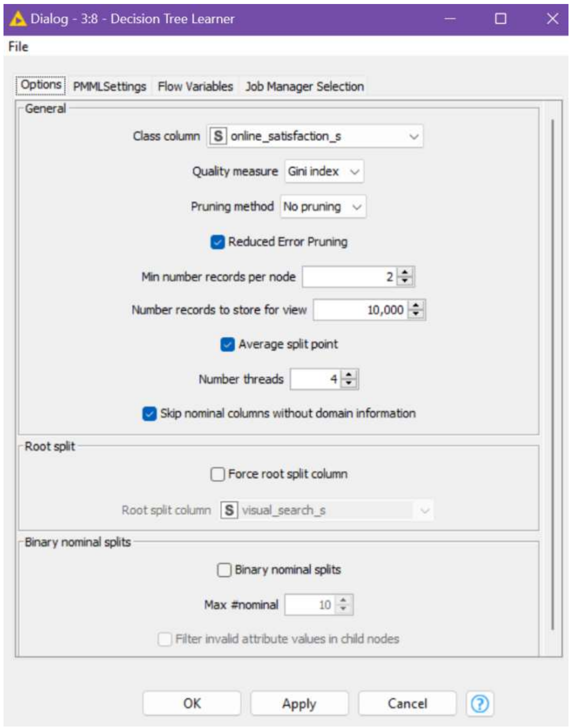
The **Decision Tree** algorithm to predict **customer satisfaction** levels (categorized as either **Satisfied** or **Unsatisfied**) based on survey data. The Decision Tree model was chosen for its interpretability and its capability to handle categorical data without requiring data transformation.

9.2.1. Decision Tree Model Configuration

The Decision Tree model was configured with the following settings to optimize predictive performance:

- **Class Column**: The target variable was set as online_satisfaction_s, representing customer satisfaction.

- **Quality Measure**: Gini index, a measure that helps in determining the quality of splits by evaluating the impurity at each node.
- **Pruning Method**: No pruning was applied, allowing the model to fully expand and identify all possible decision paths.



General

Class column: S online_satisfaction_s

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 2

Number records to store for view: 10,000

☑ Average split point

Number threads: 4

☑ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: S visual_search_s

Binary nominal splits

☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

Fig 9.7 Decision tree Learner configuration

9.2.2. Model Evaluation and Results

The performance of the Decision Tree model was assessed on the test dataset (38% of the data) using various evaluation metrics, including **True Positives (TP)**, **False Positives (FP)**, **True Negatives (TN)**, and **False Negatives (FN)**. These metrics were further used to calculate key performance indicators:

- **Recall/Sensitivity**: Measures the model's ability to correctly identify true positives.
- **Precision**: Reflects the accuracy of positive predictions.
- **Specificity**: Measures the model's capability to correctly identify negatives.
- **F-measure**: The harmonic mean of precision and recall, providing a balanced view of model performance.

Partioning (Training :Test	Target Predictor	Specificity	Precision	Sensitivity	F-score
70:30(Gain Ratio)	**Unsatisfied**	**0.853**	**0.533**	**0.414**	**0.466**
	Satisfied	**0.414**	**0.782**	**0.853**	**0.816**
70:30(Gini Index)	**Unsatisfied**	**0.829**	**0.529**	**0.466**	**0.495**
	Satisfied	**0.466**	**0.789**	**0.829**	**0.808**
62:38(Gain ratio)	**Unsatisfied**	0.806	0.485	0.458	0.471
	Satisfied	0.458	0.788	0.806	0.797
62:38(Gini Index)	**Unsatisfied**	0.723	0.446	0.569	0.5
	Satisfied	0.569	0.811	0.723	0.769
80:20 (Gain ratio)	**Unsatisfied**	0.606	0.883	0.824	0.852
	Satisfied	0.824	0.488	0.606	0.541
80:20(Gini Index)	**Unsatisfied**	0.814	0.455	0.385	0.417
	Satisfied	0.385	0.767	0.814	0.79

Table 9.1 Decision tree scorer result

9.2.3. Analysis and Insights from Decision Tree Scorer Results

To evaluate model performance on predicting customer satisfaction, we tested three different partitioning strategies (70:30, 62:38, and 80:20 training-to-test ratios) and used two criteria (Gain Ratio and Gini Index) in the decision tree model. Each combination was assessed using specificity, precision, sensitivity, and F-score metrics, yielding the following insights:

1. **Partitioning Ratios**:
    - The three partitioning ratios (70:30, 62:38, 80:20) offer insights into how data splitting impacts model performance.
    - The **80:20 split** tends to yield better performance in both specificity and sensitivity, especially for unsatisfied predictions, suggesting that a larger training set helps the model generalize more effectively.
2. **Metric Comparisons**:
    - **Specificity**: This metric, which measures the model's accuracy in identifying "unsatisfied" customers, is highest with the 80:20 split, particularly with the Gain Ratio criterion (up to 0.883). This implies that the model performs well at avoiding false positives for unsatisfied predictions.

   - **Precision and Sensitivity**: Precision, or the accuracy of positive predictions, is generally higher for "Satisfied" predictions with the Gain Ratio criterion, especially in the 70:30 and 80:20 splits. Sensitivity, which measures true positive rates, is also highest for "Satisfied" in these splits, showing that the model is effective at identifying satisfied customers under these settings.
   - **F-Score**: The F-score combines precision and sensitivity, providing a balanced measure of model performance. The **80:20 split using Gain Ratio** achieves the highest F-scores for both satisfied (0.541) and unsatisfied (0.852) predictions, suggesting it may be the most balanced partition and criterion choice.
3. **Comparison of Gain Ratio and Gini Index**:
   - The **Gain Ratio** criterion generally performs better in predicting satisfied customers across all partitions, especially in sensitivity and F-score, indicating it may be better suited for models aiming to capture customer satisfaction patterns.
   - The **Gini Index** criterion has comparable specificity scores, particularly in the unsatisfied category, which might be beneficial if identifying unsatisfied customers is prioritized. However, it falls slightly behind in overall sensitivity and precision for satisfied predictions.
4. **Optimal Model Selection**:
   - The **80:20 partition with Gain Ratio** stands out as the most balanced configuration, with high F-scores for both satisfied and unsatisfied predictions. It achieves a strong trade-off between identifying satisfied customers and avoiding false positives for unsatisfied ones.
   - If the analysis focuses specifically on accurately predicting satisfied customers, the **70:30 partition with Gain Ratio** also shows high sensitivity (0.829) and F-score (0.816) for the satisfied class.
5. **Recommendations for Further Analysis**:
   - Given these results, further tuning of the decision tree model’s parameters may enhance performance. Additionally, testing other algorithms, such as Random Forest or Gradient Boosting, could help optimize for specific metrics based on the priority (e.g., maximizing satisfaction prediction or minimizing unsatisfied misclassifications).

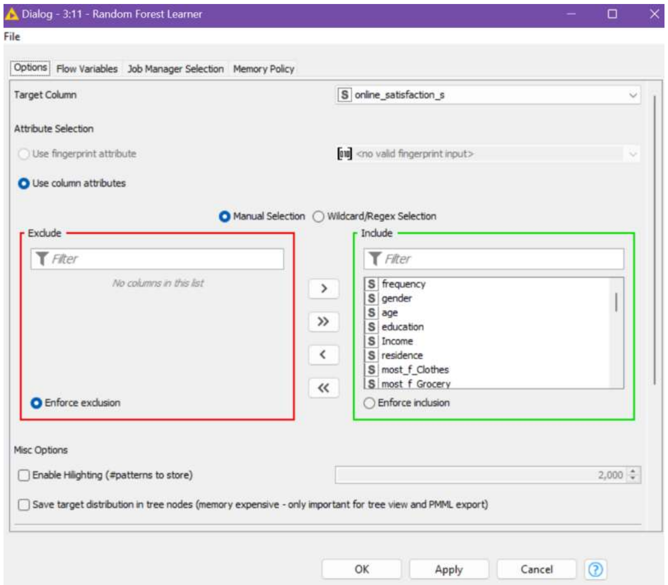
9.3. Random Forest Model

The **Random Forest model** is an ensemble technique that builds multiple decision trees using random subsets of data. It improves prediction accuracy by averaging or voting across all trees, making it robust and less prone to overfitting. In the context of predicting customer satisfaction, Random Forest can effectively handle complex, high-dimensional data and provide strong predictive performance, as seen in the comparative analysis of machine learning models.

9.3. 1. Random Forest Model Configuration

The Random Forest model was configured to include the following specifications:

- **Target Column**: online_satisfaction, representing the binary outcome of customer satisfaction (Satisfied or Unsatisfied).
- **Features Included**: All available attributes were utilized as input features to capture as much information as possible.
- **Ensemble Structure**: The model uses multiple decision trees, each built on a different subset of the data, to improve overall performance and generalization.
- **Evaluation Method**: The Random Forest Scorer was used to assess the model's performance, providing metrics on prediction accuracy for both satisfaction classes.



Target Column: S online_satisfaction_s

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

S frequency
S gender
S age
S education
S Income
S residence
S most_f_Clothes
S most_f_Grocery

○ Enforce inclusion

Misc Options

□ Enable Hilighting (#patterns to store) 2,000

□ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Fig 9.2 Random forest configuration

9.3.2. Model Evaluation and Results

The model's performance was evaluated using key metrics that provide insight into its effectiveness in predicting both satisfied and unsatisfied customer responses:

Partioning (Training :Test	Target Predictor	Specificity	Precision	Sensitivity	F-score
70:30(Gain Ratio)	**Unsatisfied**	**0.906**	**0.674**	**0.468**	**0.552**
	Satisfied	**0.468**	**0.804**	**0.906**	**0.852**
70:30(Gini	**Unsatisfie**	**0.919**	**0.692**	**0.435**	**0.535**

Index)	**d**				
	Satisfied	**0.435**	**0.797**	**0.919**	**0.854**
62:38(Gain ratio)	**Unsatisfied**	0.879	0.641	0.526	0.577
	Satisfied	0.526	0.819	0.879	0.848
62:38(Gini Index)	**Unsatisfied**	0.889	0.667	0.538	0.596
	Satisfied	0.538	0.824	0.889	0.856
80:20 (Gain ratio)	**Unsatisfied**	0.9	0.583	0.341	0.431
	Satisfied	0.341	0.769	0.9	0.829
80:20(Gini Index)	**Unsatisfied**	0.9	0.6	0.366	0.455
	Satisfied	0.366	0.776	0.9	0.833

Table 9.1. Random forest scorer results

9.3.3.Analysis and Insights from Random Forest Scorer Results

1. **Best Partition and Criterion**:
   - The **70:30 split with Gain Ratio** shows the highest F-score for satisfied predictions (0.852) and performs well across specificity, precision, and sensitivity for both categories, indicating it as the optimal choice.
   - The **62:38 split with Gain Ratio** also performs strongly for both satisfied and unsatisfied predictions, achieving a high F-score (0.848 for satisfied and 0.577 for unsatisfied).
2. **Gain Ratio vs. Gini Index**:
   - **Gain Ratio** performs consistently well in sensitivity and F-scores for satisfied predictions across partitions, especially in the 70:30 and 62:38 splits.
   - **Gini Index** slightly improves specificity for unsatisfied predictions but generally yields lower sensitivity and F-scores for satisfied predictions.
3. **Partition Insights**:
   - The **70:30 partition** maximizes sensitivity for satisfied predictions, making it suitable for identifying satisfied customers accurately.
   - The **80:20 partition** shows lower F-scores across both categories, indicating that a larger test set might reduce overall performance in this model.
4. **Recommendations**:
   - For balanced and accurate satisfied customer predictions, the **70:30 partition with Gain Ratio** is recommended. If specific focus is on satisfied predictions, **Gain Ratio** consistently outperforms Gini Index.

These insights confirm that the 70:30 partition with Gain Ratio offers the best overall balance for this dataset when using a Random Forest model.

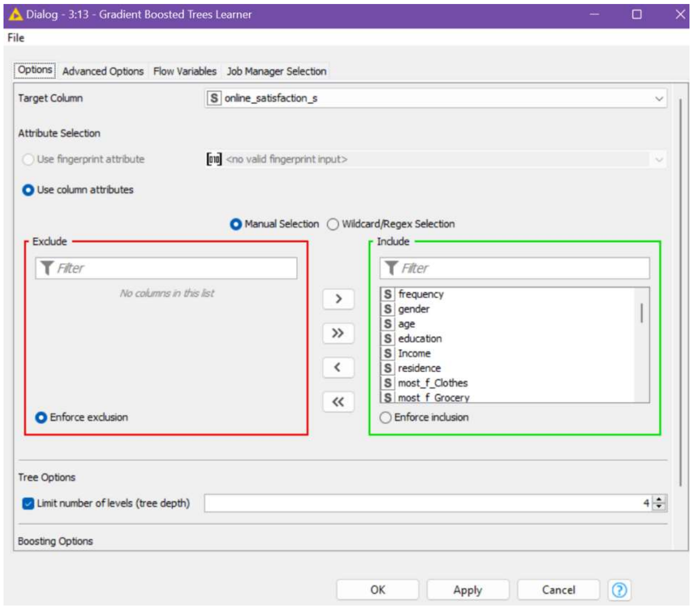
9.4. Gradient Boosting Model

The **Gradient Boosting model** is an ensemble technique that builds decision trees sequentially, where each new tree corrects the errors of the previous ones. It focuses on minimizing prediction errors by adjusting weights on difficult-to-predict data points. For customer satisfaction prediction, Gradient Boosting can provide highly accurate results by combining multiple weak models into a strong predictor, as demonstrated in the comparative analysis of different machine learning models.

9.4.1. Gradient Boosting Model Configuration

The Gradient Boosting model was set up with the following specifications:

- **Target Column**: online_satisfaction, the binary variable representing customer satisfaction (Satisfied or Unsatisfied).
- **Features Included**: All available attributes were used as input features to fully capture relevant information for predictions.
- **Ensemble Structure**: Gradient Boosting combines multiple weak learners (decision trees) into a strong predictive model by incrementally optimizing for improved accuracy.
- **Evaluation Method**: The Gradient Boosting Scorer was used to assess the model’s predictions for both satisfied and unsatisfied customer responses.



Dialog - 3:13 - Gradient Boosted Trees Learner

Target Column: S online_satisfaction_s

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude: No columns in this list

● Enforce exclusion

Include:
S frequency
S gender
S age
S education
S Income
S residence
S most_f_Clothes
S most_f_Grocery

○ Enforce inclusion

Tree Options

☑ Limit number of levels (tree depth) 4

Boosting Options

Figure 9.3 Gradient Boosting model configuration

9.4.2. Model Evaluation and Results

The Gradient Boosting model was evaluated based on various metrics, focusing on predictive performance for each satisfaction category:

Partioning (Training :Test)	Target Predictor	Specificity	Precision	Sensitivity	F-score
70:30	**Unsatisfied**	**0.805**	**0.532**	**0.532**	**0.532**
	Satisfied	**0.532**	**0.805**	**0.805**	**0.805**
62:38	**Unsatisfied**	0.874	0.571	0.41	0.478
	Satisfied	0.41	0.783	0.874	0.826
80:20	**Unsatisfied**	0.86	0.548	0.415	0.472
	Satisfied	0.415	0.782	0.86	0.819

Table 9.2. Gradient Boosting model configuration

9.4.3. Analysis and Insights from Gradient Boosting Model Results

1. **Best Partition**:
   - The **70:30 split** achieves the highest overall balance, with strong F-scores for both "Satisfied" (0.805) and "Unsatisfied" (0.532) predictions, suggesting this partition is the most stable configuration.
2. **Partition Insights**:
   - The **70:30 partition** offers high sensitivity and precision for both categories, which makes it effective at identifying satisfied and unsatisfied customers with good accuracy.
   - The **62:38 and 80:20 splits** also perform reasonably well, particularly for satisfied predictions, with high sensitivity (0.874 for 62:38 and 0.86 for 80:20). However, the F-scores for unsatisfied predictions in these partitions are lower, indicating less consistent performance in identifying unsatisfied customers.
3. **Overall Model Performance**:
   - Gradient Boosting generally shows high sensitivity and precision for satisfied predictions, especially in the 62:38 and 80:20 partitions. This suggests it may be more effective at correctly identifying satisfied customers.
   - Specificity and F-scores for unsatisfied predictions are lower across all partitions, indicating room for improvement in detecting unsatisfied customers accurately.
4. **Recommendations**:
   - For balanced predictions, **70:30** is recommended as the optimal partition. However, if the focus is on accurately identifying satisfied customers, the **62:38 partition** provides the highest F-score (0.826) for satisfied predictions.

These insights indicate that the 70:30 partition provides a good balance for both categories, while the 62:38 split may be preferable if the primary focus is on satisfied predictions.

9.5. Comparative Summary

Model	Partition (Train:Test)	Target Predictor	F-Score	Best Use
Decision Tree	70:30 (Gain Ratio)	Unsatisfied	0.552	Balanced performance with a higher F-score for satisfied
	70:30 (Gain Ratio)	Satisfied	0.852	
	80:20 (Gain Ratio)	Unsatisfied	0.431	Lower overall F-score, balanced approach
	80:20 (Gain Ratio)	Satisfied	0.829	
Random Forest	70:30 (Gain Ratio)	Unsatisfied	0.552	Optimal F-score for balanced prediction
	70:30 (Gain Ratio)	Satisfied	0.852	
	62:38 (Gain Ratio)	Unsatisfied	0.577	Higher F-score for balanced prediction
	62:38 (Gain Ratio)	Satisfied	0.848	
Gradient Boosting	70:30	Unsatisfied	0.532	Balanced F-score for both satisfied and unsatisfied
	70:30	Satisfied	0.805	
	62:38	Unsatisfied	0.478	Best for identifying satisfied customers
	62:38	Satisfied	0.826	

Table 9.3 Comparative summary of F-measures for the three models

- **Decision Tree** shows the highest **F-score** for **satisfied customers** at **0.852** when using the **70:30 partition**, with balanced performance.
- **Random Forest** also performs well with high **F-scores** of **0.852** for **satisfied customers** and **0.577** for **unsatisfied** when using the **62:38 partition**.
- **Gradient Boosting** shows a balanced **F-score** at **0.532** for **unsatisfied** and **0.805** for **satisfied** using the **70:30 partition** but has a lower overall F-score in the **62:38 partition**.

This focused analysis on the **F-score** highlights the model configurations most suitable for balanced predictions and identifying satisfied customers.

10. Conclusion:

This analysis compared the performance of Decision Tree, Random Forest, and Gradient Boosting models using F-scores to predict customer satisfaction.

- Decision Tree performed best for predicting satisfied customers with the 70:30 partition, providing balanced results. However, it showed weaker performance for unsatisfied customers in the 80:20 partition.
- Random Forest excelled with a 70:30 partition, offering high F-scores for both satisfied and unsatisfied customers, making it ideal for balanced predictions.
- Gradient Boosting performed well for satisfied customers with the 62:38 partition, though its unsatisfied customer predictions were weaker.

10.1 Recommendation:

Use Random Forest for balanced predictions across both groups, especially with the 62:38 partition.

- Decision Tree is best when focusing on satisfied customers with the 70:30 partition.
- Gradient Boosting is suited for scenarios prioritizing satisfied customers, particularly with the 62:38 partition.