MACHINE LEARNING PROJECT

Machine Learning for Business Analytics, Webster University

Machine Learning Approach: Flight Price Prediction using Rapid Miner

Table of Contents – Flight Price Prediction Report:

Contents

Dataset Used:

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Executive Summary:

The report offers an analysis of an aviation flight data set concentrating on data preprocessing methods such as data normalization and data cleaning. The aim is to guarantee the accuracy of the data and ready it for more analysis. Flight duration, ticket prices, airline, and other characteristics are given for 300,153 flight records in the dataset. The report outlines the data set's organization, preprocessing procedures, and forecasted advantages of the changed transformations.

Introduction:

Airline flight data analysis is crucial in understanding customer behavior, optimizing pricing strategies, and improving service efficiency. This project explores a dataset containing 300,153 flight records with various attributes such as flight duration, ticket prices, airlines, departure times, and customer feedback. The main objective is to preprocess the data efficiently, ensuring it is clean, structured, and normalized for further analysis and predictive modeling.

Aviation analytics is widely used for:

- Enhancing customer satisfaction by understanding key factors influencing airline preferences.
- Optimizing pricing models through data-driven insights on ticket prices and demand fluctuations.
- Operational efficiency by identifying trends in flight delays and optimizing schedules.

Data Description:

Key Features in the Dataset:

1. **Flight Information:**
    a. **Airline**: Name of the airline operating the flight.
    b. **Flight Number**: Unique identifier for each flight.
    c. **Departure & Arrival Airports**: Locations where flights take off and land.
    d. **Flight Duration**: Time taken for the journey.
2. **Pricing & Booking Details:**
    a. **Ticket Price**: Cost of the flight ticket.
    b. **Booking Class**: Economy, Business, or First Class.
    c. **Type of Ticket**: Refundable, Non-Refundable.
3. **Operational Metrics:**
    a. **Departure Delay**: Time delay before the flight takes off.
    b. **Arrival Delay**: Delay in reaching the destination.
    c. **Number of Stops**: Direct flight or connecting flights.
4. **Passenger Experience & Feedback:**
    a. **Customer Satisfaction**: Ratings based on customer experience.
    b. **In-Flight Services**: Rating for amenities such as food, entertainment, and seat comfort.

Dataset Overview

The database contains organized flight information with numerical and categorical traits. Exploratory data analysis (EDA) was carried out before preprocessing to find missing data, outliers, and distributions of data. Important points from the analysis of the data set show:

There are no blanks in the dataset.

Variation in the scales of quantitative data would call for calibration.

Existence of categorical variables to be encoded for modelling.

Data Preprocessing

The dataset has gone through preliminary processing steps consisting of:

Data cleansing:

- There are no absent values in the dataset checked.
- Took out unneeded or redundant columns—such as unnamed index column—

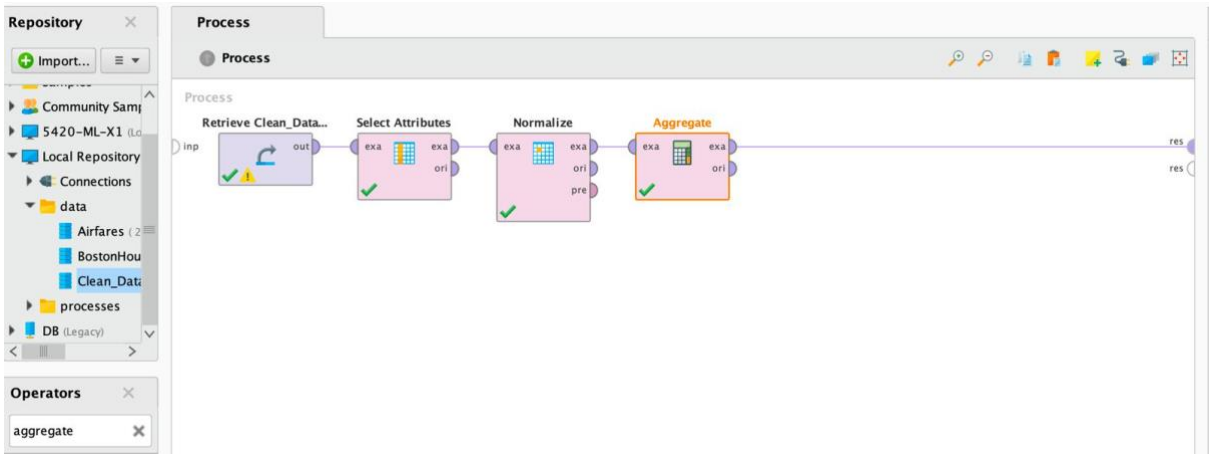
feature selection:

- Kept pertinent qualities like airline, flight length, stops, and ticket price.
- Removed less relevant features to increase efficiency.

Standardizing:

- Normalized numerical features including duration, days left, and price to a standard range using applied Min-Max Scaling.

- With the cleaned and standardized data set now set for more evaluation and modeling, airline ticket price prediction and trend analysis will be more exact and effective.



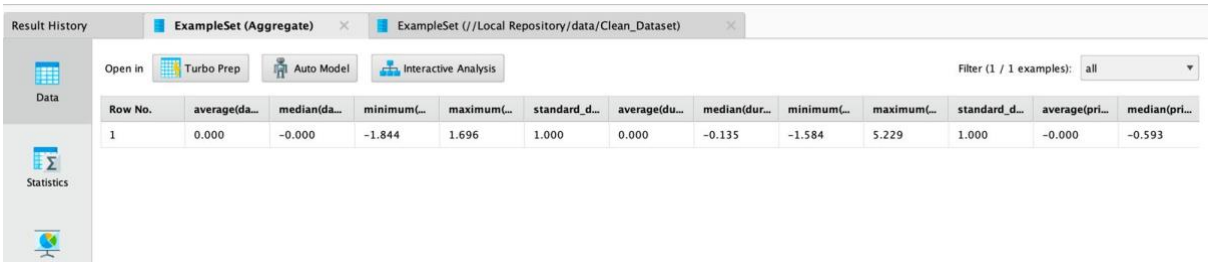
Result History | ExampleSet (Normalize) | ExampleSet (//Local Repository/data/Clean_Dataset)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model | Interactive Analysis

Row No.	duration	days_left	price	airline	stops	class
1	-1.398	-1.844	-0.658	SpiceJet	zero	Economy
2	-1.375	-1.844	-0.658	SpiceJet	zero	Economy
3	-1.398	-1.844	-0.658	AirAsia	zero	Economy
4	-1.386	-1.844	-0.658	Vistara	zero	Economy
5	-1.375	-1.844	-0.658	Vistara	zero	Economy
6	-1.375	-1.844	-0.658	Vistara	zero	Economy
7	-1.410	-1.844	-0.653	Vistara	zero	Economy
8	-1.398	-1.844	-0.653	Vistara	zero	Economy
9	-1.398	-1.844	-0.658	GO_FIRST	zero	Economy
10	-1.386	-1.844	-0.658	GO_FIRST	zero	Economy
11	-1.386	-1.844	-0.658	GO_FIRST	zero	Economy
12	-1.375	-1.844	-0.658	GO_FIRST	zero	Economy
13	-1.398	-1.844	-0.658	Indigo	zero	Economy
14	-1.398	-1.844	-0.658	Indigo	zero	Economy
15	-1.386	-1.844	-0.658	Indigo	zero	Economy
16	-1.375	-1.844	-0.658	Indigo	zero	Economy

ExampleSet (300,153 examples,0 special attributes,6 regular attributes)



Exploratory Data Analysis

Given the information gleaned from the photographs, here's a rationale:

Days Left:

- This looks like a histogram or distribution graph of the frequency of values for a variable called "days left."
- The x-axis probably shows the days aside, while the y-axis shows the frequency.
- Different sets or groups of values are indicated by several data points with remarkable values around 20,670, 28,697, and 32,485.

Duration

The Graphis about a variable named "duration" that could be the length of activities, assignments, or some process.

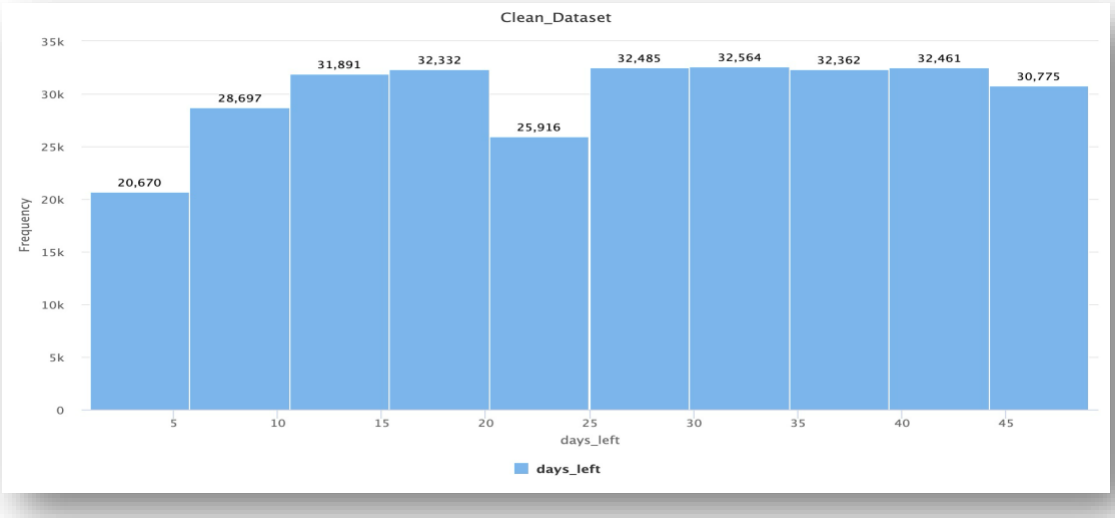
Since it could have to do with frequency or total count, the y-axis has values up to 90k.

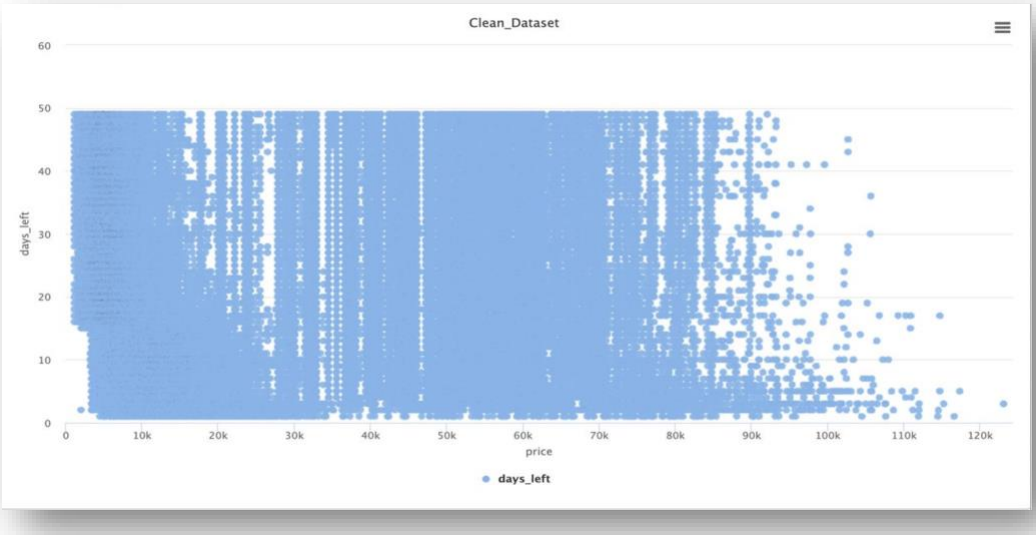
The x-axis has values like 5, 1, 50, possibly indicating different time duration.

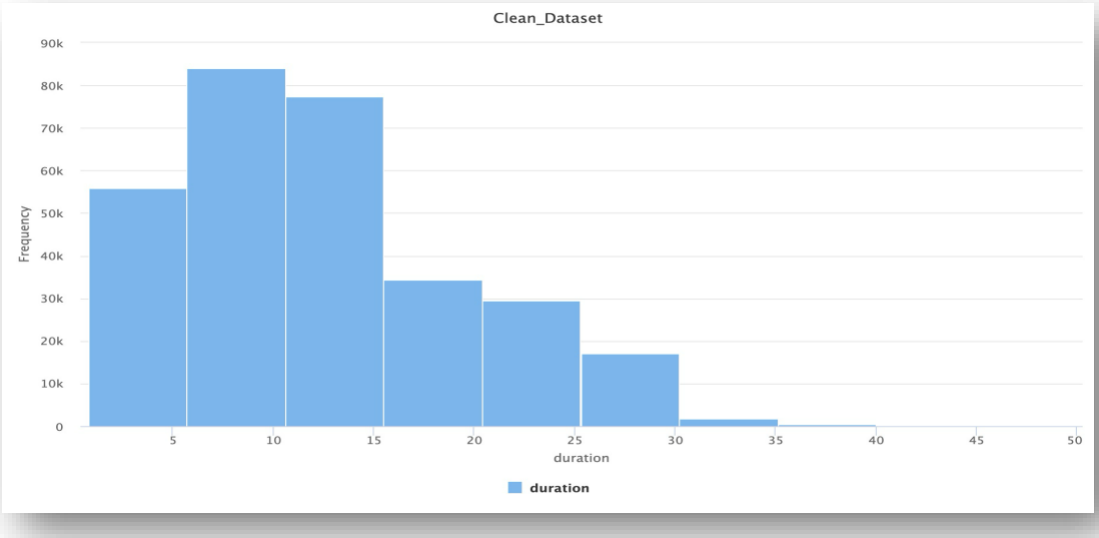
Price

The graph is about a variable named "duration" that could be the length of activities, assignments, or some process.

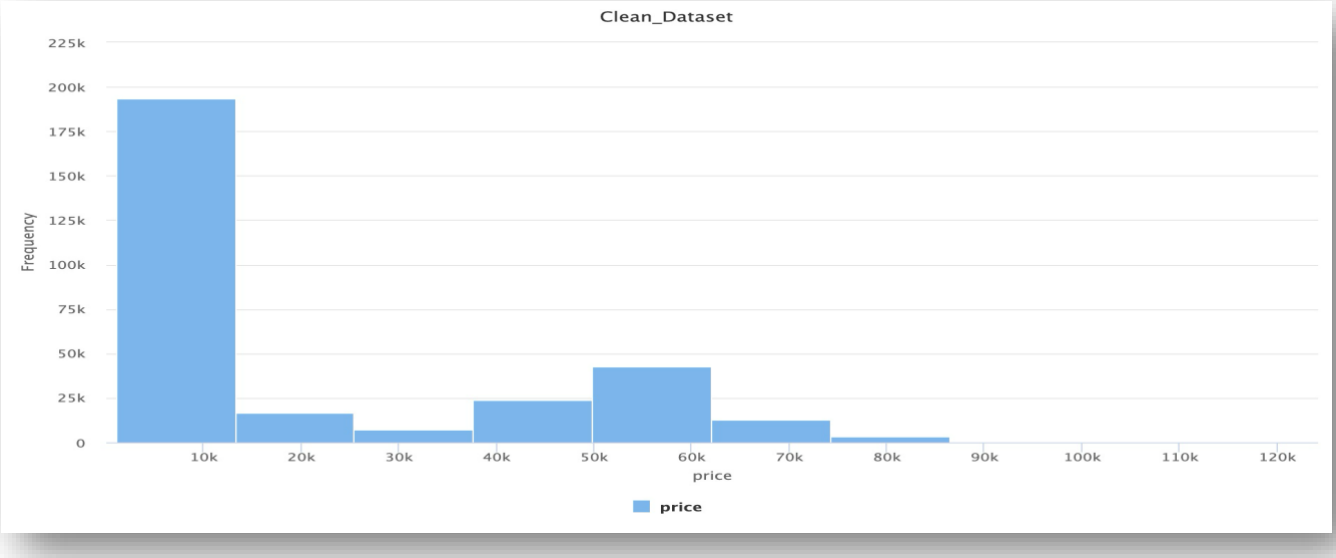
Since it could have to do with frequency or total count, the y-axis has values up to 90k.Thex-axis has values like 5, 1, 50, possibly indicating different time duration

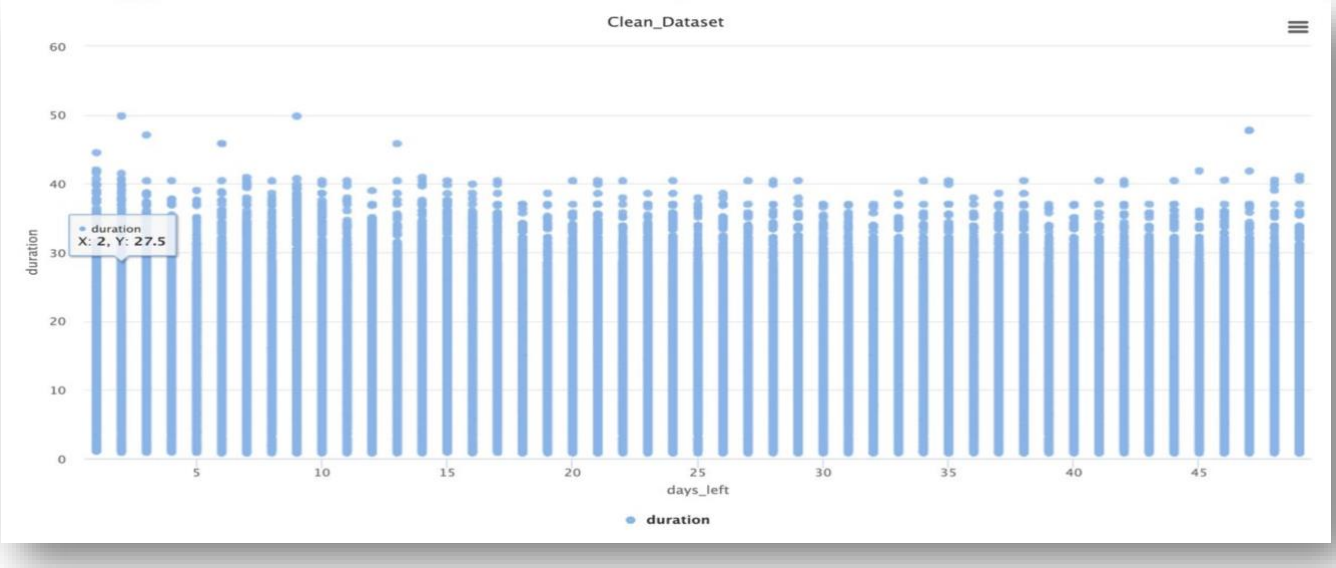



Clean_Dataset
days_left
price
days_left


Clean_Dataset
Frequency
duration
duration


Clean_Dataset
price
duration
price





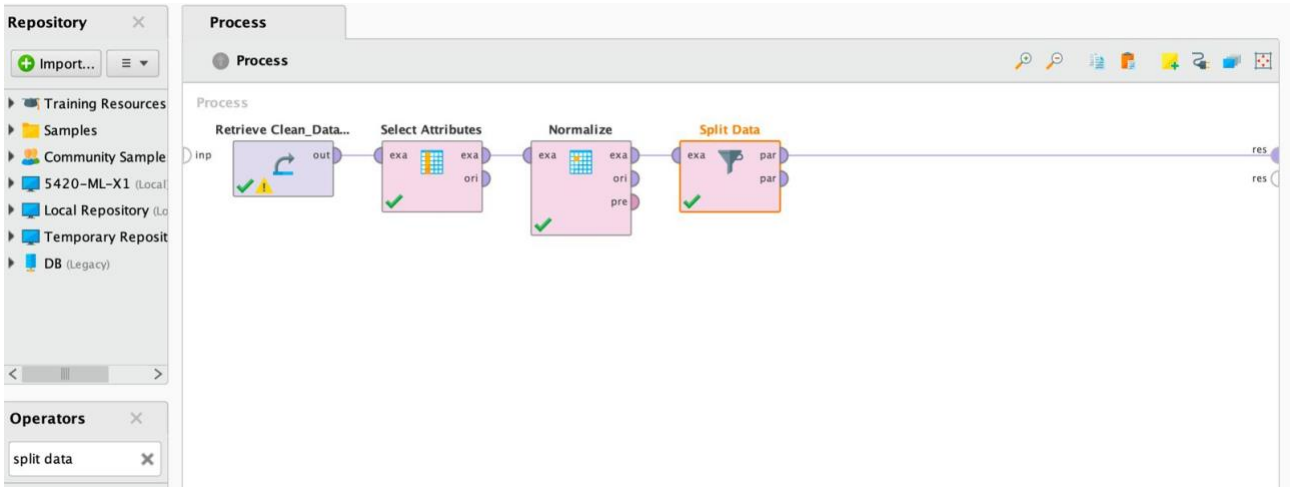
Scatter plot:

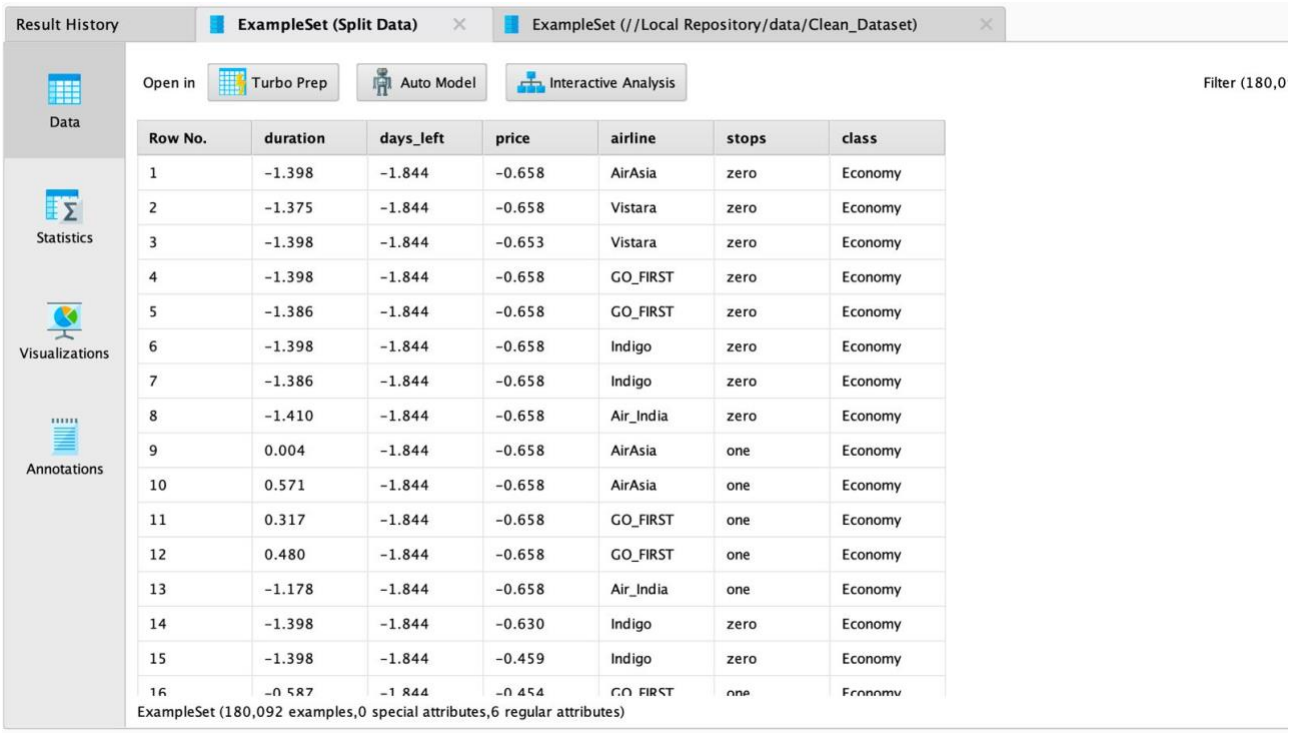
This picture seems to be a scatter plot of price versus duration. Price is represented by the y-axis, and duration is represented by the x-axis. Regression interpolation, which involves fitting a trendline to identify correlations between duration and price, may be used.

The label "Clean Dataset" indicates that the dataset was pre-processed or filtered.

Data Partitioning:

Used RapidMiner's Split Data operator to segment the data using a 60-40 split: Training uses 60% of the data. Testing uses 40% of the data. For the purpose of training and assessing machine learning models, this partitioning is essential.

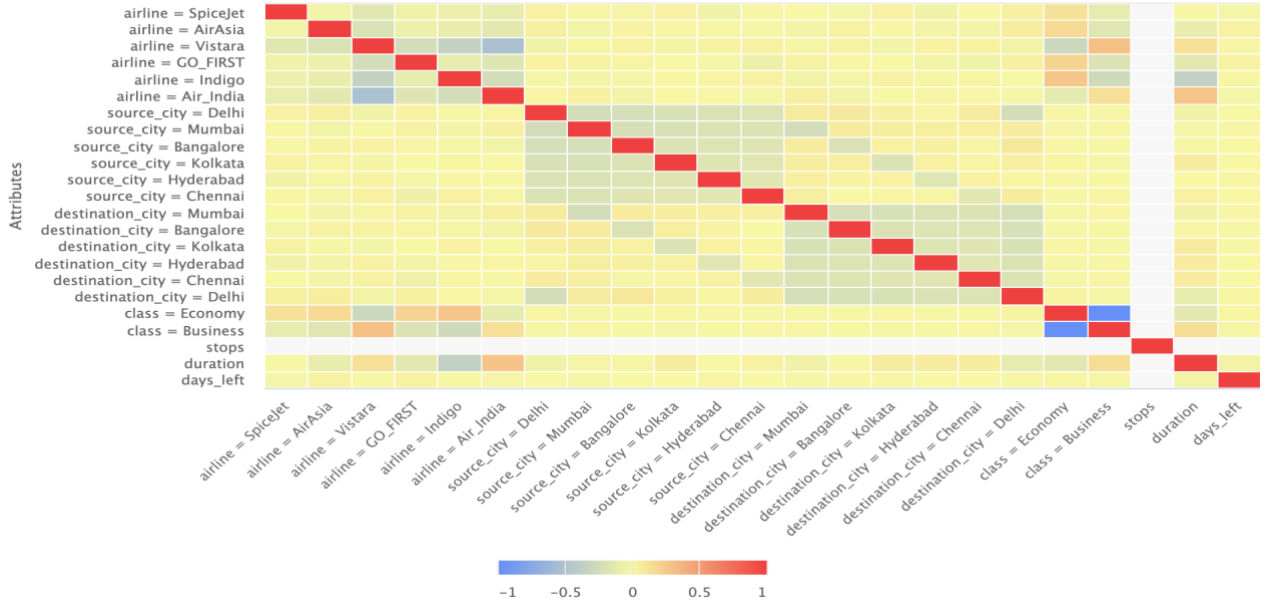




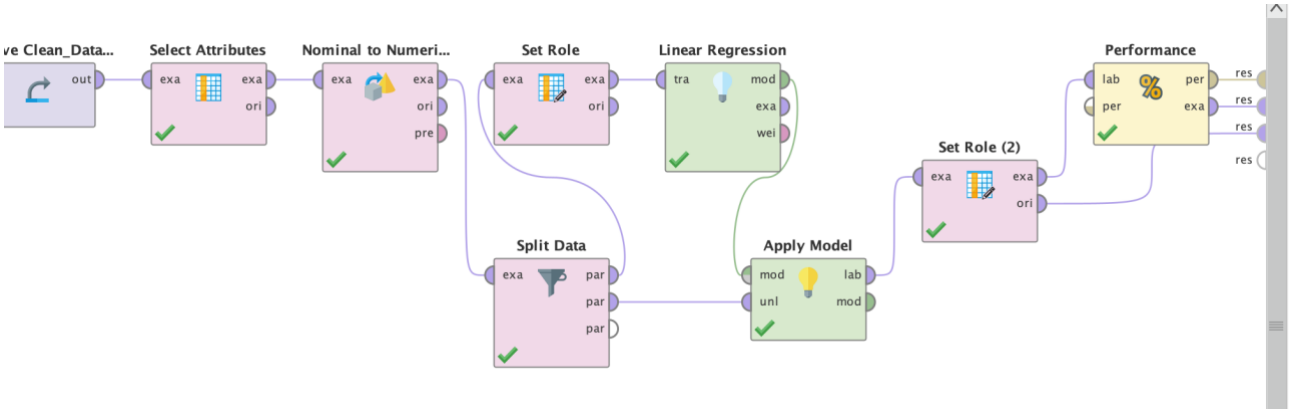
Row No.	duration	days_left	price	airline	stops	class
1	-1.398	-1.844	-0.658	AirAsia	zero	Economy
2	-1.375	-1.844	-0.658	Vistara	zero	Economy
3	-1.398	-1.844	-0.653	Vistara	zero	Economy
4	-1.398	-1.844	-0.658	GO_FIRST	zero	Economy
5	-1.386	-1.844	-0.658	GO_FIRST	zero	Economy
6	-1.398	-1.844	-0.658	Indigo	zero	Economy
7	-1.386	-1.844	-0.658	Indigo	zero	Economy
8	-1.410	-1.844	-0.658	Air_India	zero	Economy
9	0.004	-1.844	-0.658	AirAsia	one	Economy
10	0.571	-1.844	-0.658	AirAsia	one	Economy
11	0.317	-1.844	-0.658	GO_FIRST	one	Economy
12	0.480	-1.844	-0.658	GO_FIRST	one	Economy
13	-1.178	-1.844	-0.658	Air_India	one	Economy
14	-1.398	-1.844	-0.630	Indigo	zero	Economy
15	-1.398	-1.844	-0.459	Indigo	zero	Economy
16	-0.587	-1.844	-0.454	GO_FIRST	one	Economy

Correlation Matrix:

Attribu...	airline ...	airline ...	airline ...	airline ...	airline ...	airline ...	source...	source...	source...	source...	source...	source...	destina...	destina...	destina...	destina...
airline ...	1	-0.042	-0.152	-0.051	-0.072	-0.107	0.034	-0.006	-0.017	0.032	-0.052	0.003	0.002	-0.024	0.029	-0.049
airline ...	-0.042	1	-0.205	-0.069	-0.098	-0.145	0.042	-0.040	0.022	0.012	-0.017	-0.024	-0.030	0.029	0.004	-0.030
airline ...	-0.152	-0.205	1	-0.249	-0.353	-0.523	-0.049	-0.007	0.032	-0.027	0.028	0.032	-0.016	0.030	-0.033	0.035
airline ...	-0.051	-0.069	-0.249	1	-0.118	-0.176	0.032	0.026	0.015	0.001	-0.024	-0.063	0.026	0.013	-0.001	-0.026
airline ...	-0.072	-0.098	-0.353	-0.118	1	-0.249	-0.014	-0.028	-0.011	0.018	0.011	0.033	-0.034	-0.014	0.033	0.003
airline ...	-0.107	-0.145	-0.523	-0.176	-0.249	1	0.013	0.037	-0.040	-0.003	0.003	-0.013	0.044	-0.036	-0.003	0.009
source_...	0.034	0.042	-0.049	0.032	-0.014	0.013	1	-0.255	-0.233	-0.218	-0.200	-0.194	0.069	0.075	0.040	0.014
source_...	-0.006	-0.040	-0.007	0.026	-0.028	0.037	-0.255	1	-0.231	-0.216	-0.199	-0.193	-0.250	0.058	0.057	0.045
source_...	-0.017	0.022	0.032	0.015	-0.011	-0.040	-0.233	-0.231	1	-0.197	-0.182	-0.176	0.059	-0.208	0.035	0.037
source_...	0.032	0.012	-0.027	0.001	0.018	-0.003	-0.218	-0.216	-0.197	1	-0.170	-0.165	0.053	0.047	-0.191	0.035
source_...	-0.052	-0.017	0.028	-0.024	0.011	0.003	-0.200	-0.199	-0.182	-0.170	1	-0.152	0.050	0.024	0.031	-0.161
source_...	0.003	-0.024	0.032	-0.063	0.033	-0.013	-0.194	-0.193	-0.176	-0.165	-0.152	1	0.042	-0.000	0.018	0.015
destinat...	0.002	-0.030	-0.016	0.026	-0.034	0.044	0.069	-0.250	0.059	0.053	0.050	0.042	1	-0.225	-0.220	-0.202
destinat...	-0.024	0.029	0.030	0.013	-0.014	-0.036	0.075	0.058	-0.208	0.047	0.024	-0.000	-0.225	1	-0.201	-0.185
destinat...	0.029	0.004	-0.033	-0.001	0.033	-0.003	0.040	0.057	0.035	-0.191	0.031	0.018	-0.220	-0.201	1	-0.181
destinat...	-0.049	-0.030	0.035	-0.026	0.003	0.009	0.014	0.045	0.037	0.035	-0.161	0.015	-0.202	-0.185	-0.181	1
destinat...	0.001	-0.030	0.028	-0.059	0.017	0.005	0.063	0.047	-0.016	0.010	0.026	-0.152	-0.196	-0.179	-0.175	-0.161



Linear Regression:



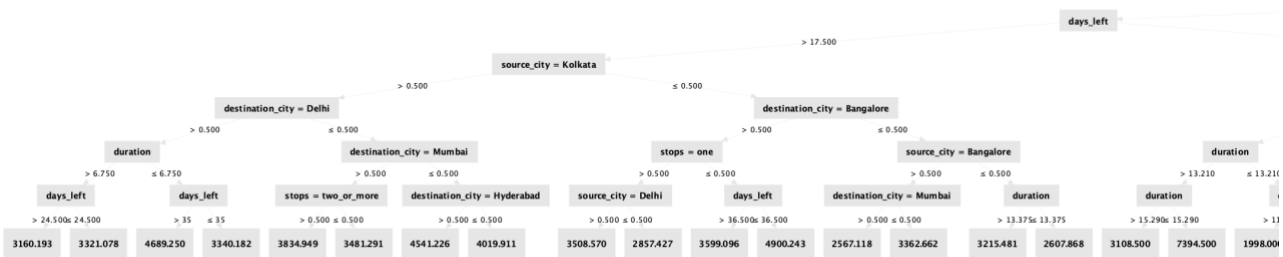
PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 6777.487 +/- 0.000
absolute_error: 4575.245 +/- 5000.146
squared_error: 45934333.248 +/- 122635399.471
correlation: 0.954
squared_correlation: 0.911
```

Performance Metrics:

- **Root Mean Squared Error (RMSE):** 6777.487
    - This is the average difference between predicted and actual prices.
    - Lower RMSE is better.
- **Absolute Error:** 4575.245
    - This is the average absolute difference between actual and predicted values.
- **Correlation:** 0.954
    - **This is very high**, meaning the model predicts flight prices quite well.
- **Squared Correlation (R^2):** 0.911
    - **91.1% of the variance in flight prices is explained by the model**, which is very good.

Decision Tree:



PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 4721.286 +/- 0.000
absolute_error: 2657.747 +/- 3902.169
squared_error: 22290542.844 +/- 78000545.499
correlation: 0.978
squared_correlation: 0.957
```

Decision Tree Regression, and here is how it compares to Linear Regression model:

Metric	Linear Regression	Decision Tree Regression	Better Model?
Root Mean Squared Error (RMSE)	6777.487	4721.286	Decision Tree
Absolute Error	4575.245	2657.747	Decision Tree
Correlation	0.954	0.978	Decision Tree
R^2 (Squared Correlation)	0.911	0.957	Decision Tree

Conclusion:

Decision Tree model is better than Linear Regression because:

- **Lower RMSE (4721.286 vs. 6777.487)** → More accurate predictions.
- **Higher R^2 (0.957 vs. 0.911)** → Explains more variance in flight prices.
- **Lower Absolute Error (2657.747 vs. 4575.245)** → More precise predictions.
- **The model is now ready for deployment to predict new flight prices.**

Decision Tree Regression is the best model for flight price prediction task