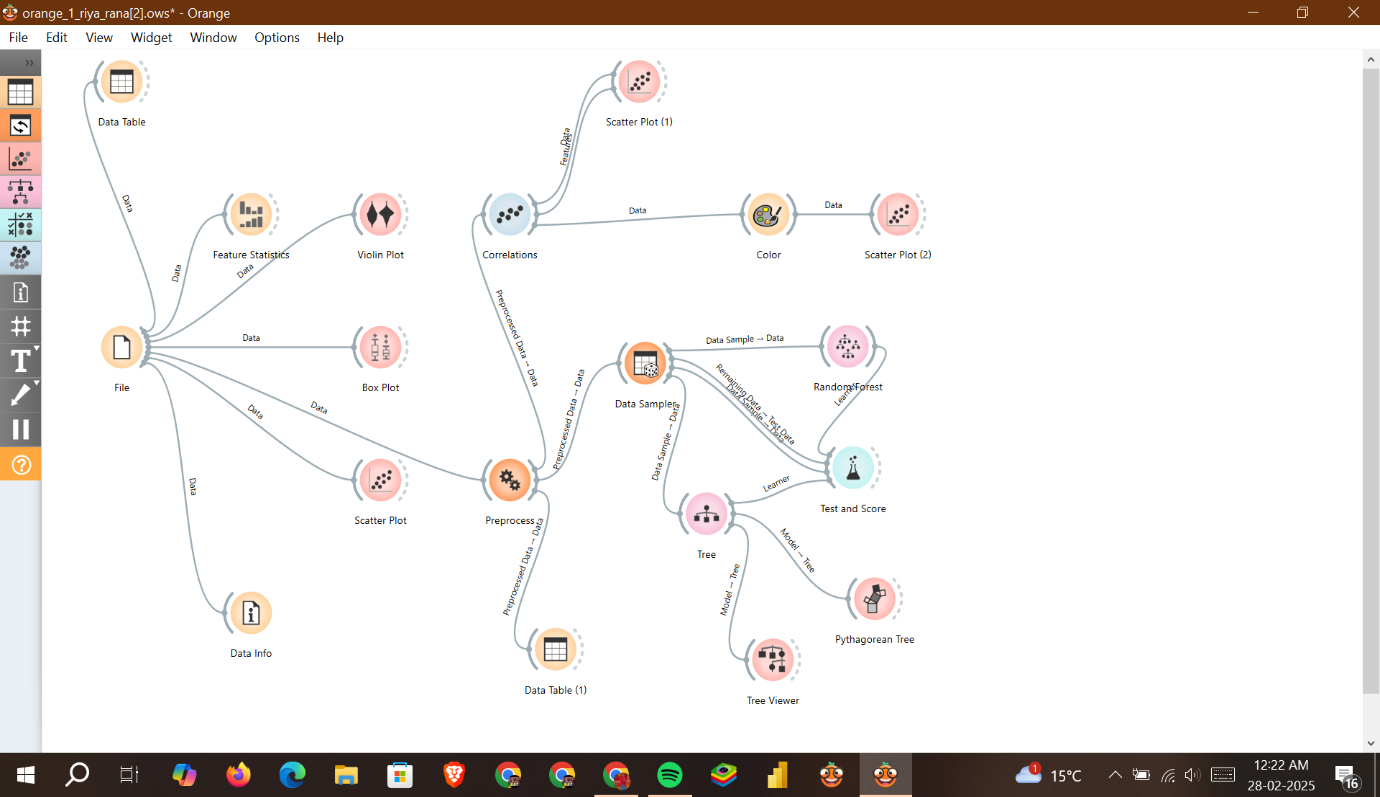
**PROJECT REPORT ON HR ANALYTICS DATASET USING ORANGE SOFTWARE**

**RIYA RANA**



**Preliminary Analysis of the Orange Workflow (.ows File)**

**This** **Orange workflow** consists of multiple **nodes (widgets)** that define a data pipeline. From are the XML structure:

**1. Identified Nodes (Widgets)**

* **File Input** → Loads data into Orange.
* **Data Table** → Displays raw data.
* **Data Info** → Provides dataset summary.
* **Feature Statistics** → Computes statistics for features.
* **Scatter Plot** → Visualizes data relationships.
* **Preprocess** → Likely used for cleaning, normalization, or feature selection.

**Analysis of the Orange Workflow**

**1 Workflow Breakdown**

**Data Input & Exploration**

* **File**: Loads the dataset into Orange.
* **Data Table**: Displays the dataset in tabular form.
* **Data Info**: Provides basic statistics about the dataset.
* **Feature Statistics**: Computes summary statistics for attributes.
* **Box Plot & Violin Plot**: Visualizes data distribution and variance.
* **Scatter Plot (1 & 2)**: Analyses relationships between variables.
* **Correlations**: Computes correlations between variables.

**Purpose:** This phase is focused on **data understanding and feature analysis** before modeling.

**2️ Data Preprocessing & Sampling**

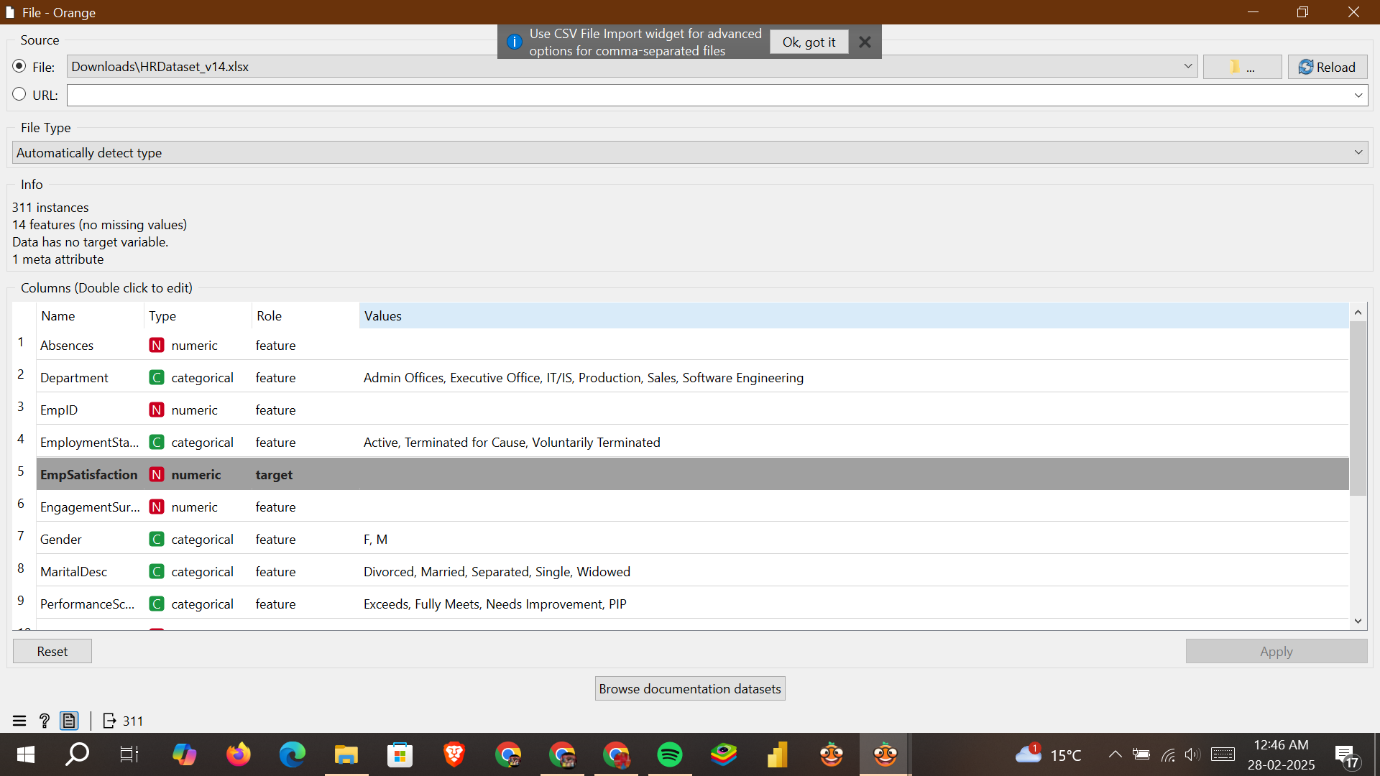
* **Preprocess**: Likely used for normalization, missing value handling, or feature selection.
* **Data Samples**: Splits data into training and testing sets.

**Purpose:** Ensures data is clean and ready for modelling.

**3️ Machine Learning Models**

* **Random Forest**: Trains a Random Forest model for classification or regression.
* **Decision Tree & Tree Viewer**: Constructs and visualizes decision trees.
* **Test and Score**: Evaluates model performance using metrics like accuracy or F1-score.

**Purpose:** Applies **supervised learning models (Random Forest, Decision Tree)** to make predictions.



**Dataset Analysis (HR Dataset for Employee Satisfaction Prediction)**

The dataset consists of **311 instances** and **14 features**, with no missing values. The **target variable** is **"EmpSatisfaction"**, which is numeric. The dataset includes a mix of **categorical** and **numeric** features related to employee demographics and performance.

|  |
| --- |
| Key Features in the Dataset |
| |  |  |  | | --- | --- | --- | | **Feature Name** | **Type** | **Description** | | **Absences** | Numeric | Number of absences recorded | | **Department** | Categorical | Employee's department (e.g., IT, Sales, Production) | | **EmpID** | Numeric | Unique identifier (likely not useful for modeling) | | **EmploymentStatus** | Categorical | Active, Terminated for Cause, Voluntarily Terminated | | **EmpSatisfaction** | Numeric (Target) | Employee satisfaction score (prediction goal) | | **EngagementSurvey** | Numeric | Employee engagement score | | **Gender** | Categorical | Male (M) or Female (F) | | **MaritalDesc** | Categorical | Marital status (Married, Single, etc.) | | **PerformanceScore** | Categorical | Performance rating (Exceeds, Meets, Needs Improvement, etc.) | |

**Feature Statistics Analysis (HR Dataset for Employee Satisfaction Prediction)**

**Key Observations from the Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Median** | **Mode** | **Min-Max** | **Dispersion** | **Remarks** |
| **Absences** | 10.24 | 10 | 4 | 1 - 20 | 0.57 | Slightly right-skewed, some employees have high absences. |
| **EngagementSurvey** | 4.11 | 4.28 | 5 | 1.12 - 5 | 0.19 | High engagement scores, minimal variation. |
| **PerfScoreID** | 2.98 | 3 | 3 | 1 - 4 | 0.20 | Most employees have a performance score of 3. |
| **Salary** | 69020.68 | 62810 | 57815 | 45046 - 250000 | 0.36 | Large salary variation, possibly influencing satisfaction. |
| **SpecialProjects** | 1.22 | 0 | 0 | 0 - 8 | 1.92 | Many employees have no special projects, but a few have up to 8. |
| **EmpSatisfaction** (Target) | 3.89 | 4 | 3 | 1 - 5 | 0.23 | Majority have satisfaction ratings between 3-5. |

**Insights for Employee Satisfaction Prediction**

**EmpSatisfaction is on a scale of 1-5**, meaning it could be **treated as both a regression and classification problem**:

* **Regression:** If we predict exact values (e.g., 3.89, 4.1).
* **Classification:** If grouped into categories (e.g., Low = 1-2, Medium = 3, High = 4-5).

**Salary variation is high**, which might mean salary could significantly impact employee satisfaction.  
**Absences might correlate with lower satisfaction**—further analysis needed with scatter plots and correlation matrices.

**EngagementSurvey scores are generally high**, meaning dissatisfaction may come from other factors like salary, workload, or job roles.

**Test and Score:**

**Test and Score** results compare the **Random Forest** and **Decision Tree** models in predicting **Employee Satisfaction**. Here’s a breakdown of key evaluation metrics:

|  |
| --- |
|  |
| **Model Performance Comparison** |
| | **Metric** | **Random Forest** | **Decision Tree** | **Remarks** | | --- | --- | --- | --- | | **MSE (Mean Squared Error)** | 0.812 | 1.323 | Lower is better → RF performs better. | | **RMSE (Root Mean Squared Error)** | 0.901 | 1.150 | RF has better predictive accuracy. | | **MAE (Mean Absolute Error)** | 0.775 | 0.925 | RF predictions are closer to actual values. | | **MAPE (Mean Absolute Percentage Error)** | 0.221 | 0.257 | RF has lower relative error. | | **R² Score (Coefficient of Determination)** | 0.031 | -0.07 | Both models have very low explanatory power. | |

**Insights from the Results**

* **Random Forest outperforms Decision Tree** in all metrics, meaning it's more reliable for prediction.
* **Low R² values (0.031 for RF, -0.07 for DT)** suggest the model **does not explain much variance** in satisfaction scores.
* **High RMSE (0.901 for RF)** means there is still considerable error in predictions.