**ORANGE PROJECT REPORT**

**DIYA RANA**

**Orange Workflow Analysis (Finance Classifier)**

Machine learning pipeline, incorporating: Logistic Regression , Confusion Matrix for evaluation  
 Feature selection & preprocessing improvements

**1 Data Preparation & Exploration**

* **File Input & Data Table** → Loads and displays the dataset.
* **Feature Statistics & Correlations** → Analyses variable relationships.
* **Box Plot & Scatter Plot (1 & 2)** → Visualizes distributions.

**2️ Preprocessing & Sampling**

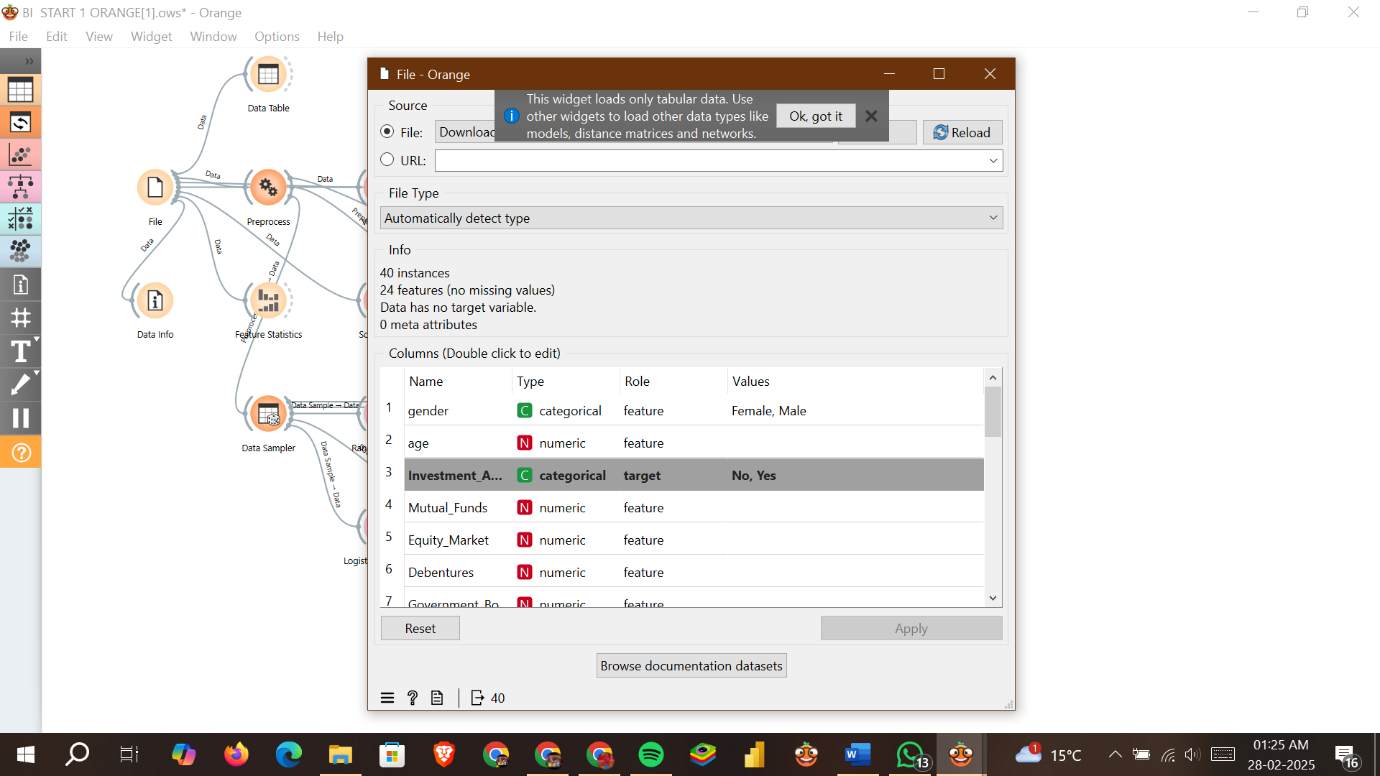
* **Preprocess** → Feature selection, normalization, encoding.
* **📊Data Sampler** → Splits data into training and test sets.

**3️ Machine Learning Models**

* **Random Forest**
* **Logistic Regression**
* **Decision Tree (with Tree Viewer & Pythagorean Tree for interpretation)**

**4️ Model Evaluation & Performance Testing**

* **Test & Score** → Compares model accuracy, error rates.
* **Confusion Matrix** → Evaluates classification results.



**Dataset Analysis (Investment Adoption Prediction)**

From your new **Orange dataset screenshot**, I see that this dataset is focused on predicting **Investment Adoption** (Yes/No) based on various financial and demographic features.

**Key Dataset Details**

40 instances (small dataset)  
24 features (mix of categorical & numeric variables)  
 Target Variable: Investment\_A (Yes/No) → Binary Classification Problem

**Key Features in the Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Type** | **Role** | **Description** |
| **Gender** | Categorical | Feature | Male/Female |
| **Age** | Numeric | Feature | Age of individual |
| **Mutual\_Funds** | Numeric | Feature | Investment in mutual funds |
| **Equity\_Market** | Numeric | Feature | Investment in equity markets |
| **Debentures** | Numeric | Feature | Investment in debentures |
| **Government\_Bonds** | Numeric | Feature | Investment in government bonds |
| **Investment\_A** | Categorical | Target | **(Yes/No)** – Whether the individual adopts investment |

**In Sights from the Dataset**

* **Binary Classification Task** → Logistic Regression, Random Forest, or Decision Trees are good choices.
* **Financial Variables Play a Key Role** → Investments in **mutual funds, equity, bonds, and debentures** likely influence the target outcome.
* **Small Sample Size (40 instances)** → Might lead to **overfitting** with complex models; cross-validation is crucial.

**Key Feature Insights (Feature Statistics)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Median** | **Mode** | **Dispersion** | **Min-Max** | **Remarks** |
| **Age** | 27.80 | 27 | 27 | 0.13 | 21 - 35 | Mostly young investors. |
| **Mutual Funds** | 2.55 | 2 | 2 | 0.46 | 1 - 7 | Some investors have high mutual fund investments. |
| **Equity Market** | 3.48 | 4 | 4 | 0.32 | 1 - 6 | Popular among investors. |
| **Debentures** | 5.75 | 6.5 | 7 | 0.29 | 1 - 7 | High average investment. |
| **Government Bonds** | 4.65 | 5 | 5 | 0.29 | 1 - 7 | Common investment choice. |
| **Fixed Deposits** | 3.58 | 3.5 | 3 | 0.50 | 1 - 7 | Moderate preference. |
| **PPF** | 2.02 | 1 | 1 | 0.78 | 1 - 6 | Least preferred investment. |

**Key Observations**

* Young investors (21-35 age group) dominate the dataset.
* Equity Markets, Debentures, and Government Bonds are the most common investments.
* Fixed Deposits & Mutual Funds have moderate interest, while PPF has the least.
* Target variable (Investment Avenue: Yes/No) suggests a binary classification approach**.**