A red and black logo

Description automatically generated

|  |
| --- |
| **Team:** |
| **Mahmoud Sayed** |
| **Omar Mohamed** |
| **Anas Ahmed** |

**Loan Approval**

**Classification**

**AI Final Project**

Table of Contents

[Introduction to AI Project 1](#_Toc186220046)

[Introduction to Dataset 1](#_Toc186220047)

[Dataset Attributes 2](#_Toc186220048)

[Project phases 3](#_Toc186220049)

[CSV data processing 4](#_Toc186220050)

[1. Data analysis 4](#_Toc186220051)

[ Distributions 4](#_Toc186220052)

[ Future statistics 4](#_Toc186220053)

[ Correlation 5](#_Toc186220054)

[ FreeViz 5](#_Toc186220055)

[2. Data Modeling 6](#_Toc186220056)

[ Preparing the data 6](#_Toc186220057)

[ Data Training 7](#_Toc186220058)

[ Data Testing 7](#_Toc186220059)

[3. Data Prediction 10](#_Toc186220060)

[10](#_Toc186220061)

[Image Processing 11](#_Toc186220062)

[1. Image Embedding and Training 11](#_Toc186220063)

[2. Data Testing 11](#_Toc186220064)

[3. Image Classification 12](#_Toc186220065)

[Results and Conclusion 13](#_Toc186220066)

[ Loan Approval Classification 13](#_Toc186220067)

[ Brain Tumor Image Classification 13](#_Toc186220068)

[References 14](#_Toc186220069)

Table of Figures

[1.The balance of the target 4](#_Toc186220070)

[2. The basic statistics for each feature 4](#_Toc186220071)

[3.The Correlation 5](#_Toc186220072)

[4. FreeViz 5](#_Toc186220073)

[5.Data Preparing 7](#_Toc186220074)

[6. Testing the training efficiency 8](#_Toc186220075)

[7.The confusion 8](#_Toc186220076)

[8.Remaining Data Prediction 9](#_Toc186220077)

[9. The Prediction 10](#_Toc186220078)

[10. Data Image Training 11](#_Toc186220079)

[11.Image Predication 12](#_Toc186220080)

[12.Image Classification 12](#_Toc186220081)

# Introduction to AI Project

**Artificial Intelligence** is the simulation of human intelligence processes by machines that are programmed to think and act like humans. In this AI project we will get more into machine learning but first let us differentiate between AI and machine learning.

**Machine Learning** is a subfield of artificial intelligence, which enables machines to learn from a dataset or experiences without being programmed. We will get deeper into machine learning by using **orange.**

Wesupply orange with large amounts of labeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states.

# Introduction to Dataset

**The first step in orange is to get a dataset, there’s two types of datasets:**

**csv dataset:** We will use **the Loan Approval Classification** Dataset so we can develop a model that can predict loan approval decisions. By analyzing the dataset and training it to predict whether a loan application will be approved or rejected based on a set of input features.

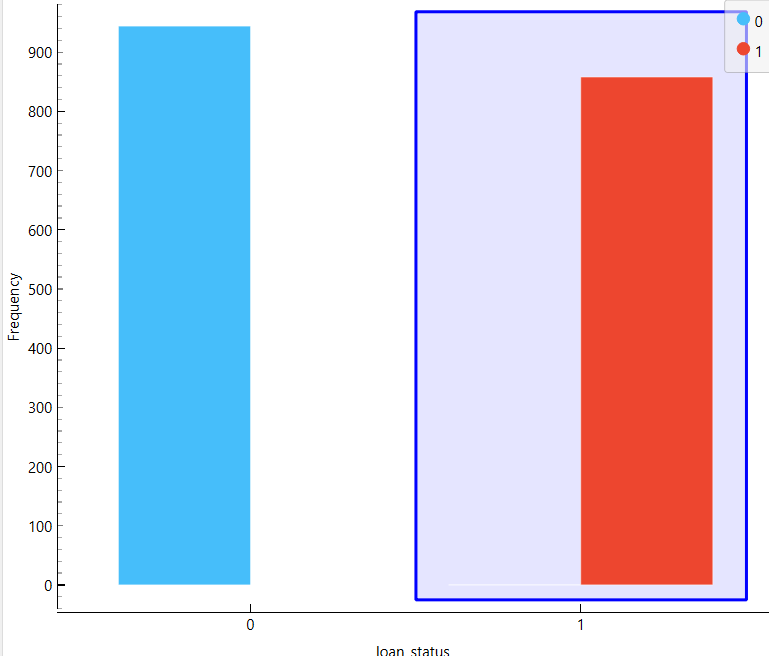
**Image dataset:** We will use **The Brain Tumor** Image Dataset, in this dataset we have a collection of MRI scans categorized into two classes: "Yes" (brain tumor detected) and "No" (no brain tumor detected). It is structured to help binary classification tasks for medical image analysis and can serve as a resource for training deep learning models aimed at identifying brain tumors from MRI scans.

# Dataset Attributes

1. **Person Age:** Younger people may have less financial stability, which could lead to higher rejection rates. Older applicants with a stable history are generally seen as lower risk.
2. **Person Gender:** Gender can affect the loan status indirectly due to many societal and economic factors.
3. **Person Education:** Higher education levels could correlate with better earning potential and financial stability, increasing loan approval chances.
4. **Person Income:** The higher the income, the higher the approval rate. On the other hand, low income can increase the chance of rejection.
5. **Person Employment Experience (Person Emp Exp):** More work experience can increase the approving rate.
6. **Person Home Ownership:** Homeowners might have a better approval chance due to their financial stability.Renters have a higher chance of being rejected.
7. **Loan Amount (Loan Amnt):** Higher loan amounts increase financial risk to the lender.
8. **Loan Intent:** the chance of getting approvedVaries by the intent type. Certain loan intents might have higher rejection rates due to the risks it comes with.
9. **Loan Interest Rate (Loan Int Rate):** Higher interest rates can lead to riskier loans that are more likely to be rejected.
10. **Loan Percent Income:** If the loan payment forms a large percentage of income, it leads to a higher rejection rate.
11. **Credit Bureau Person Credit History Length (CB Person Cred Hist Length):** A longer credit history provides lenders with more data to know the repayment behavior**,** people with shorter credit histories might get rejected.
12. **Credit Score:** Higher credit scores lower the risk and it increases the approving rate.
13. **Previous Loan Defaults on File:** Defaults indicate leads to a higher risk that increases the chances of rejection.
14. **loan status:** The target variable, representing whether a loan application was approved or rejected.

# Project phases

# CSV data processing



1.The balance of the target

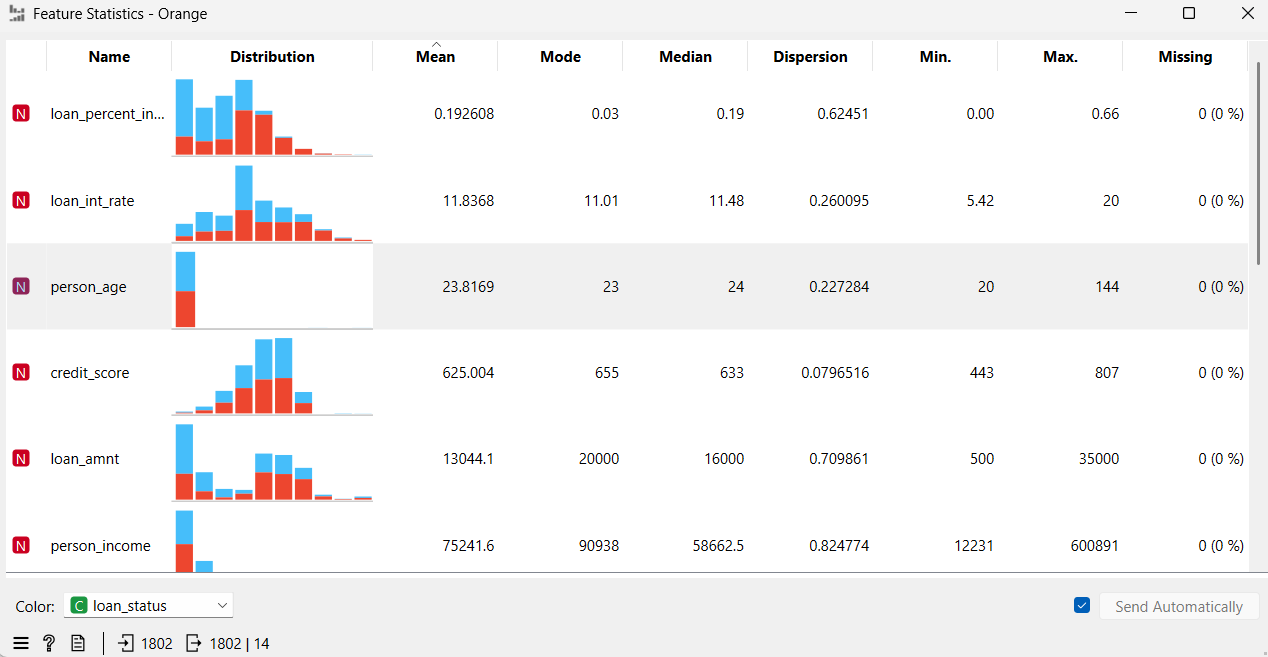
## Data analysis

The first thing we need to do is to analysis the dataset using specific widgets so we can understand it.

### Distributions

Using the Distributions widget, we can check the Balancing of the dataset as shown in the given Figure. Instances with (NO) Target represent 52% of the total number of instances while 48% of represent the instances with (YES) Target.

### Future statistics

From the Future statistics widget we Show the basic statistics for each feature like max, min, mean, median and missing values so it can help you in understanding the dataset.

2. The basic statistics for each feature

### A screenshot of a computer Description automatically generatedCorrelation

3.The Correlation

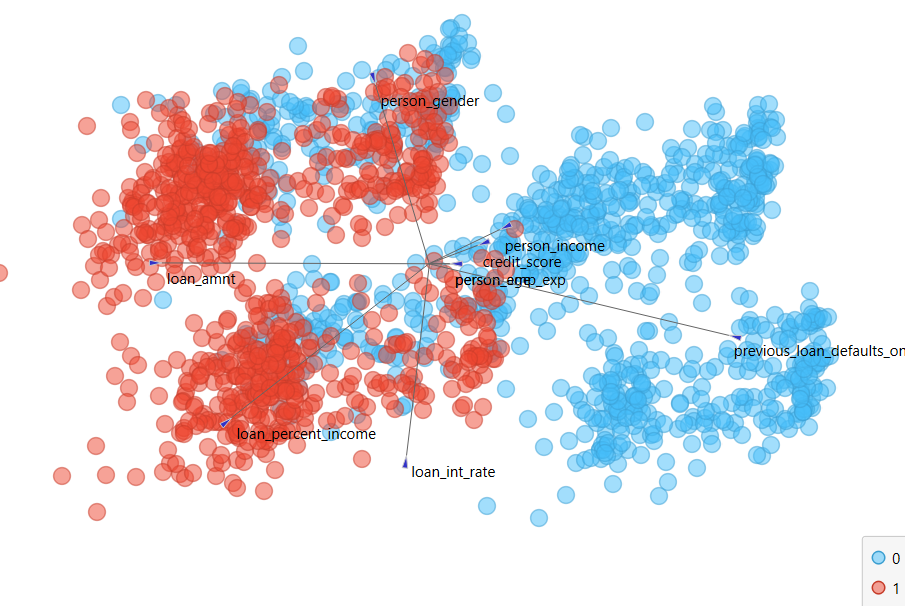
We can use Correlation widget to States whether there is a direct or inverse relation between features.

The +ve values means **direct** relation while –ve values means **inverse** relation, and if the correlation is greater than +0.7 or less than -0.7, that means it is a **strong** correlation.

For example, in this dataset, Loan Int Rate and Loan Amnt, as we see the correlation is +ve, that means it is a direct relation, but the correlation is less than +0.7 that means it is **weak** correlation.

### FreeViz

FreeViz analyses the features and state the most effective features on target. The arrow length presents how the feature is affecting the target as shown in figure below.



4. FreeViz

## Data Modeling

### Preparing the data

In this phase we prepare the dataset for the training and predictions.

there’re some required widgets we can use to prepare it for the training phase.

#### Impute

This widget Substitute any missing value using several methods.

This step will help you not to lose any record.

#### Preprocess

This step will increase the efficiency of analyzing the dataset and increase the accuracy of prediction by Normalizing any numeric field to a range from 0 to 1. This step will allow orange application to analyze the features more efficiently and will increase the accuracy of predictions.

#### Outliers

It removes any abnormal records that will affect both training and prediction phases by dividing the dataset into two proportions as follows:

**Inliers** / **Outliers**

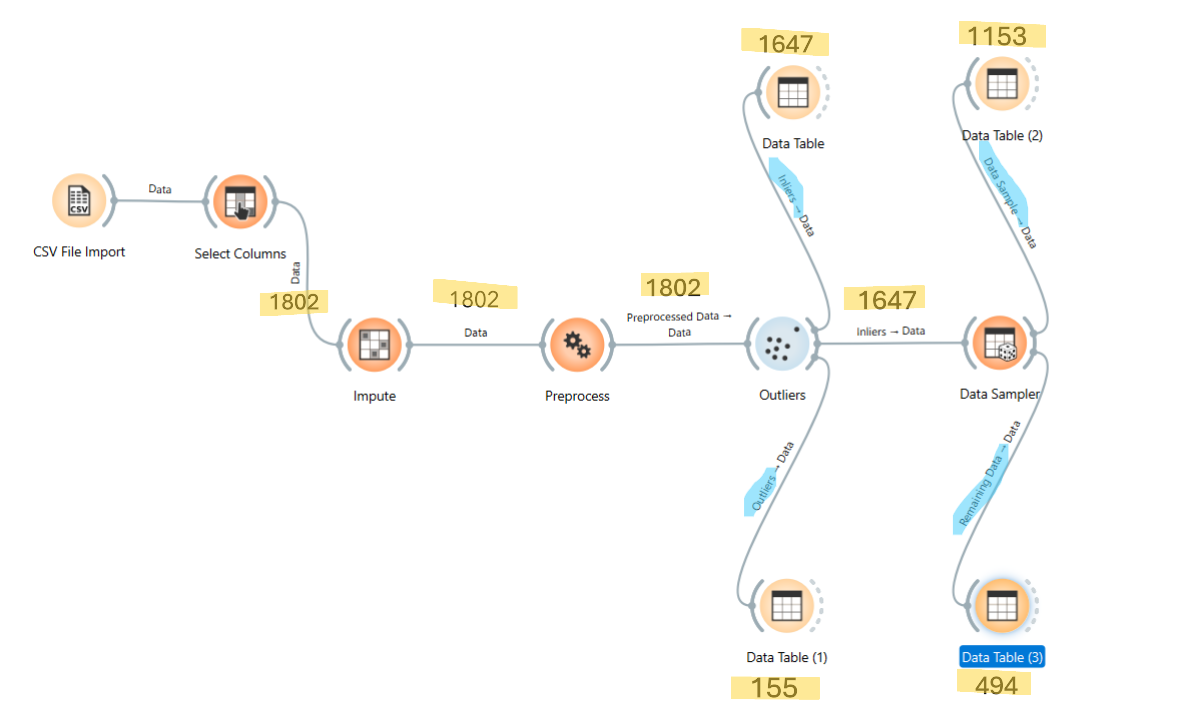
This step will increase the accuracy of prediction.

#### Data sampler

It divides the dataset into two proportions:

**Sample:** It will be used to train the model in the training phase.

**Remaining:** Used in prediction and test the prediction accuracy.



5.Data Preparing



After preparing the data, we need to train it and test its training efficiency before moving on to the prediction phase.

### Data Training

To train the data we will use Models, there’s so many models we that can use, in this project we will use: **KNN – SVM – Random Forest – Tree –**

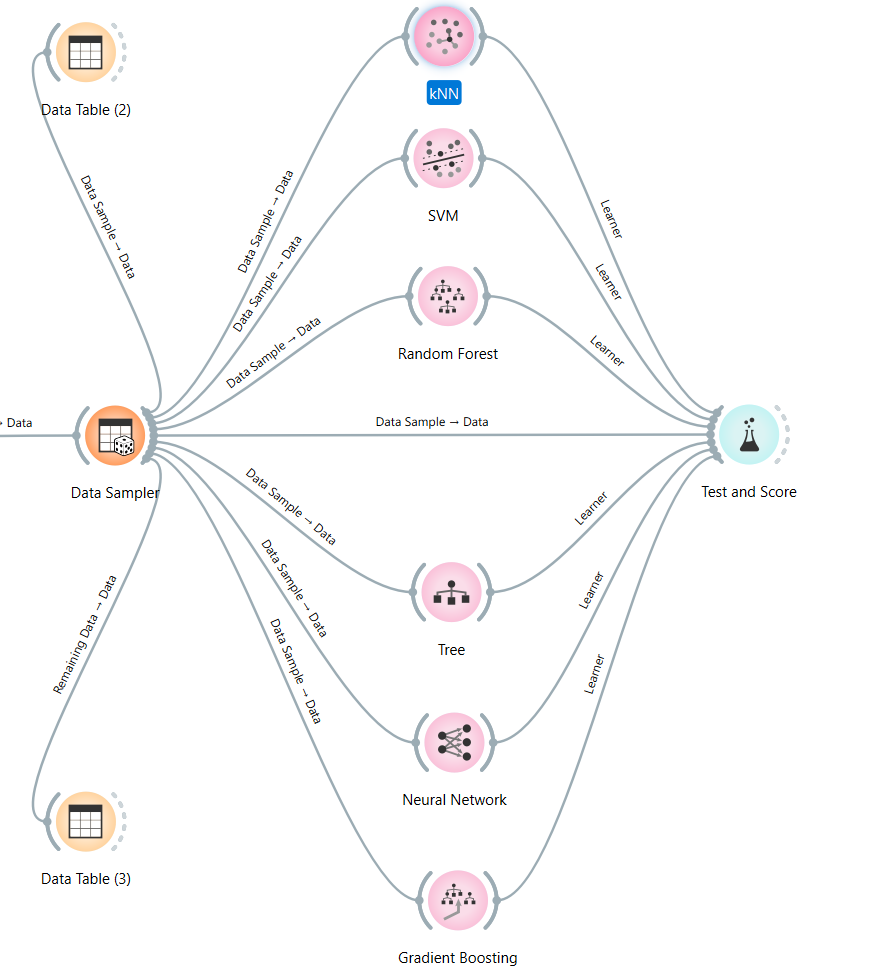
**Natural Network –** **Gradient Boosting.**

### Data Testing

To test the training efficiency, we going to use **“Test and Score”** widget, this widget will compare between **the trained data** and **the data sample**

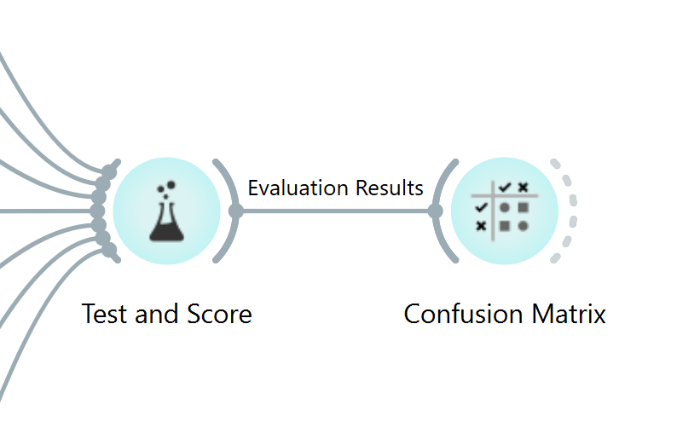
A screenshot of a computer

Description automatically generated

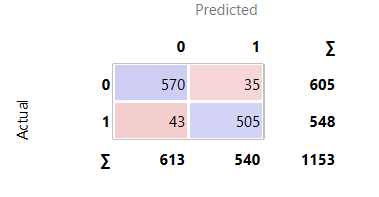


6. Testing the training efficiency



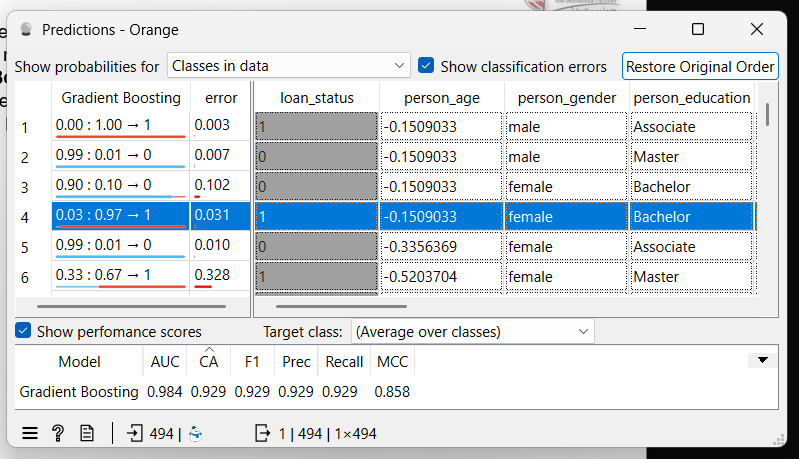
Now, we can notice that **Gradient Boosting** model has the highest **“CA”** that’s why we will rely on this model in training the sample data.

We can use "**Confusion Matrix”** to show the differences between actual data and predicted data. It displays how many instances were classified and misclassified. As shown in the below figure.



7.The confusion

**Now,** we need to predict the remaining instances whether their loan will get approved or not. To do so, we will use **“Predictions”** widget and connect it with **“Gradient Boosting”** model as it achieved the highest classification accuracy **“CA”,** and we will connect it also to **“Data Sampler”** and send only the “**Remaining Data**” to test the learning accuracy**.**

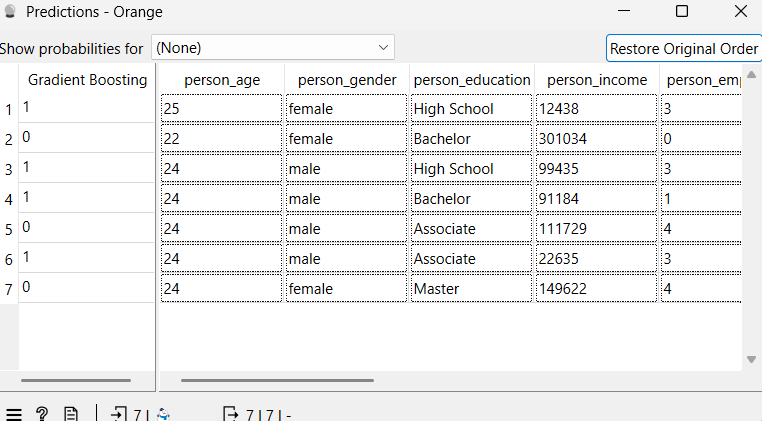


8.Remaining Data Prediction

Now, we save this model to use later in the prediction phase.

## Data Prediction

Now, we will open a new orange file to load the saved model and apply it on some new records.

from the figure below, we can see the predictions. **Notice** that there is no target field, as we will be predicting it,

9. The Prediction



# Image Processing

In this phase we are going to train orange to learn from image-based data and turn these raw images into suitable data that we can predict the type of it.

We are going to import “**Brain Tumor**” Image Dataset, in this dataset we have a collection of MRI scans that we are going to use to predict the type of the image: "Yes" (brain tumor detected) and "No" (no brain tumor detected).

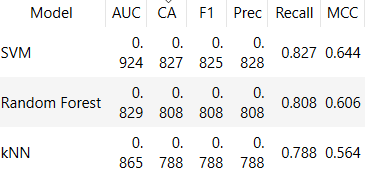
## A diagram of a networkImage Embedding and Training

10. Data Image Training

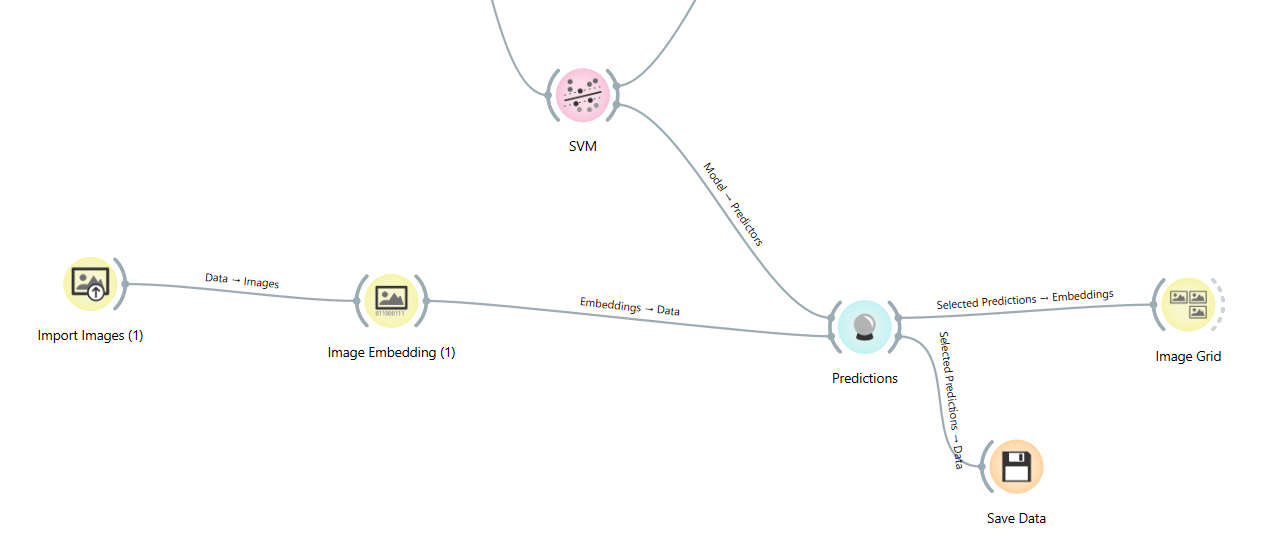
After importing the data, we are going to store the images as bits and bytes so we can train it using the models.

## Data Testing

To test the training efficiency, we going to use **“Test and Score”** widget

As we notice from the figure below “**SVM”** model has the highest **“CA”** that’s why we will rely on this model in training the data.

## Image Classification

Now that the data is fully trained, we will use a different brain tumor image than those used during the training to detect the presence of a brain tumor.

11.Image Predication

As we notice from the figure above, We connected **“Predictions”** widget and connected it with **“SVM”** model as it achieved the highest classification accuracy **“CA”,** and we connected it also to the new imported images after that we connect the predictions to **“image grid”** to see the prediction on each picture.



12.Image Classification

# Results and Conclusion

### Loan Approval Classification

We analyzed a dataset to build a model that can predict loan approval decisions. Key insights included:

1. **Balanced Dataset:** 52% of instances were labeled as "NO" and 48% as "YES".
2. **Preparing The Data:**

* Missing values were imputed
* Data was normalized
* Outliers were removed to increase prediction accuracy
* Data was sampled into training and prediction sets

1. **Training & Testing:** We used several models, with Gradient Boosting showing the highest accuracy. This model was then used for predictions.

### Brain Tumor Image Classification

We used MRI scans to train a model that can identify brain tumors. Highlights included:

1. **Image Processing:** MRI images were used to train various models.
2. **Best Model:** SVM model was the highest accuracy.
3. **Predictions:** The trained SVM model was then used to predict the presence of brain tumors in new images.

In summary, we successfully built a machine learning model that predict loan approval and detect brain tumor.

# **A blue letters on a transparent background Description automatically generated**References

Dataset from

orange datamining website