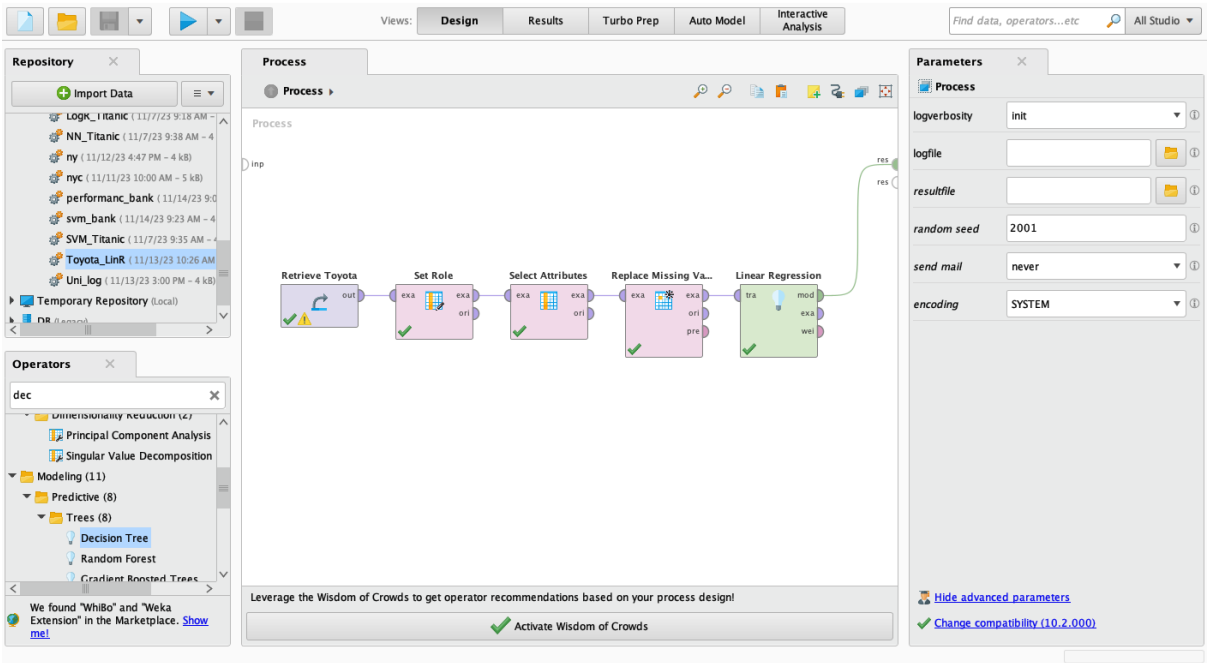
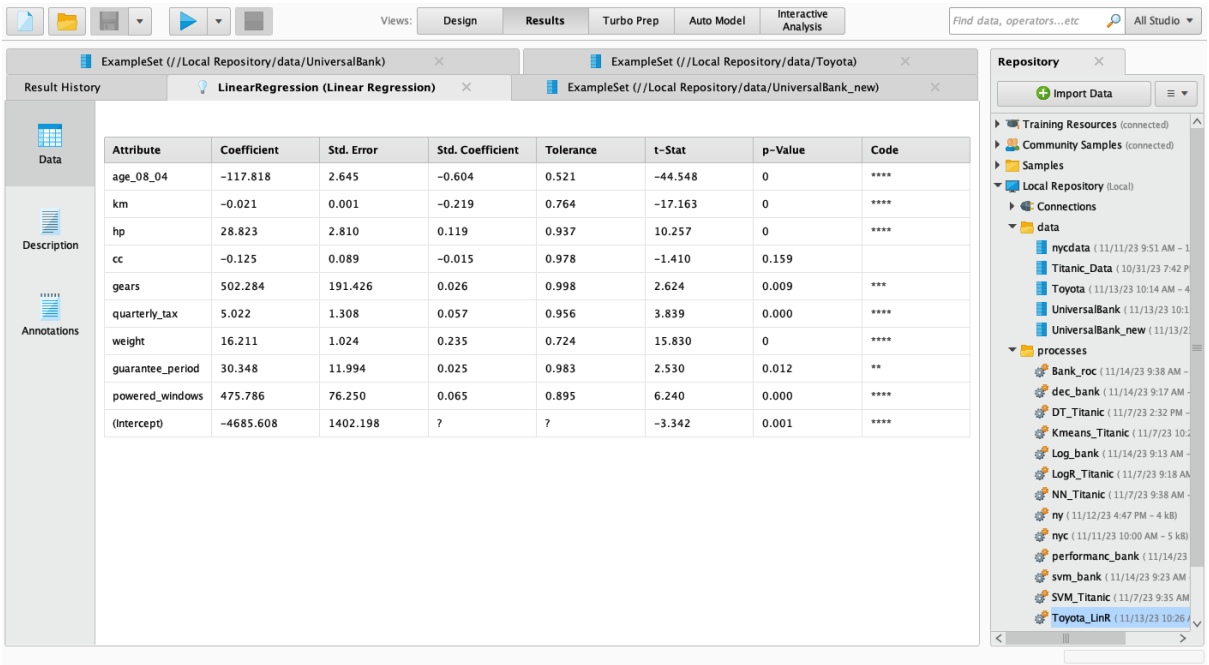
Assignment 5

Name: Komal Patil
G30080769

Task 1: Car Sales Data
Target variable: price
Selected Attributes: age_08_04, km, hp, cc, gears, quarterly_tax, weight, guarantee_period, powered_windows
Model 1: Linear Regression



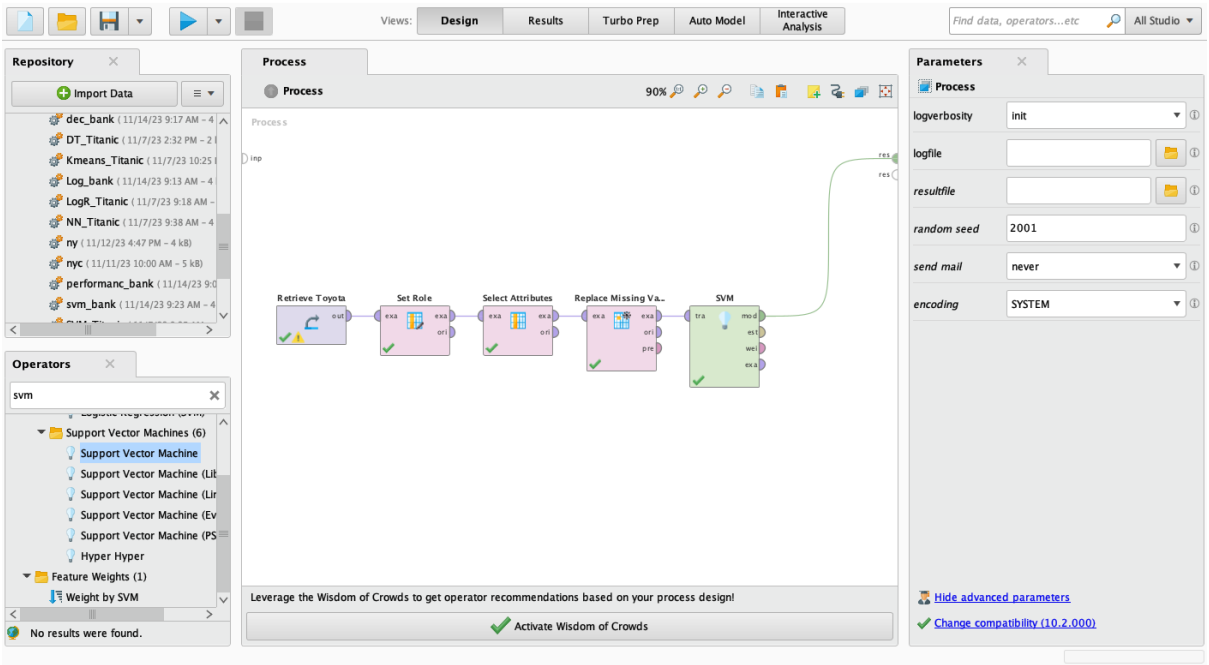
Result:



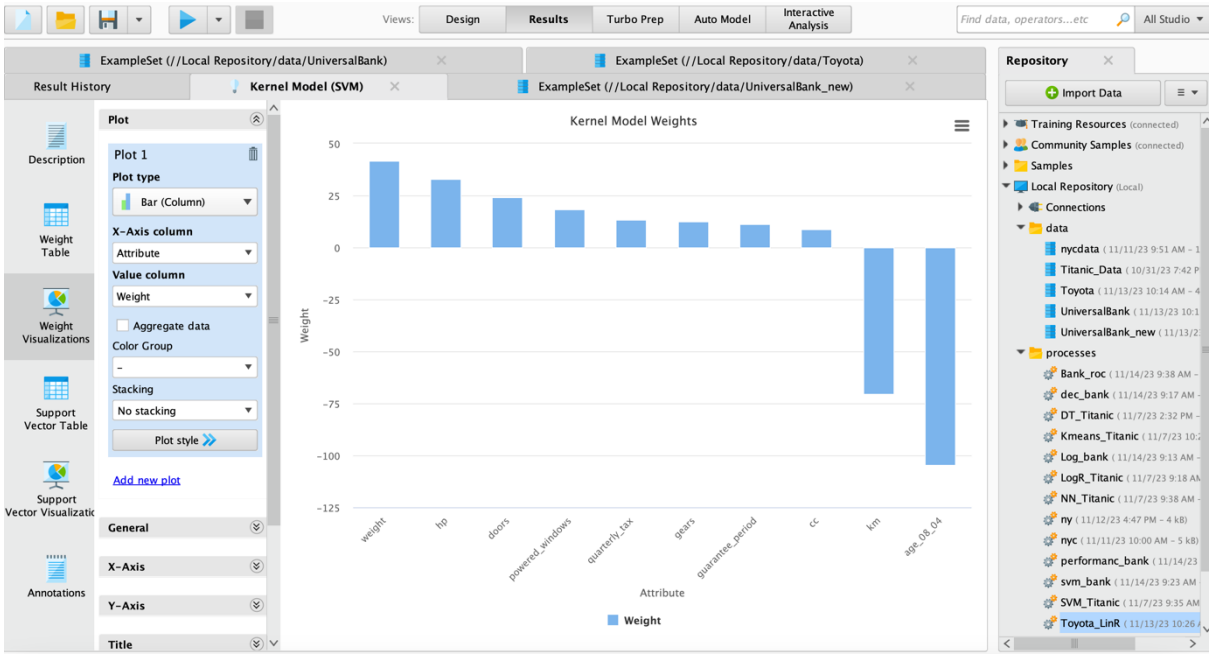
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
age_08_04	-117.818	2.645	-0.604	0.521	-44.548	0	****
km	-0.021	0.001	-0.219	0.764	-17.163	0	****
hp	28.823	2.810	0.119	0.937	10.257	0	****
cc	-0.125	0.089	-0.015	0.978	-1.410	0.159	
gears	502.284	191.426	0.026	0.998	2.624	0.009	***
quarterly_tax	5.022	1.308	0.057	0.956	3.839	0.000	****
weight	16.211	1.024	0.235	0.724	15.830	0	****
guarantee_period	30.348	11.994	0.025	0.983	2.530	0.012	**
powered_windows	475.786	76.250	0.065	0.895	6.240	0.000	****
(Intercept)	-4685.608	1402.198	?	?	-3.342	0.001	****

Interpretation:
All the attributes have p-values less than zero, indicating that they are all significant. HP is the most significant attribute among all of them due to its highest coefficient. It is normal to expect greater pricing for cars with additional gears, power windows, and horsepower. It is anticipated that cars with higher engine displacement, mileage, and quarterly taxes will cost less. It is anticipated that a car will start with 4685.608 points.

Model 2: Support Vector Machine (SVM)



Result:



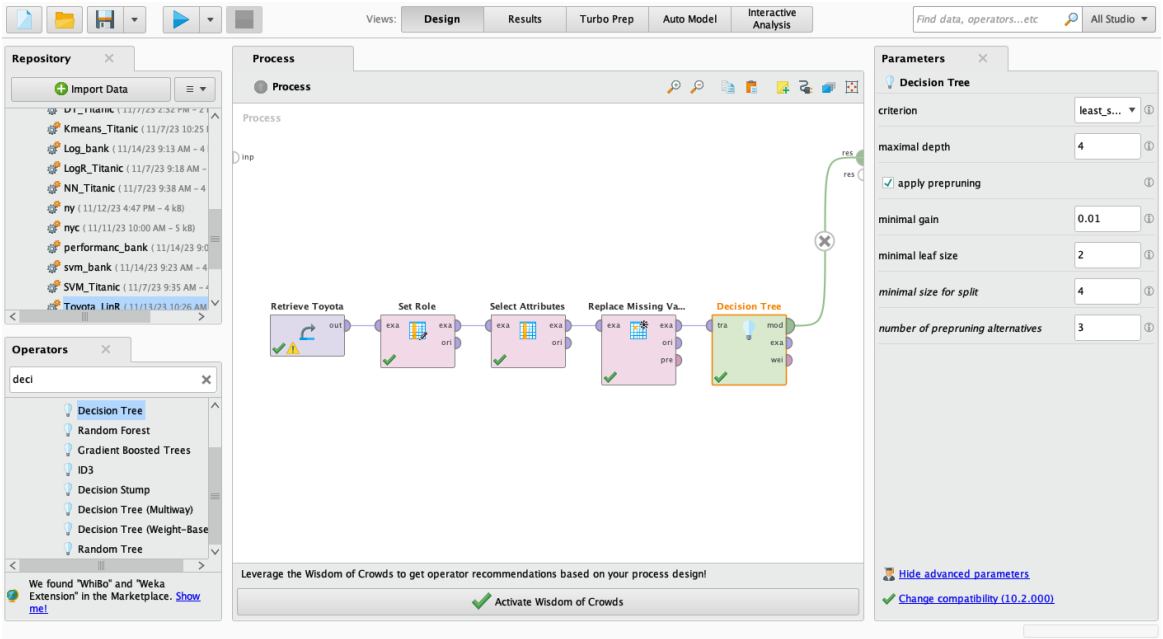
Interpretation:

- The age of the car (age_08_04) has the largest negative impact on the predicted price, followed by mileage (km) and horsepower (hp). This suggests that older cars with higher mileage and less powerful engines are expected to have lower prices.

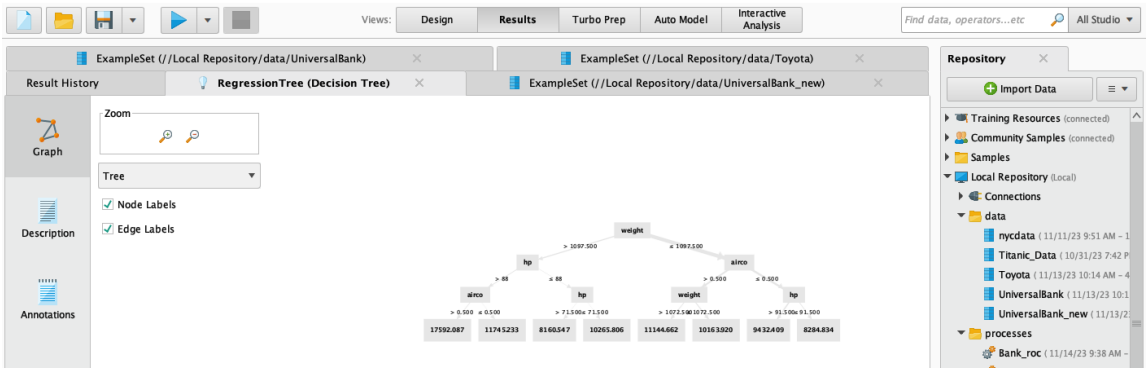
- Doors (doors), gears (gears), quarterly tax (quarterly_tax), weight (weight), and guarantee period (guarantee_period) also have some influence on the predicted price, but their impacts are smaller compared to the features mentioned earlier. This suggests that these features have a weaker impact on the predicted price.

- Cars with power windows (powered_windows) are expected to have a higher predicted price. This suggests that potential buyers are willing to pay a premium for this convenience feature.

Model 3: Decision Tree



Result:



Interpretation:

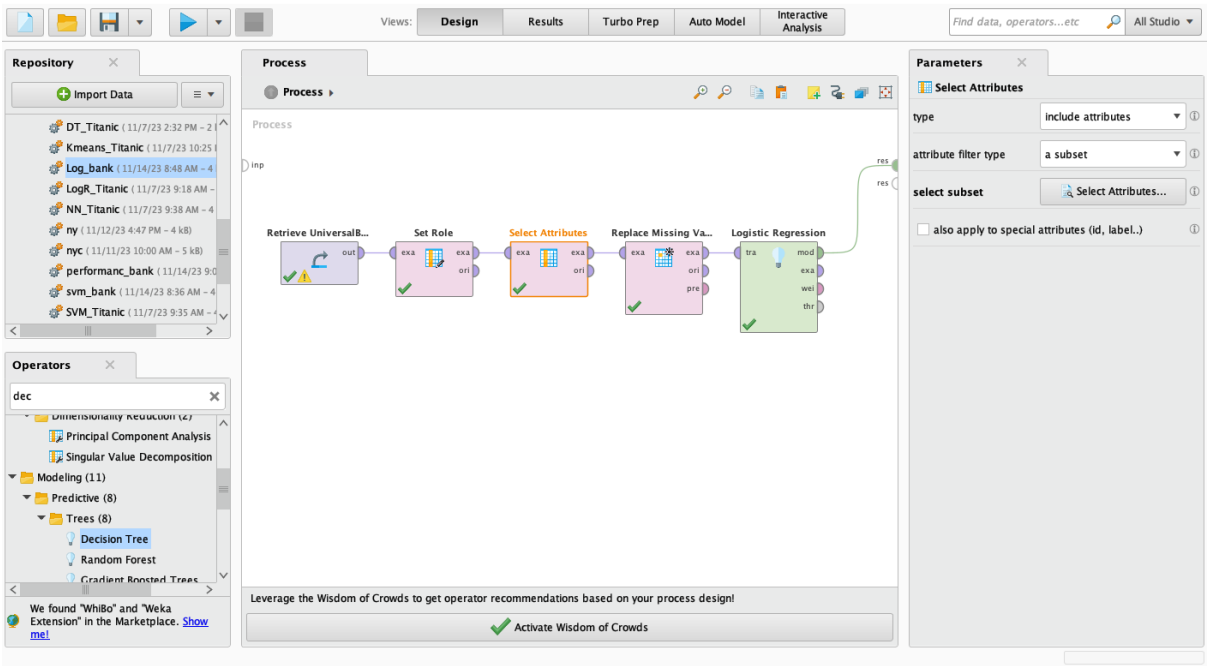
- Cars with a higher weight are expected to have a higher price.
- For cars with a higher weight, the presence of air conditioning is associated with a further increase in price.
- For cars with a lower weight, horsepower plays a more significant role in determining the price.

Choose your best model complexity and provide your logic:

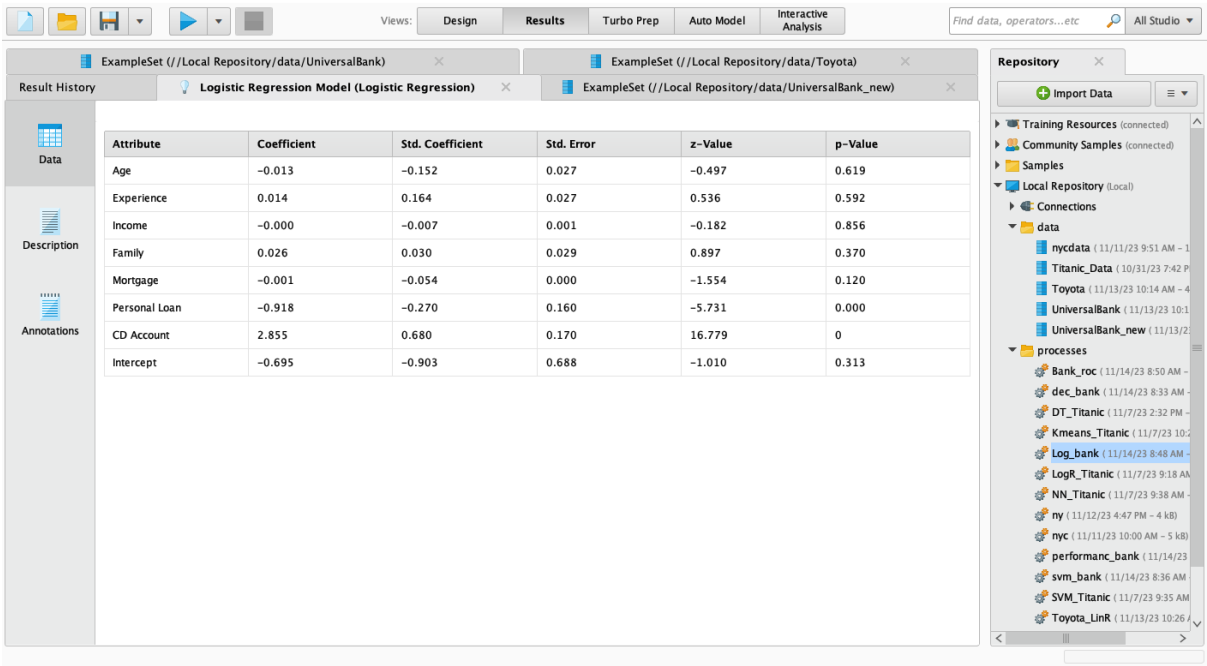
Since the features of the Toyota data were numerical in nature, simple linear regression, SVM and Decision tree were used for analysis. The intention was to avoid overfitting by keeping the model's complexity low. The **linear regression model** seems to be the most appropriate for estimating the cost of a car, based on the data presented. It provides interpretability with statistically significant coefficients and p-value, so that each attribute's effect on the target variable may be clearly understood.

Task 2: Bank Customer Data
Target variable: credit card
Selected Attributed: Age, Cd account, Experience, Family, income, Mortgage, Personal Loan

Model 1: Logistic Regression



Result:

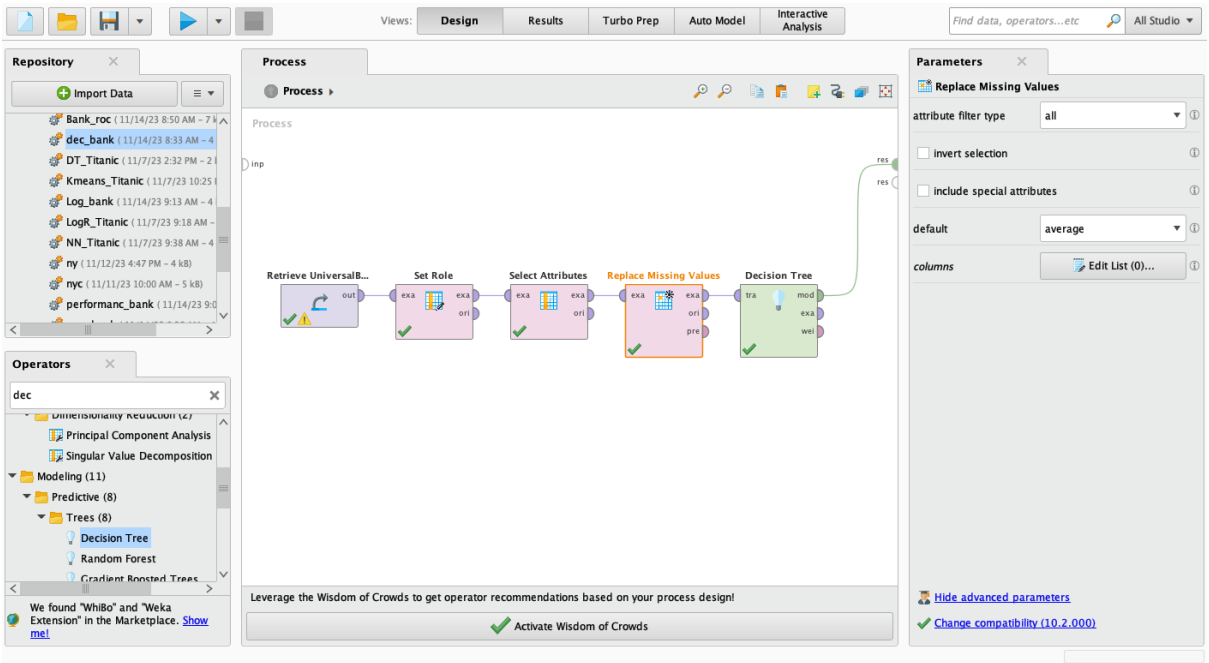


Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Age	-0.013	-0.152	0.027	-0.497	0.619
Experience	0.014	0.164	0.027	0.536	0.592
Income	-0.000	-0.007	0.001	-0.182	0.856
Family	0.026	0.030	0.029	0.897	0.370
Mortgage	-0.001	-0.054	0.000	-1.554	0.120
Personal Loan	-0.918	-0.270	0.160	-5.731	0.000
CD Account	2.855	0.680	0.170	16.779	0
Intercept	-0.695	-0.903	0.688	-1.010	0.313

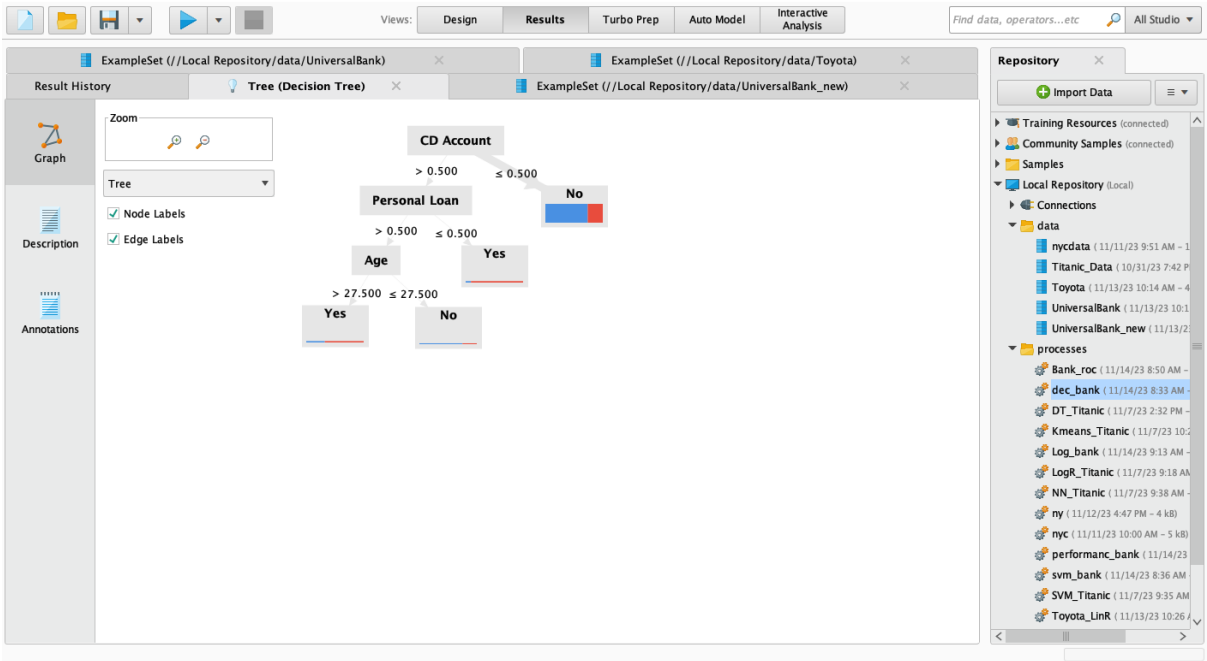
Interpretation:

The attributes of CD account, mortgage, family, and personal loan have p values less than 0.5, indicating their significance. Of these, CD account has the highest coefficient, making it the most significant. Given that the CD Account attribute's z-value is positive, it is an effective predictor of the customer's credit card membership.

Model 2: Decision Tree



Result:



Interpretation:

- Customers with a CD account are more likely to say "Yes" than customers without a CD account.

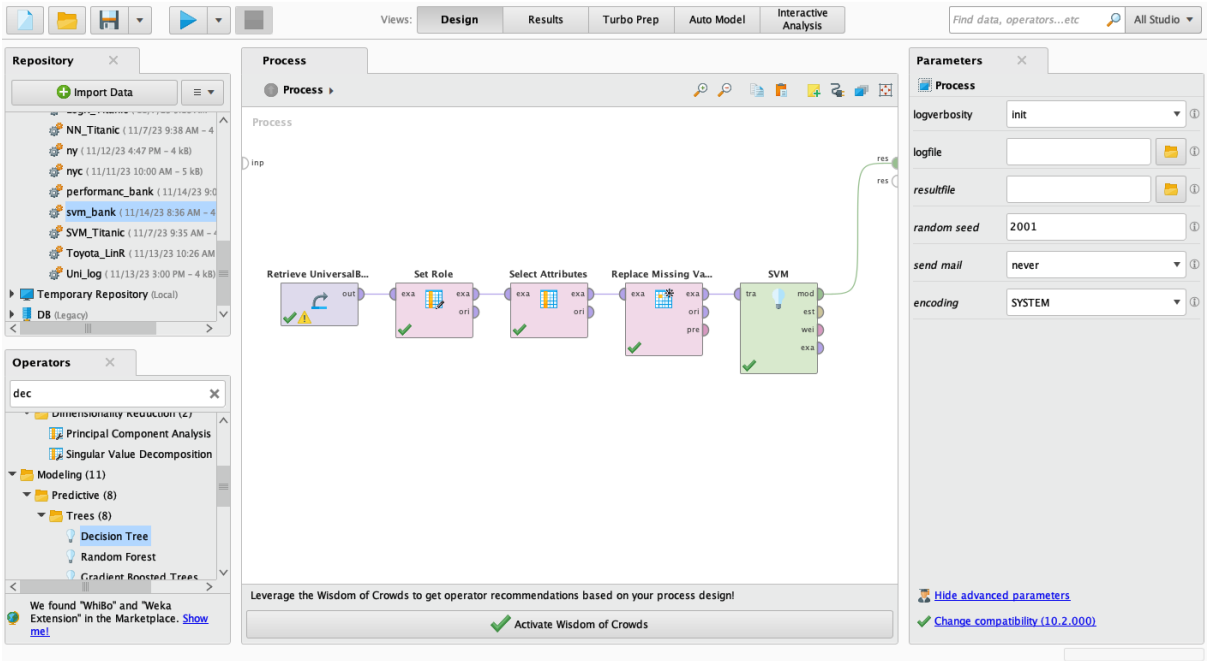
- Of the customers with a CD account, those who also have a personal loan are even more likely to say "Yes".

- For customers with a CD account and a personal loan, age is the deciding factor. Customers over 27.5 years old are more likely to say "Yes", while those 27.5 years old or younger are more likely to say "No".

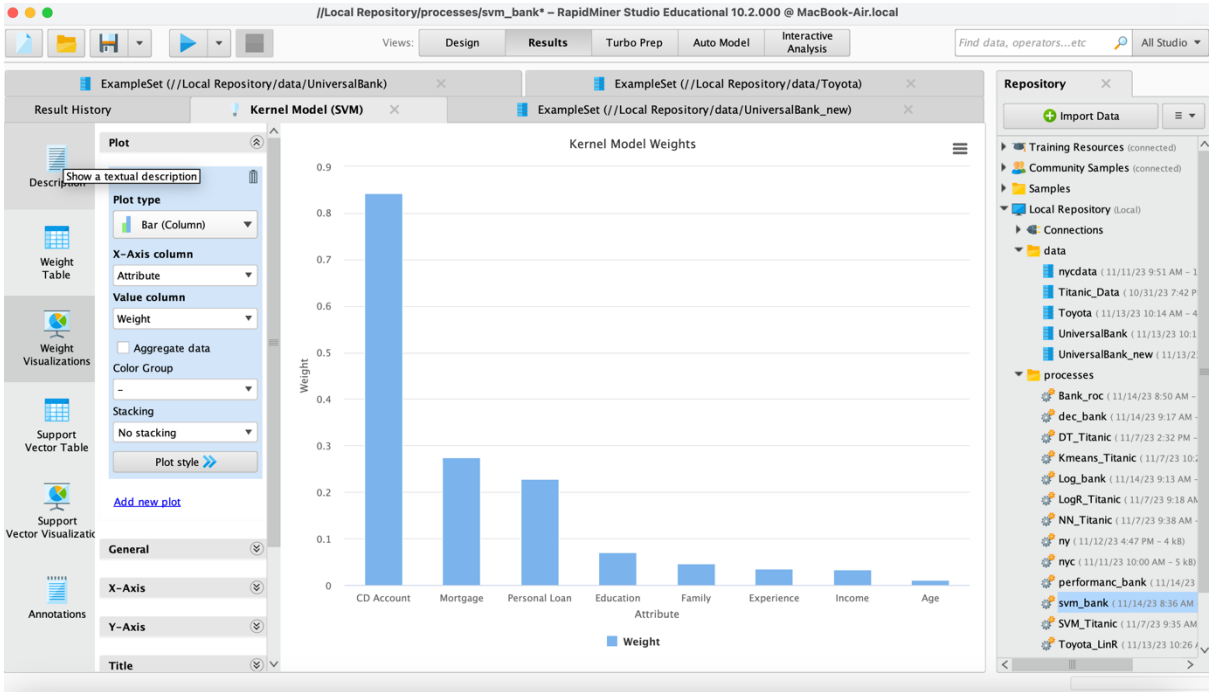
- Customers with a CD account and no personal loan are almost guaranteed to say "Yes".

- Customers without a CD account are more likely to say "No", but there are still a significant number who say "Yes".

Model 3: Support Vector Machine (SVM)



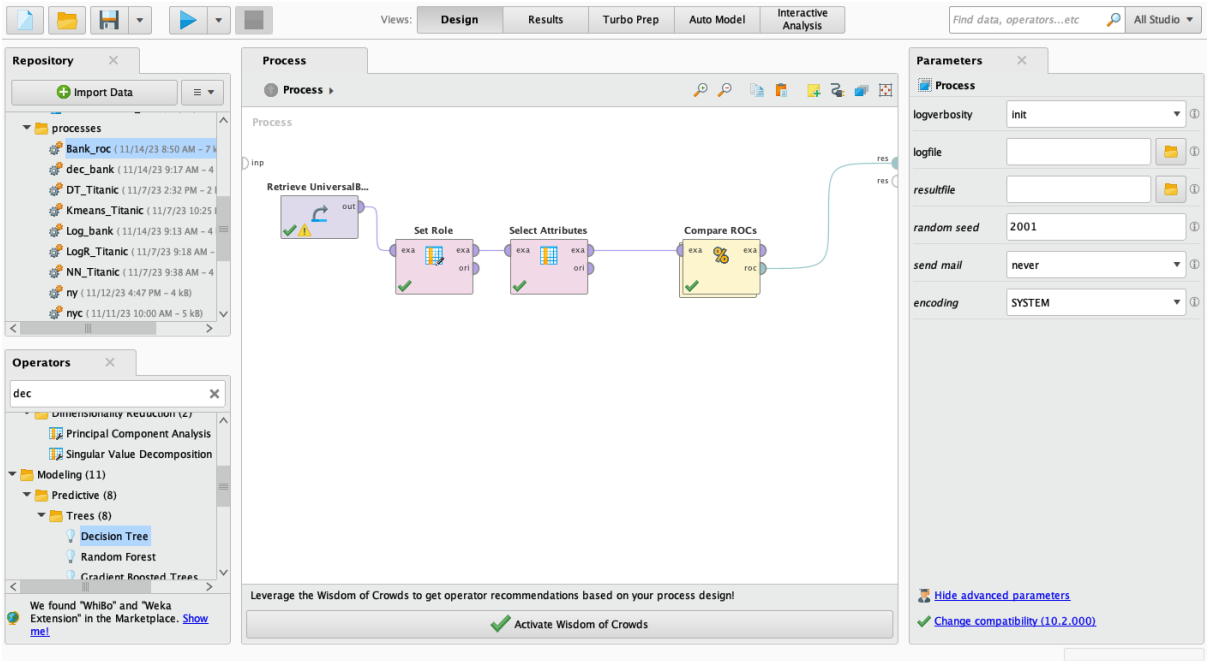
Result:

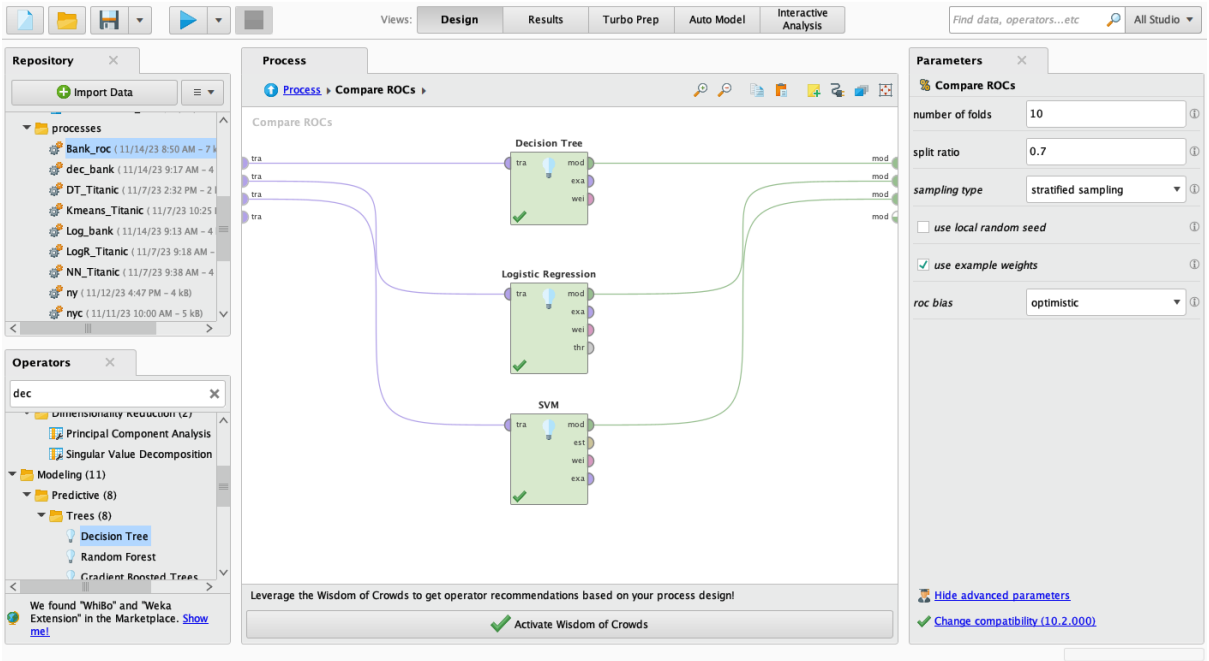


Interpretation:

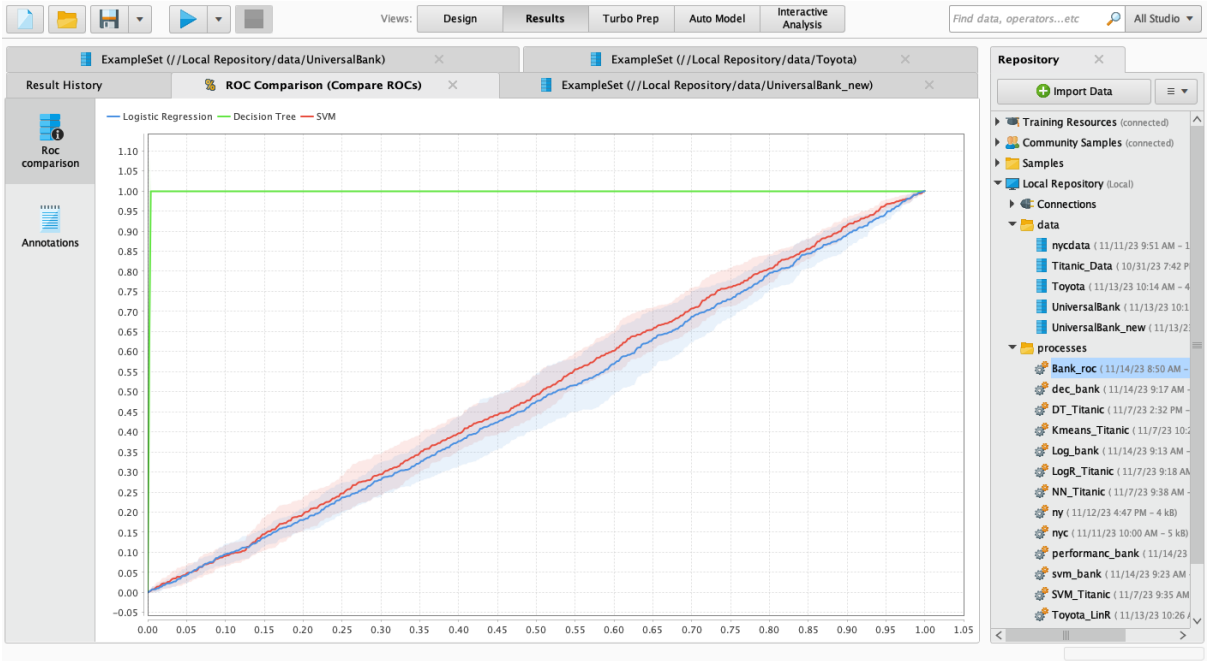
- CD account is the largest positive value (0.843), indicating that it has the most significant positive impact on the predicted probability of saying "Yes". This suggests that customers with a CD account are much more likely to respond positively to a particular offer or campaign compared to those without a CD account.
- Financial factors like having a mortgage or personal loan, higher income, and education also contribute to the likelihood of a positive response. Age and experience have a weaker but still statistically significant impact on the prediction.

ROC Curve



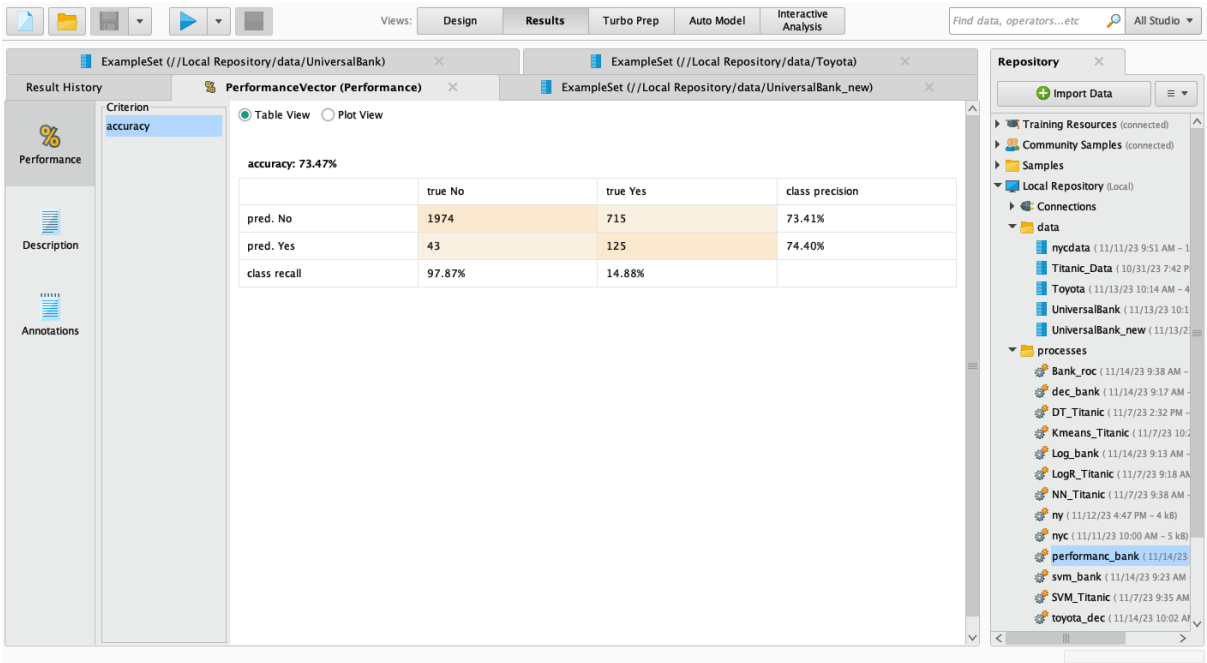


Result:



Interpretation:

The ROC curves order indicates that the decision tree, SVM, and logistic regression are the top three methods for categorizing positive and negative situations, respectively. Decision Trees have the greatest AUC, indicating that their accuracy is high at 73.47% and the class recall is 97.87%.



	true No	true Yes	class precision
pred. No	1974	715	73.41%
pred. Yes	43	125	74.40%
class recall	97.87%	14.88%	