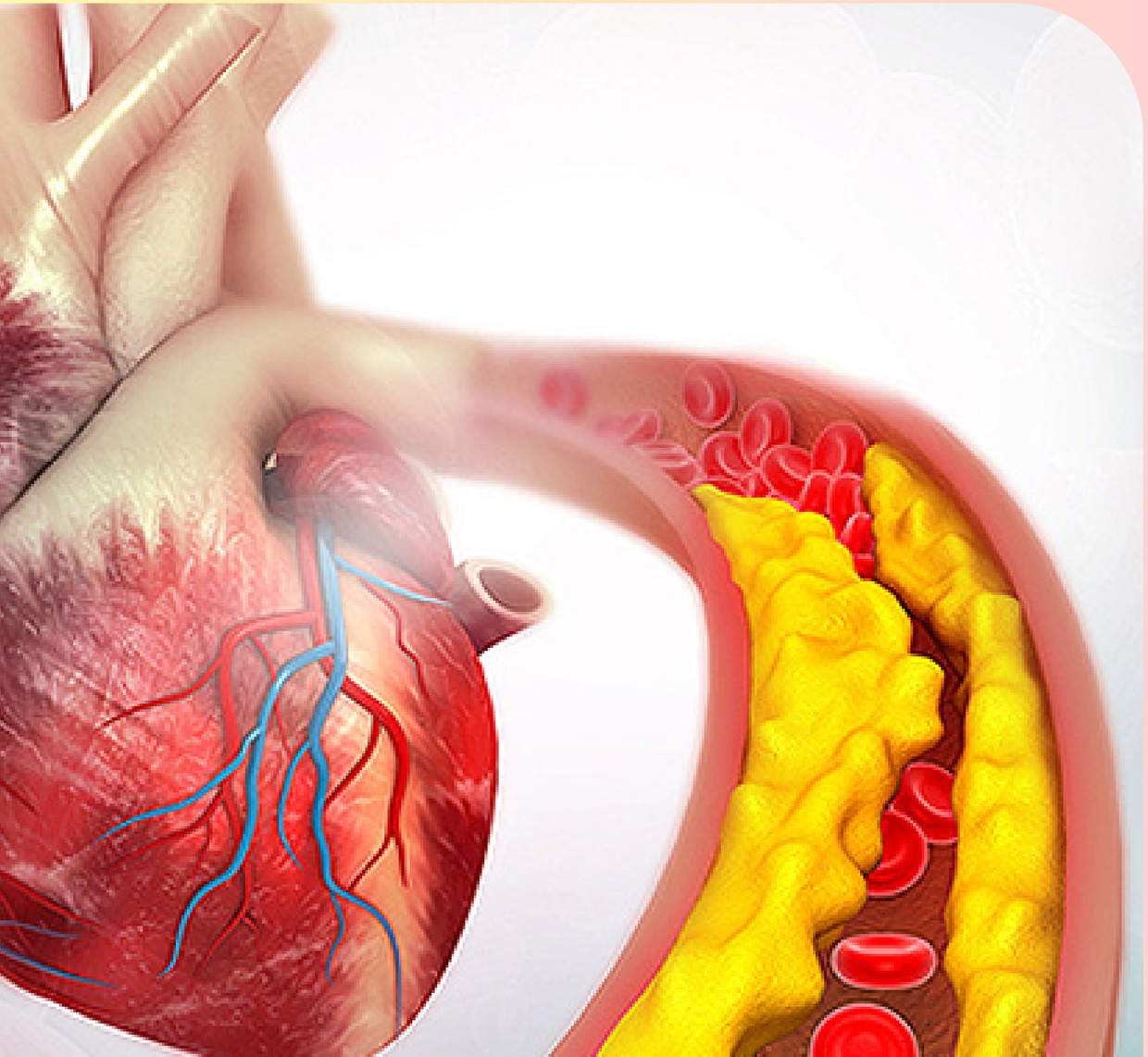
BUILDING PREDICTIVE MODELS FOR HEART DISEASE

Using Various Machine Learning Models

PRESENTED TO
MASTER_NEURON



Why should you care about Heart Disease?

- Heart disease is the leading cause of death in the U.S.
- 1 in 4 deaths
- $219 Billion
- All across the country

Back to Agenda

Back to Agenda

Proposed Solutions

Solution : 1

To Build a Model using Random Forest

Solution : 2

To Build a Model using Neural Network

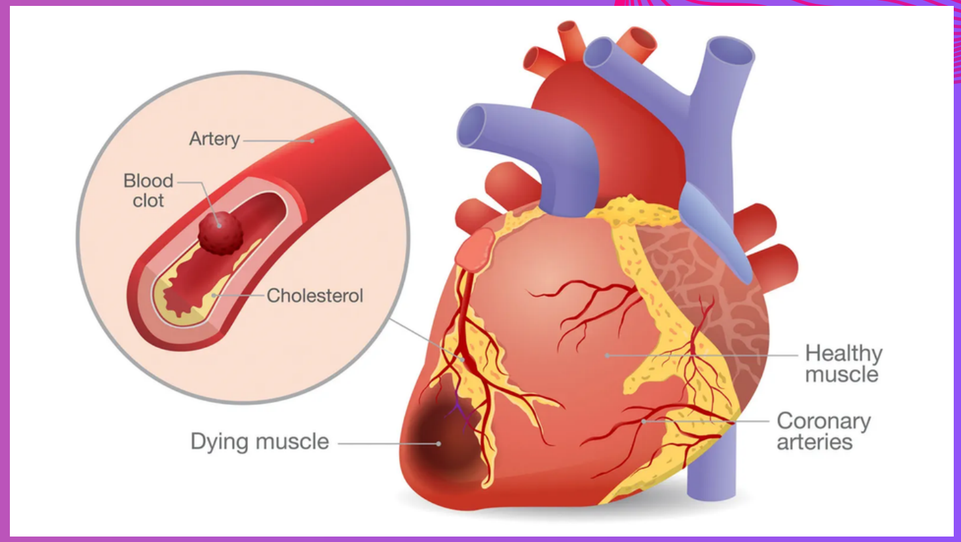
Solution : 3

To Build a model using Logistic Regression

Objective:

The goal of this project was to build predicitve models for coronary heart disease using the 2015 BRFSS dataset. I go through the process of getting the 2015 BRFSS dataset, selecting features for exploration in my predictive models based risk factors identified in past heart disease research, exploratory data analysis, model testing, and reporting on results. Methods explored in this, are: Random Forests, SVM,Decision Tree,Logistic Regression, and Neural Networks.

1



V/S

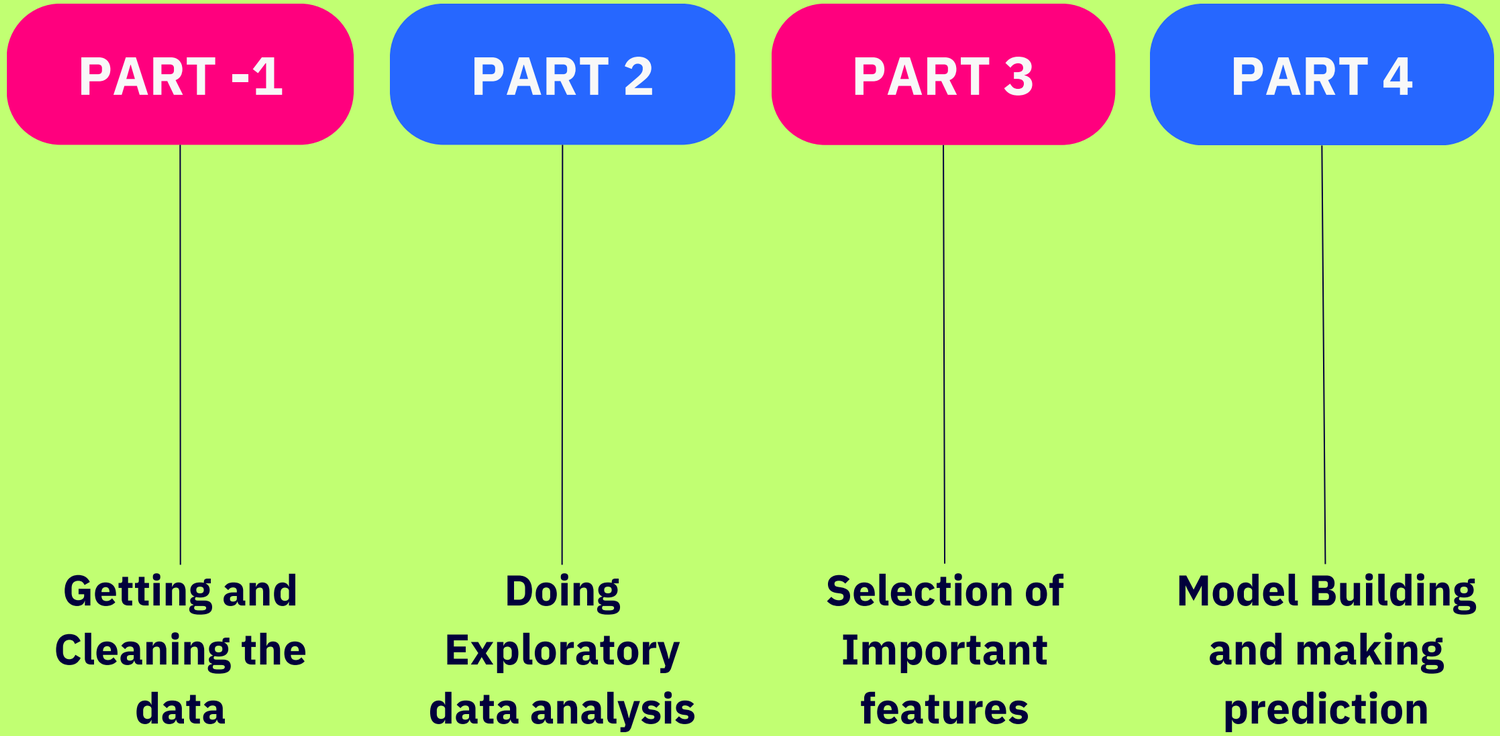
To make a model to predict the Heart Attack

0

Back to Agenda

Add Company Name

Proposed Model Work Flow



Back to Agenda

Know Your Dataset:

Back to Agenda

The Centers for Disease Control and Prevention (CDC) has developed the Behavioral Risk Factor Surveillance System (**BRFSS),** which is an annual telephone surveillance system conducted on a statewide basis. The primary objective of BRFSS is to monitor modifiable risk behaviors and other factors that contribute to the primary causes of morbidity and mortality within the population.

It contains 441,456 records and 330 columns.

Know Your Dataset:

Back to Agenda

Studies conducted in this field have identified crucial factors that play a significant role in the development of chronic illnesses such as diabetes and heart disease.

- blood pressure (high)
- cholesterol (high)
- smoking
- diabetes
- obesity
- age
- sex
- race
- Time since last checkup
- Mental Health
- diet
- exercise
- alcohol consumption
- BMI
- Household Income
- Marital Status
- Sleep
- Education
- Health care coverage

Preprocessing:

Back to Agenda

- ## Cleaning Of the data

- **Handing the Missing value:** We Drop Missing Values - knocks 100,000 rows out right away.
- **Do Encoding for Ordinal data:** We convert all the ordinal Data into numeric Form .
- **For BMI:** _BMI5 (no changes, just note that these are BMI * 100. So for example a BMI of 4018 is really 40.18)

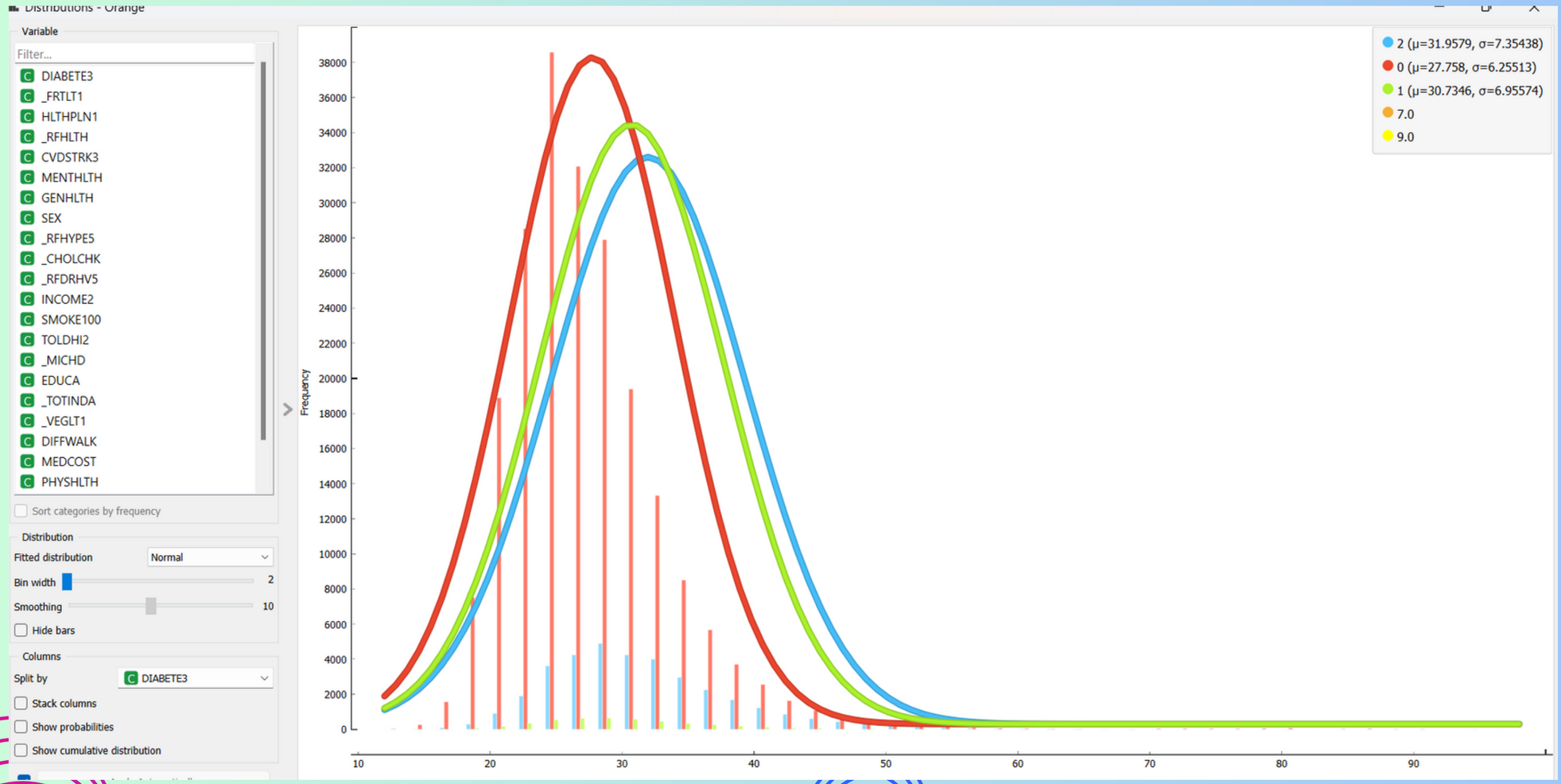
Look the Table:

Back to Agenda

	DIABETE3	_FRTLT1	HLTHPLN1	_RFHLTH	CVDSTRK3	MENTHLTH	GENHLTH	SEX	_RFHYPE5	_CHOLCHK	_RFDRHV5	INCOME2	SMOKE100	TOLDHI2	_
1	0	0	1.0	0	0	18.0	5.0	0	0	1.0	1.0	3.0	1.0	1.0	0
2	0	0	0	1.0	0	0	3.0	0	1.0	0	1.0	1.0	1.0	0	0
3	0	1.0	1.0	0	0	30.0	5.0	0	0	1.0	1.0	8.0	0	1.0	0
4	0	1.0	1.0	1.0	0	0	2.0	0	0	1.0	1.0	6.0	0	0	0
5	0	1.0	1.0	1.0	0	3.0	2.0	0	0	1.0	1.0	4.0	0	1.0	0
6	0	1.0	1.0	1.0	0	0	2.0	1.0	0	1.0	1.0	8.0	1.0	1.0	0
7	0	0	1.0	1.0	0	0	3.0	0	0	1.0	1.0	7.0	1.0	0	0
8	0	0	1.0	1.0	0	0	3.0	0	0	1.0	1.0	4.0	1.0	1.0	0
9	1	1.0	1.0	0	0	30.0	5.0	0	0	1.0	1.0	1.0	1.0	1.0	1.0
10	0	0	1.0	1.0	0	0	2.0	1.0	1.0	1.0	1.0	3.0	0	0	0
11	1	1.0	1.0	1.0	0	0	3.0	1.0	1.0	1.0	1.0	8.0	1.0	0	0
12	0	1.0	1.0	1.0	0	0	3.0	0	0	1.0	1.0	1.0	1.0	1.0	0
13	0	0	1.0	1.0	0	0	3.0	0	1.0	1.0	1.0	7.0	1.0	0	0
14	1	0	1.0	0	0	0	4.0	0	0	1.0	1.0	6.0	0	1.0	0
15	0	0	1.0	0	1.0	30.0	4.0	0	1.0	1.0	1.0	2.0	1.0	1.0	0
16	0	0	1.0	1.0	0	5.0	2.0	0	0	1.0	1.0	8.0	0	0	0
17	0	1.0	1.0	1.0	0	0	3.0	0	0	1.0	1.0	3.0	0	1.0	0
18	1	0	1.0	1.0	0	0	2.0	1.0	1.0	1.0	1.0	6.0	1.0	0	0
19	0	0	1.0	1.0	0	15.0	2.0	0	1.0	0	1.0	7.0	0	0	0
20	0	0	1.0	1.0	0	10.0	2.0	1.0	1.0	1.0	0	8.0	0	1.0	0
21	0	1.0	1.0	1.0	1.0	30.0	3.0	0	0	1.0	1.0	4.0	0	1.0	1.0
22	0	1.0	1.0	0	0	15.0	5.0	0	0	1.0	1.0	3.0	1.0	1.0	0
23	0	0	1.0	1.0	0	0	3.0	1.0	1.0	1.0	1.0	5.0	1.0	0	0
24	1	1.0	1.0	1.0	0	0	1.0	0	0	1.0	1.0	4.0	0	0	0
25	0	1.0	1.0	1.0	0	6.0	3.0	0	0	1.0	1.0	6.0	1.0	1.0	0
26	0	1.0	1.0	1.0	0	0	2.0	0	1.0	1.0	1.0	8.0	0	0	0
27	1	0	1.0	0	1.0	0	5.0	1.0	0	1.0	1.0	5.0	1.0	1.0	1.0
28	1	0	1.0	0	0	0	4.0	1.0	0	1.0	1.0	4.0	1.0	1.0	1.0
29	1	1.0	1.0	0	0	20.0	4.0	0	0	1.0	1.0	7.0	1.0	1.0	0
30	0	1.0	1.0	1.0	0	0	1.0	1.0	1.0	1.0	1.0	8.0	1.0	1.0	0
31	1	0	1.0	0	1.0	0	4.0	0	0	1.0	1.0	4.0	1.0	1.0	0
32	0	1.0	0	1.0	0	0	1.0	1.0	0	1.0	1.0	3.0	1.0	0	0
33	0	1.0	1.0	1.0	0	2.0	1.0	0	1.0	1.0	1.0	8.0	0	0	0
34	0	0	1.0	1.0	0	0	3.0	1.0	0	1.0	1.0	2.0	0	0	0
35	1	0	1.0	1.0	0	0	2.0	0	0	1.0	1.0	3.0	1.0	1.0	0
36	0	1.0	1.0	1.0	0	0	2.0	0	0	1.0	1.0	4.0	0	0	0

Exploratory data analysis (EDA):

Back to Agenda



Exploratory data analysis (EDA):

Back to Agenda

Handling Outliers

Data Table (2) (1) - Orange

Info
311289 instances (no missing data)
23 features
No target variable.
4 meta attributes

Variables
- Show variable labels (if present)
- Visualize numeric values
- Color by instance classes

Selection
- Select full rows

	IMONTH	IDAY	IYEAR	PCDMDECN	_FRTLT1	HLTHPLN1	_RFHLTH	CVDSTRK3	MENTHLTH	GENHLTH	SEX	_RFHYPE5	_CHOL(
1	b'01'	b'29'	b'2015'	b''	2	1	2	2	18	5	2	2	
2	b'01'	b'20'	b'2015'	b''	2	2	1	2	88	3	2	1	
3	b'01'	b'14'	b'2015'	b''	1	1	2	2	30	5	2	2	
4	b'01'	b'14'	b'2015'	b''	9	1	2	2	88	5	2	1	
5	b'01'	b'14'	b'2015'	b''	1	1	1	2	88	2	2	2	
6	b'01'	b'05'	b'2015'	b''	1	1	1	2	3	2	2	2	
7	b'01'	b'30'	b'2015'	b''	1	1	1	2	88	2	1	2	
8	b'01'	b'22'	b'2015'	b''	2	1	1	2	88	3	2	2	
9	b'01'	b'16'	b'2015'	b''	2	1	2	1	88	5	2	2	
10	b'01'	b'20'	b'2015'	b''	9	1	1	2	88	3	2	2	
11	b'01'	b'14'	b'2015'	b''	2	1	1	2	88	3	2	2	
12	b'01'	b'04'	b'2015'	b''	1	1	2	2	30	5	2	2	
13	b'01'	b'20'	b'2015'	b''	9	1	1	2	88	3	1	2	
14	b'01'	b'27'	b'2015'	b''	1	1	1	1	88	3	2	2	
15	b'01'	b'06'	b'2015'	b''	2	1	1	2	88	2	1	1	
16	b'01'	b'28'	b'2015'	b''	1	1	1	2	88	3	1	1	
17	b'01'	b'17'	b'2015'	b''	1	1	1	2	88	3	2	2	
18	b'01'	b'03'	b'2015'	b''	2	1	1	2	30	3	2	1	
19	b'01'	b'14'	b'2015'	b''	1	2	1	2	88	2	1	2	
20	b'01'	b'03'	b'2015'	b''	2	1	1	2	88	3	2	1	
21	b'01'	b'10'	b'2015'	b''	2	1	2	2	88	4	2	2	
22	b'01'	b'05'	b'2015'	b''	2	1	2	1	30	4	2	1	
23	b'01'	b'20'	b'2015'	b''	2	1	1	2	5	2	2	2	
24	b'01'	b'06'	b'2015'	b''	1	1	1	2	88	3	2	2	
25	b'01'	b'30'	b'2015'	b''	2	1	1	2	88	2	1	1	
26	b'01'	b'22'	b'2015'	b''	2	1	1	2	15	2	2	1	
27	b'01'	b'15'	b'2015'	b''	9	1	1	2	88	2	1	1	
28	b'01'	b'03'	b'2015'	b''	1	1	2	2	88	5	1	2	
29	b'01'	b'29'	b'2015'	b''	2	1	1	2	10	2	1	1	
30	b'01'	b'15'	b'2015'	b''	1	1	1	1	30	3	2	2	
31	b'01'	b'07'	b'2015'	b''	2	1	1	2	5	2	2	2	
32	b'01'	b'20'	b'2015'	b''	2	1	1	2	88	3	2	2	
33	b'01'	b'03'	b'2015'	b''	1	1	2	1	88	4	2	2	
34	b'01'	b'20'	b'2015'	b''	1	1	2	2	15	5	2	2	
35	b'01'	b'28'	b'2015'	b''	2	1	1	2	88	3	1	1	

Exploratory data analysis (EDA):

Back to Agenda

Feature Statistics:

Feature Statistics (1) - Orange

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
_FRTLT1		1.65163	1	0.925054	1	9	0 (0%)
HLTHPLN1		1.05463	1	0.269512	1	9	0 (0%)
_RFHLTH		1.19022	1	0.369692	1	9	0 (0%)
CVDSTRK3		1.96317	2	0.140713	1	9	0 (0%)
MENTHLTH		64.6324	88	0.555416	1	99	0 (0%)
GENHLTH		2.55453	2	0.427734	1	9	0 (0%)
SEX		1.5717	2	0.31484	1	2	0 (0%)
_RFHYPE5		1.44297	1	0.367786	1	9	0 (0%)
_CHOLCHK		1.07536	1	0.545289	1	9	0 (0%)

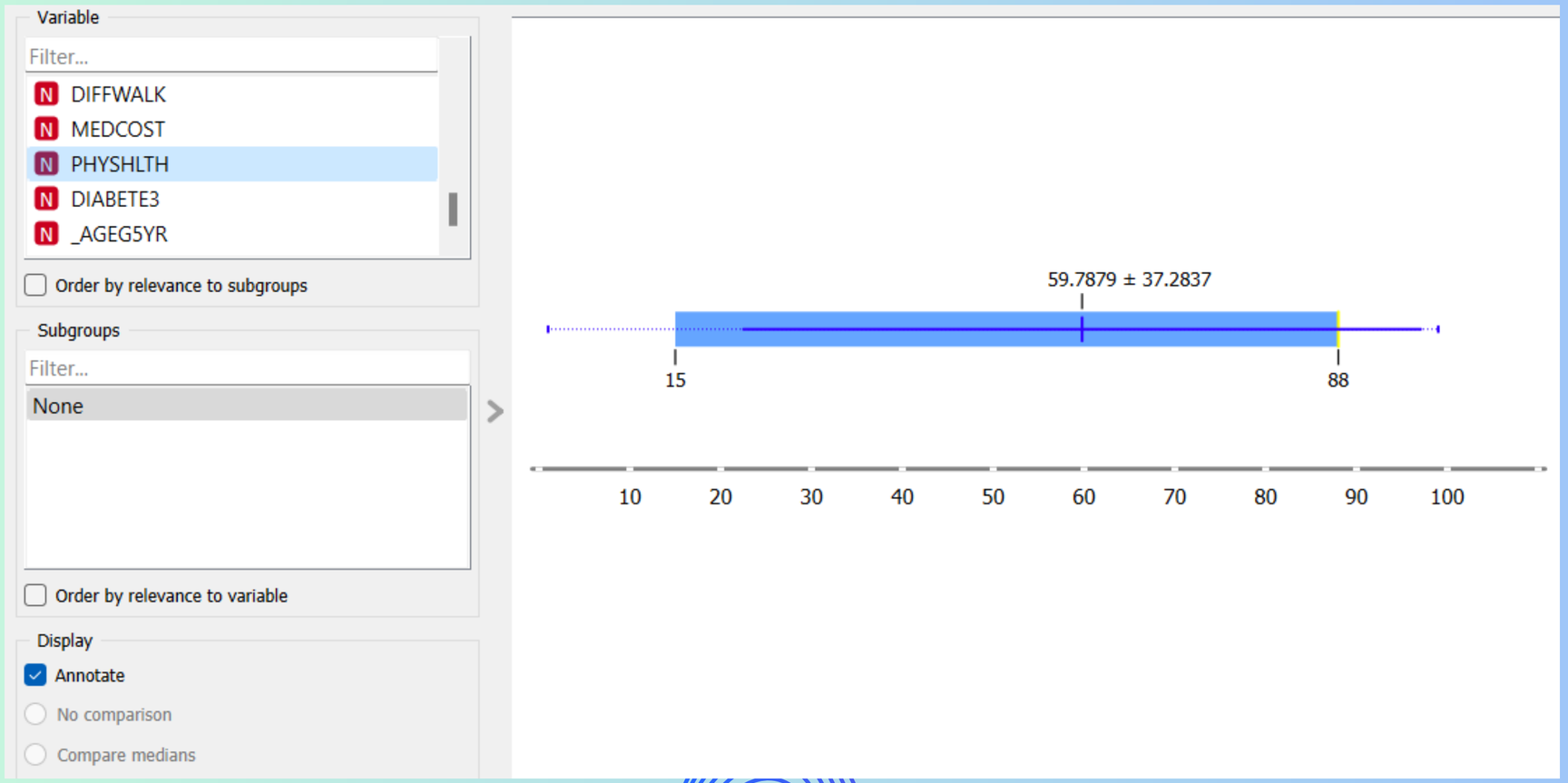
Color: _RFHLTH

Send Automatically

Exploratory data analysis (EDA):

Back to Agenda

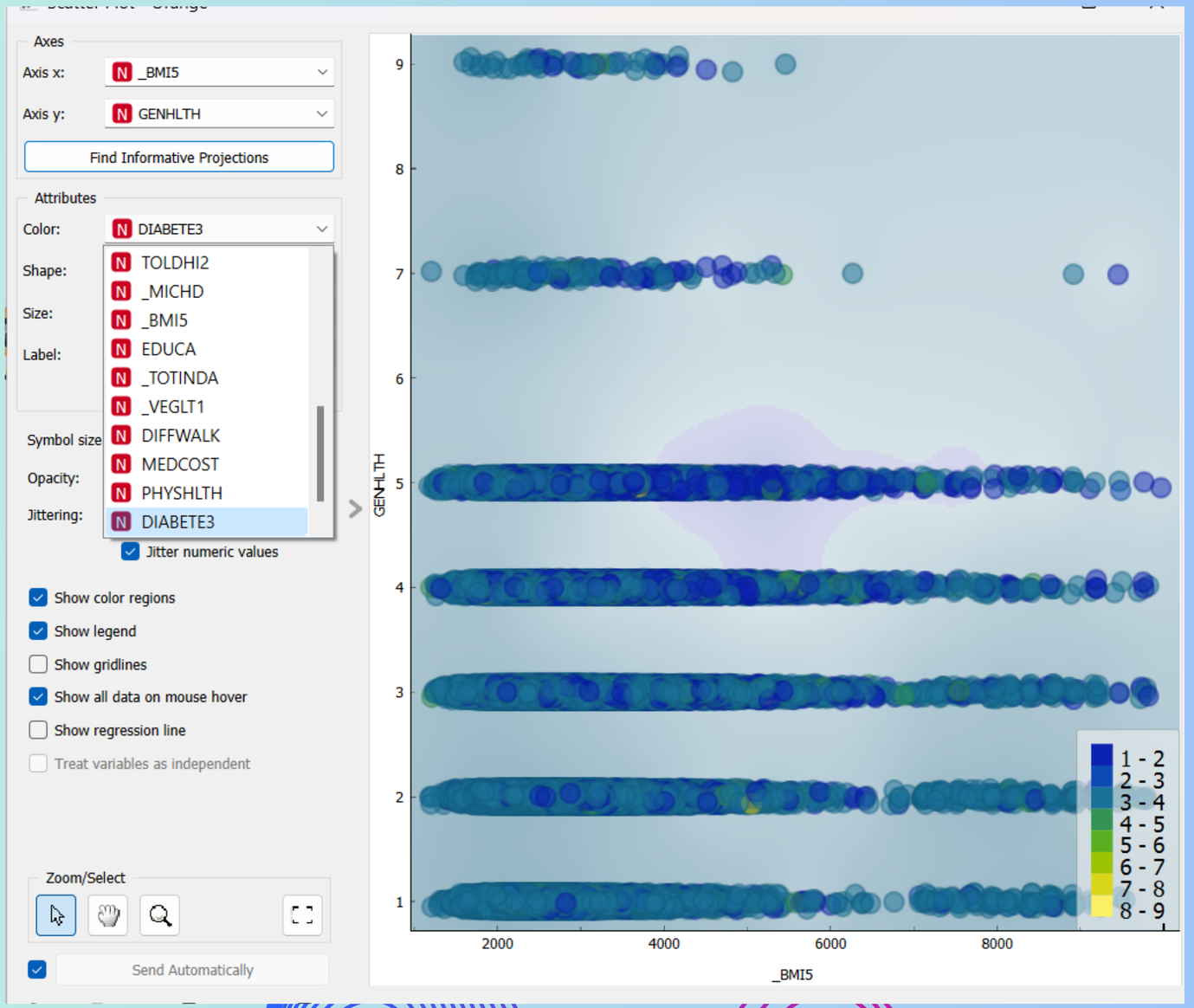
Box Plot:



Exploratory data analysis (EDA):

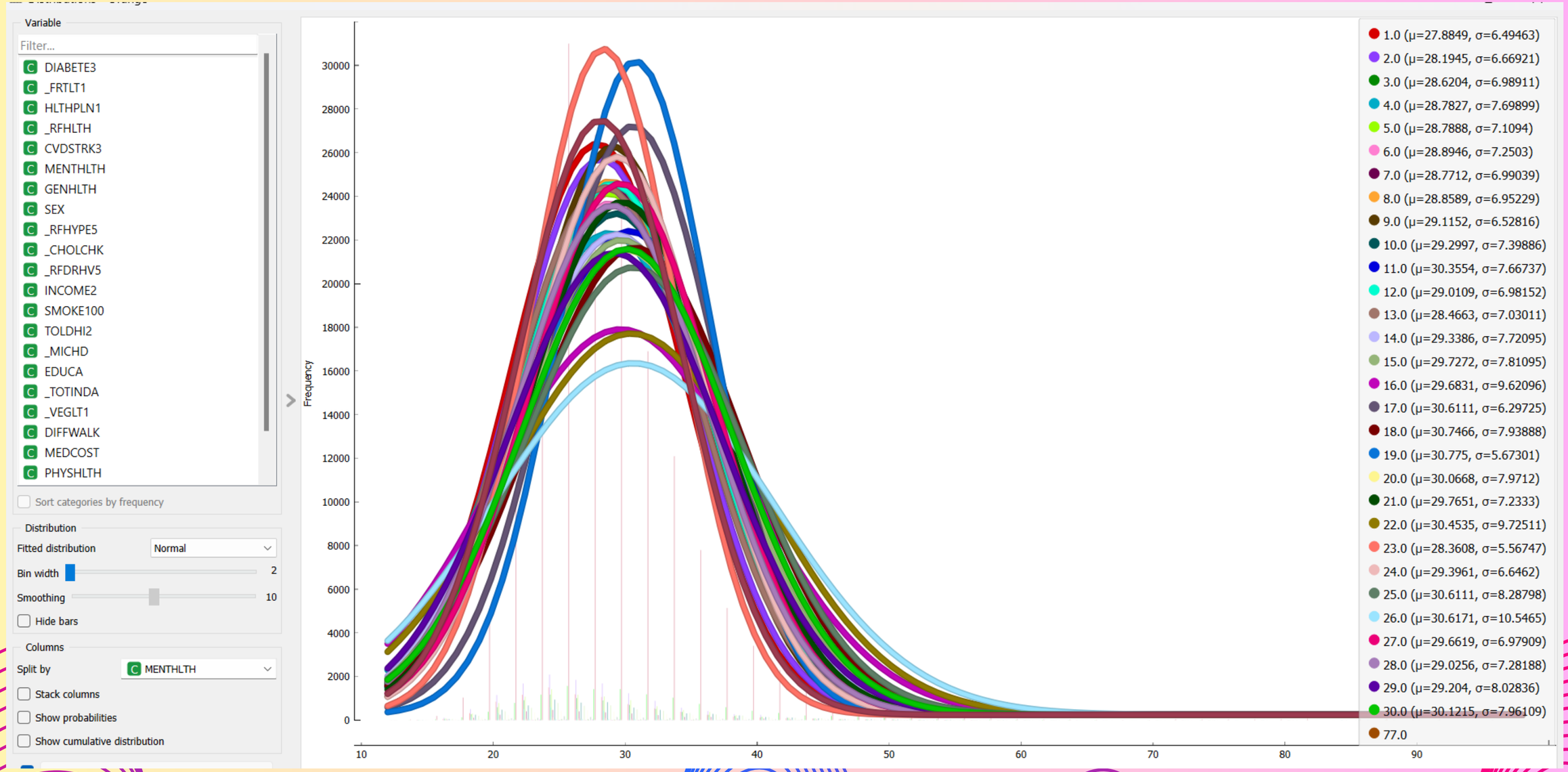
Back to Agenda

Scatter Plot:



Exploratory data analysis (EDA):

Back to Agenda



Exploratory data analysis (EDA):

Back to Agenda

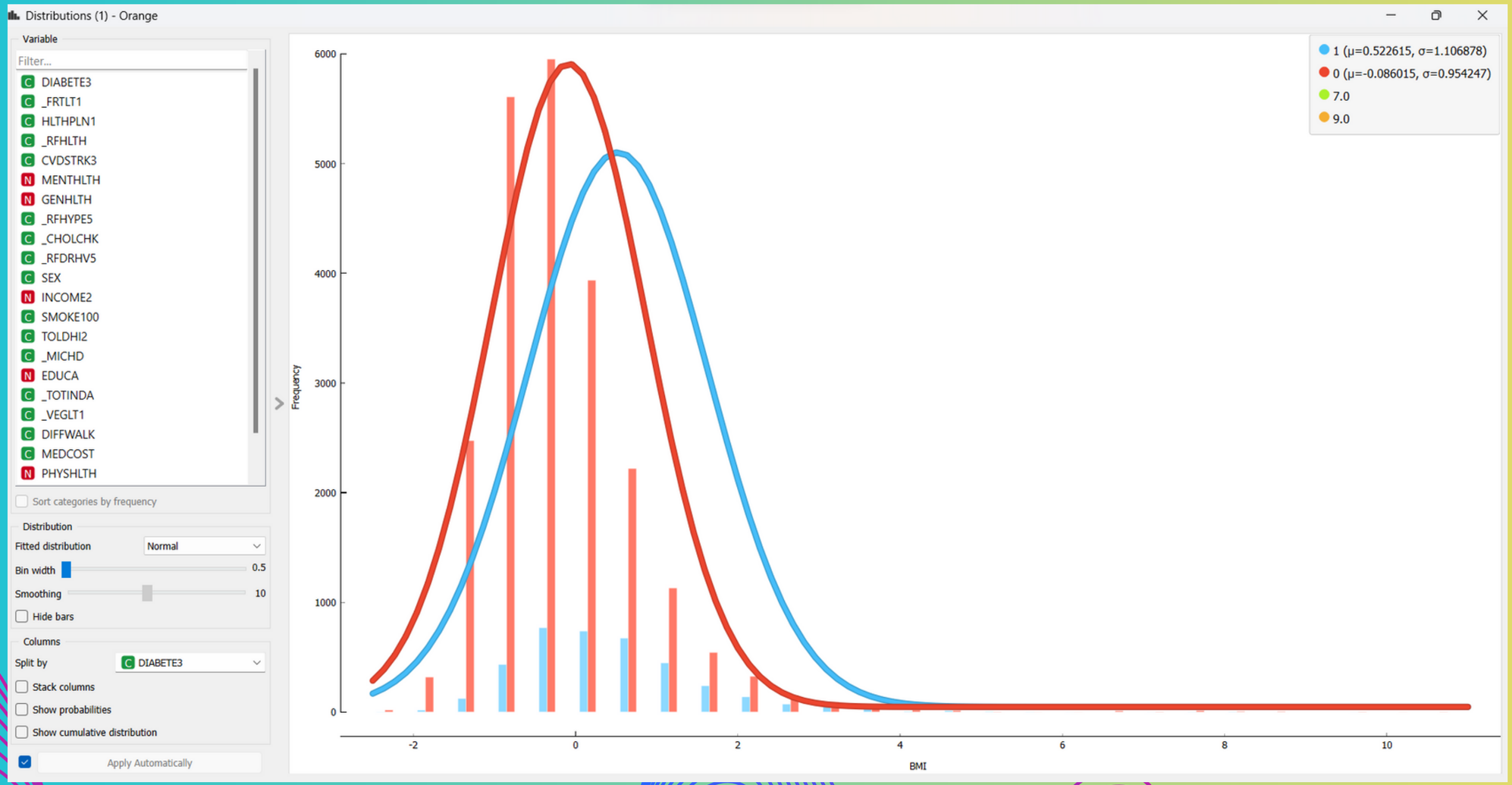
CONCLUSION:

- The brfss dataset is clearly imbalanced. When training my models, I get about 90% accuracy on many models with AUC between 70 and 80.
- And as in previous Figures , we see that none of features follow Normal Distribution, so, it is Imbalanced data.
- This may be caused by the models are learning the distribution in the data.

Exploratory data analysis (EDA):

Back to Agenda

Now ,we got the Normalized data



Exploratory data analysis (EDA):

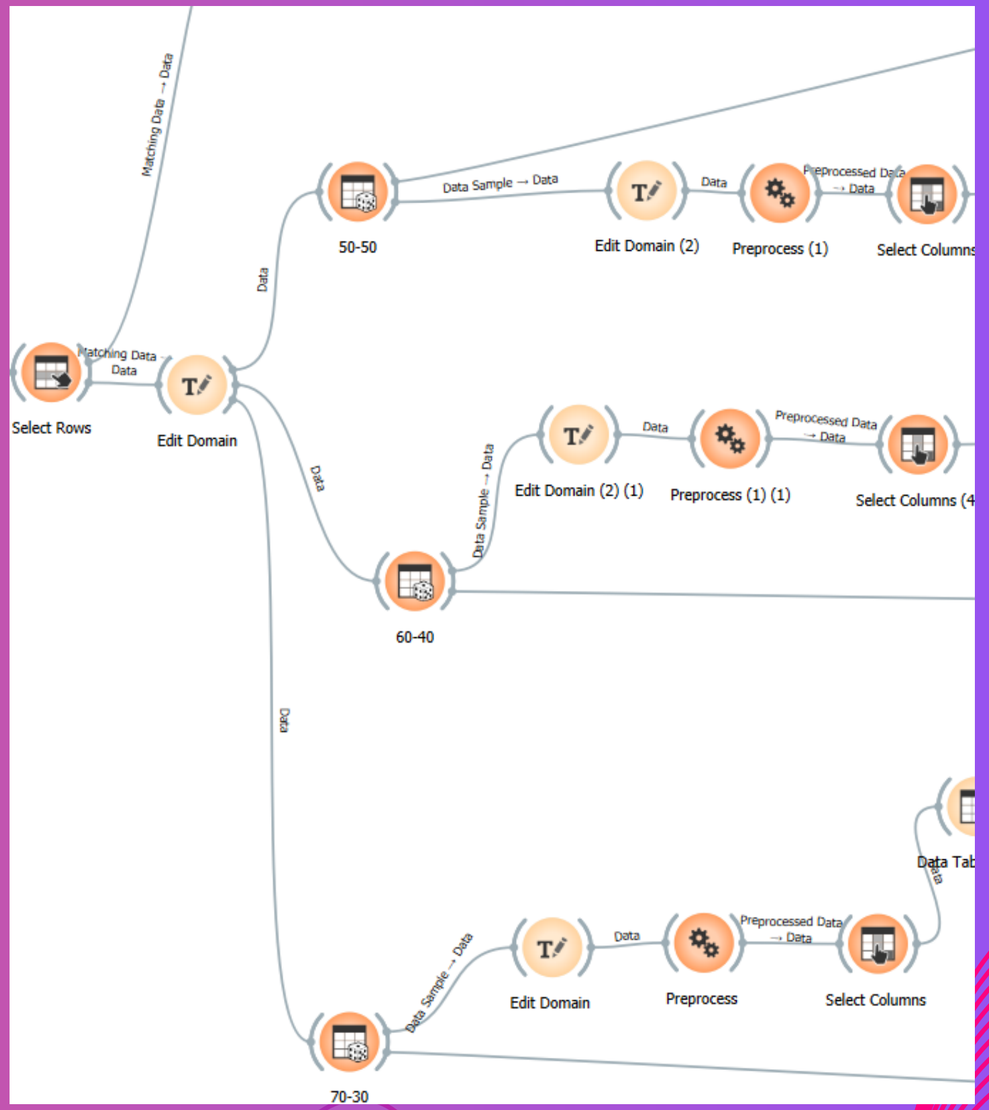
Back to Agenda

CONCLUSION:

- To handle the imbalance data, we have divide whole the data in three format 60-40,70-30,50-50 ratio ;
- To do this, I will take a random sample of 23,893 instances of the 0 (or No heart Disease / Attack) and all of the 23,893 instances of the 1 (or Yes Heart Disease / Attack).
- The if the new dataset performs comparably, then I can rest assured that it
- With roughly 48,000 datapoints, I hope that this is sufficient to train the model and that the random selection will not greatly change the results. I have the random seed set to 1.

Exploratory data analysis (EDA):

Its looks like this:

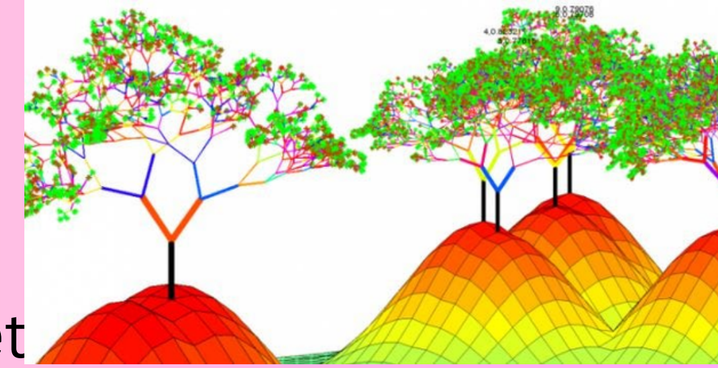


Model Building:

We are using Four Different Model:

- **Random Forest**
- **Logistic Regression**
- **Neural Network**
- **Support Vector Machine**
- **Decision Tree**

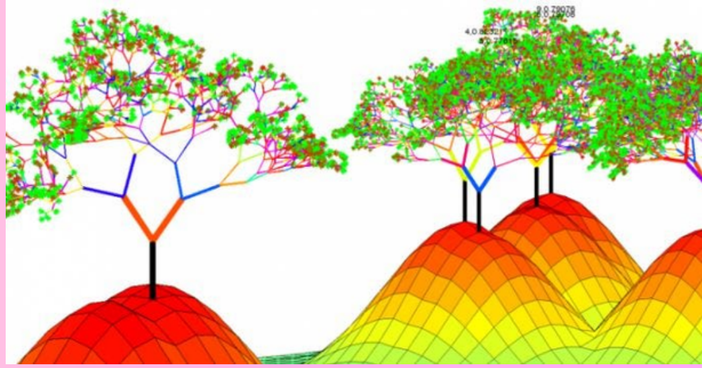
Random Forest:



Random Forest - w/ Feature Selection - Full Dataset
10 trees & 50 trees Tested (n_estimator changes)

- RF 10 trees - 5-fold cv - with Feature Selection : 0.89 (+/- 0.00) | AUC: 0.71 (+/- 0.01) | Runtime: 9.93 seconds
- RF 50 trees - 5-fold cv - with Feature Selection ACC: 0.89 (+/- 0.00) | AUC: 0.74 (+/- 0.01) | Runtime: 48.17 seconds
- RF 50 trees - 10-fold cv - with Feature Selection ACC: 0.89 (+/- 0.00) | AUC: 0.74 (+/- 0.01) | Runtime: 103.57 seconds
- RF Selected Features: ['DIABETE3', '_RFHYPE5', 'TOLDHI2', '_CHOLCHK', '_BMI5', 'SMOKE100', 'CVDSTRK3', '_MICHD', '_TOTINDA', '_FRTLT1', '_VEGLT1', '_RFDRHV5', 'HLTHPLN1', 'MEDCOST', 'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK', 'SEX', '_AGEG5YR', 'EDUCA', 'INCOME2']

Random Forest:



Cross validation

Number of folds: 5

Stratified

Cross validation by feature

Random sampling

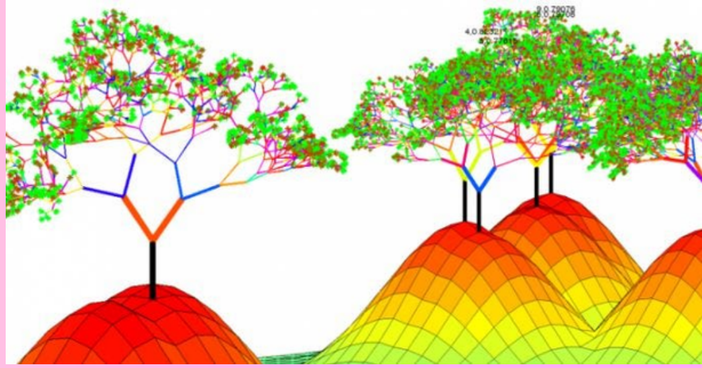
Repeat train/test: 10

Training set size: 66 %

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.757	0.853	0.818	0.810	0.853	0.176

Random Forest:



Plot

Target 1

Classifiers

Random Forest

Curves

Merge Predictions from Folds

Show convex ROC curves

Show ROC convex hull

Analysis

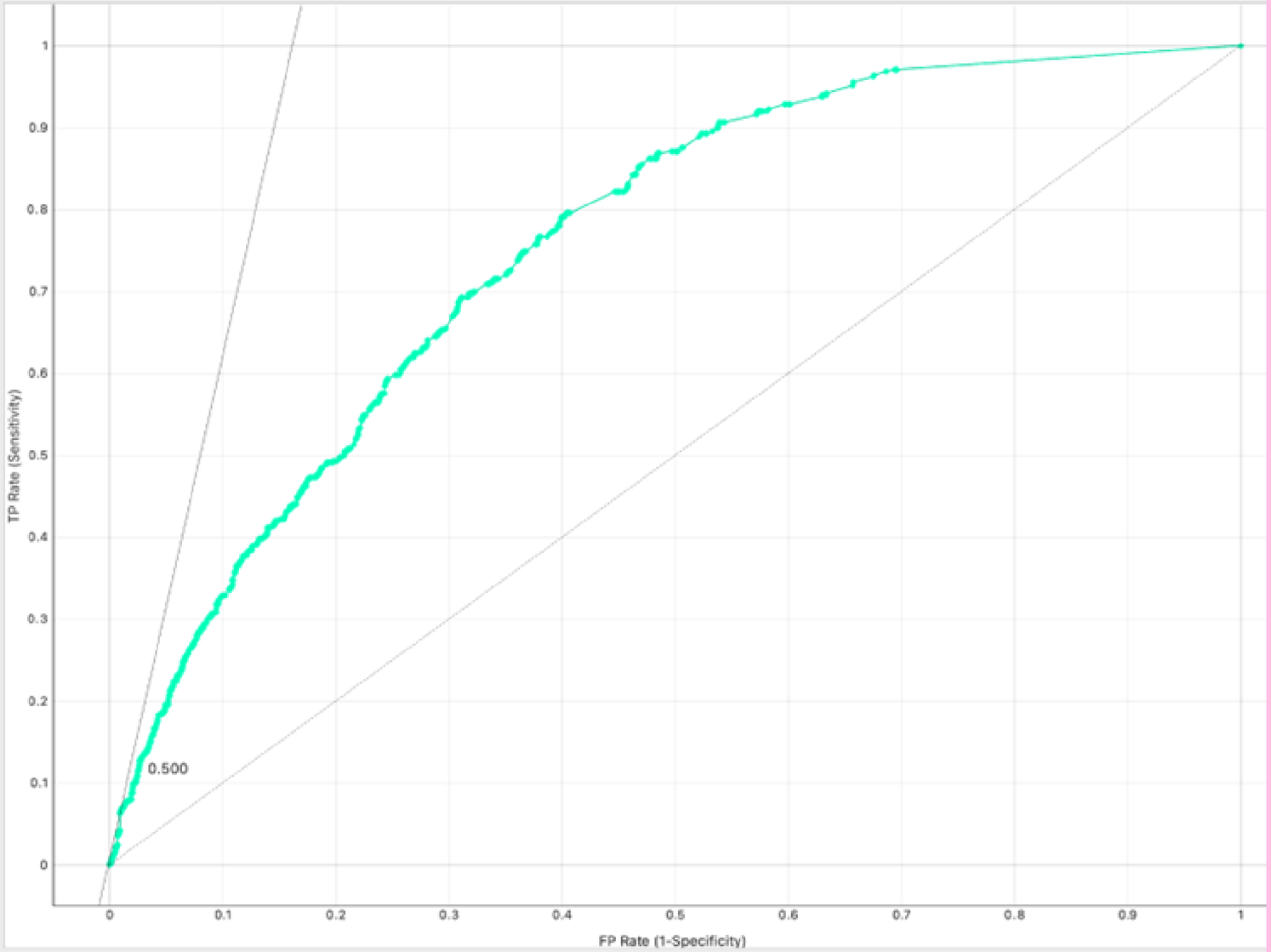
Default threshold (0.5) point

Show performance line

FP Cost: 500

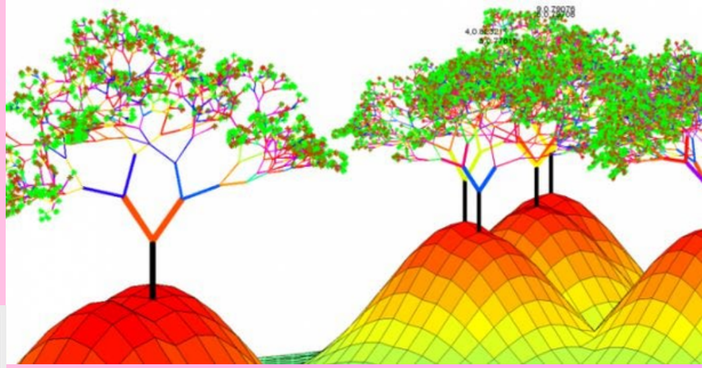
FN Cost: 500

Prior probability: 14 %



1×3200

Random Forest:



Curve

Target: 1

- Lift Curve
- Cumulative Gains
- Precision Recall

Models

- Random Forest

Settings

- Show thresholds
- Show points

Area under the curve

- Random Forest: 0.755

Apply Automatically

Probability threshold(s):
0.073

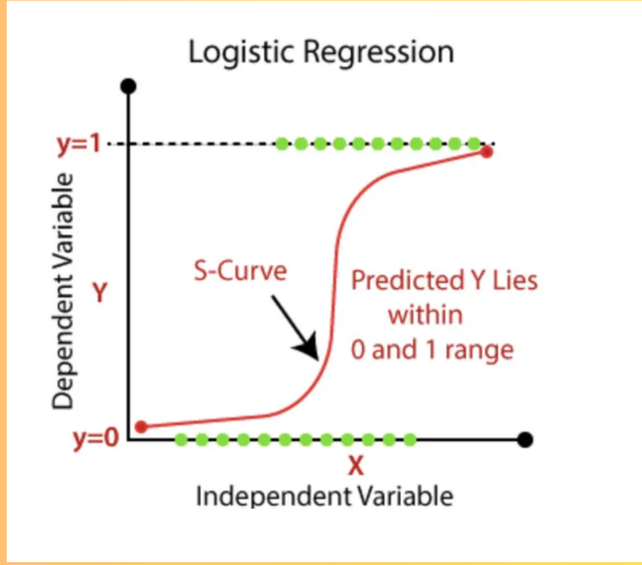
Precision

Recall

1×3200

Can't output a model:

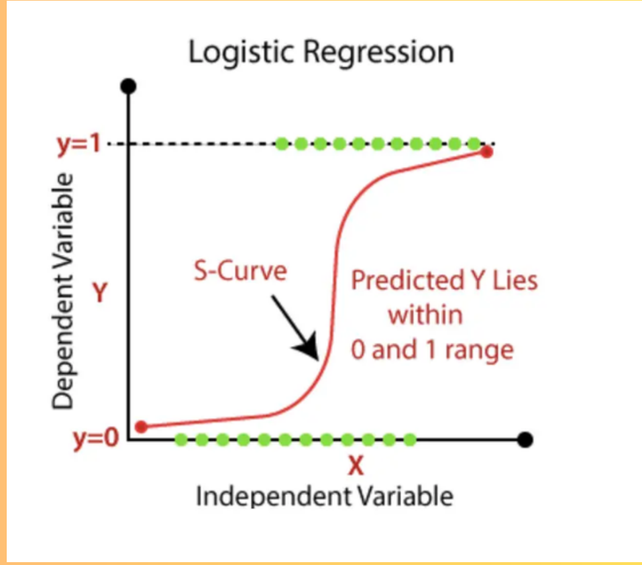
Logistic Regression:



Logistic Regression - w/ Feature Selection - Full Dataset

- Logistic Regression with Feature Selection : 0.809 (+/- 0.00) | AUC: 0.803 (+/- 0.01) |
- Logistic Regression with Feature Selection : 0.823 (+/- 0.00) | F1 score: 0.829 (+/- 0.01) |
- RF Selected Features: ['DIABETE3', '_RFHYPE5', 'TOLDHI2', '_CHOLCHK', '_BMI5', 'SMOKE100', 'CVDSTRK3', '_MICHD', '_TOTINDA', '_FRTLT1', '_VEGLT1', '_RFDRHV5', 'HLTHPLN1', 'MEDCOST', 'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK', 'SEX', '_AGEG5YR', 'EDUCA', 'INCOME2']

Logistic Regression:



Cross validation

Number of folds: 5

Stratified

Cross validation by feature

Random sampling

Repeat train/test: 10

Training set size: 66 %

Stratified

Leave one out

Test on train data

Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.803	0.861	0.823	0.824	0.861	0.208

Compare models by: Area under ROC curve

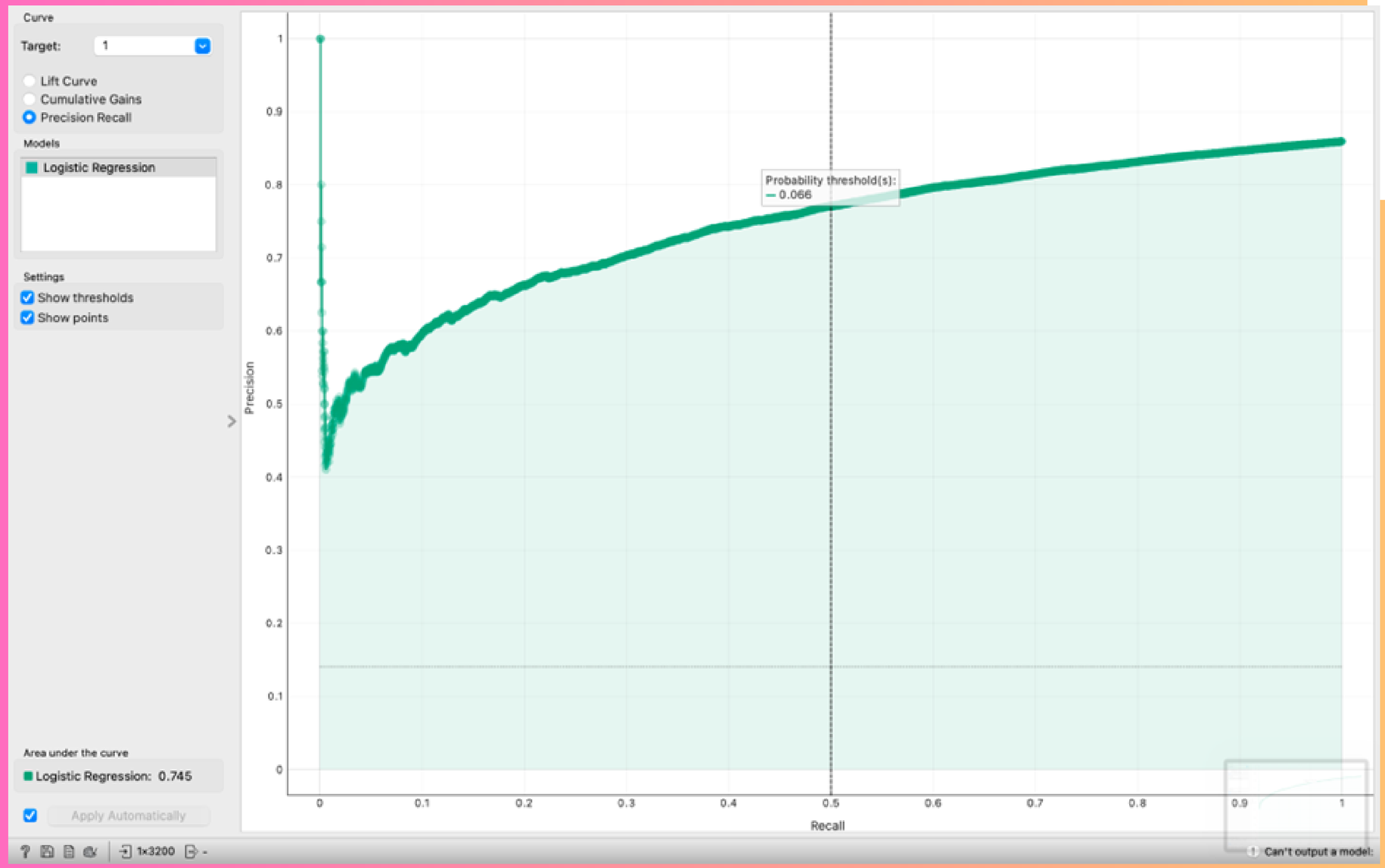
Negligible diff.: 0.1

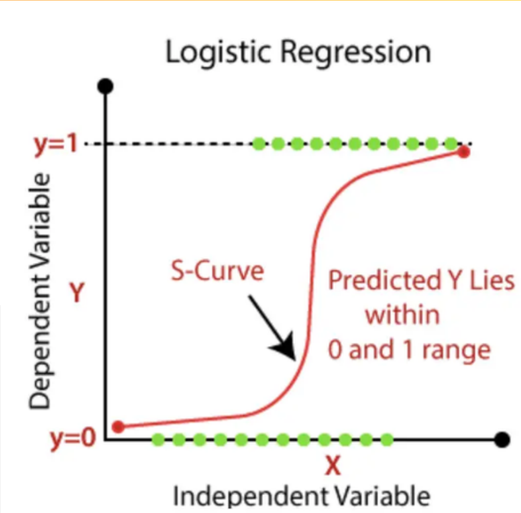
	Logistic Regression
Logistic Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

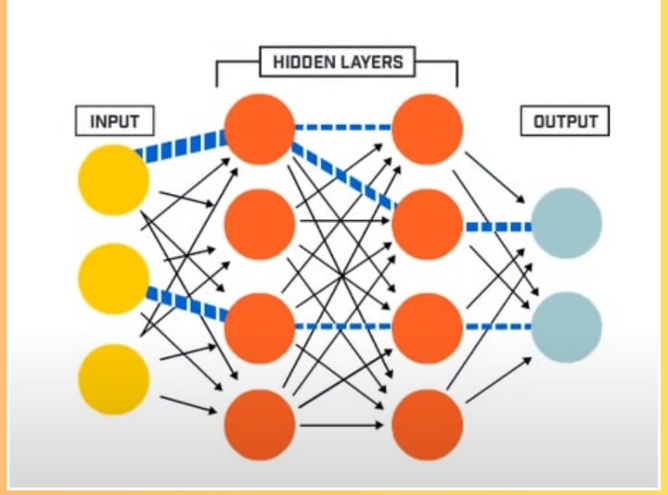
3200 | - | 0 | - 3200 | 1×3200

Logistic Regression:





Neural Network:

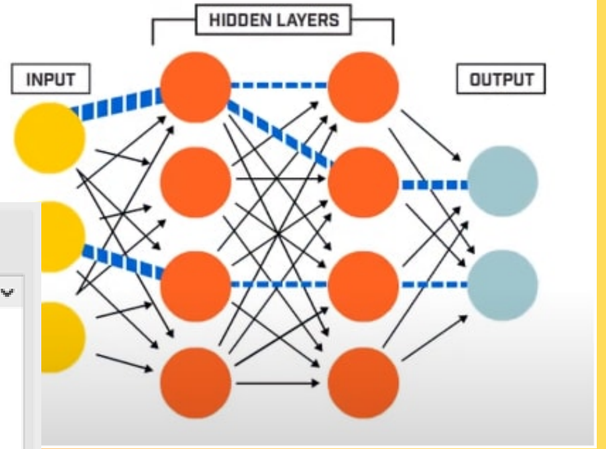


- Neural Network - Acc: 0.76 (+/- 0.01)
- Neural Network - AUC: 0.74 (+/- 0.01)

Notes:

- Selected Features:same as Logistic Regression
- w/o feature selection was also tested, but there was no significant change in ACC or AUC.
-

Neural Network:



Cross validation

Number of folds: 5

Stratified

Cross validation by feature

Random sampling

Repeat train/test: 10

Training set size: 66 %

Stratified

Leave one out

Test on train data

Test on test data

Evaluation results for target 1

Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network	0.745	0.829	0.246	0.324	0.198	0.161

Compare models by: Area under ROC curve

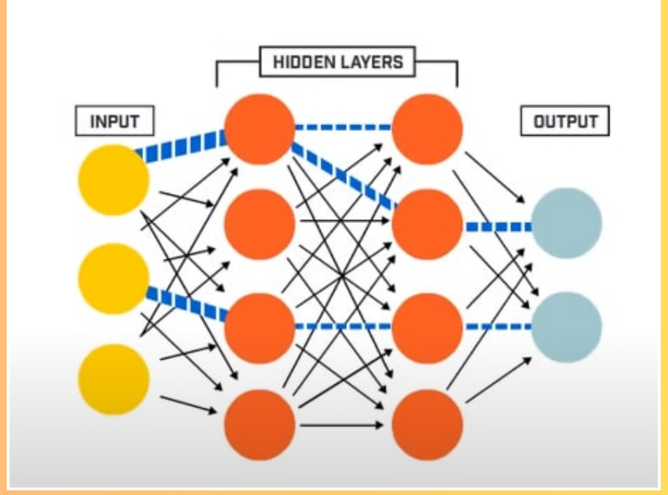
Negligible diff.: 0.1

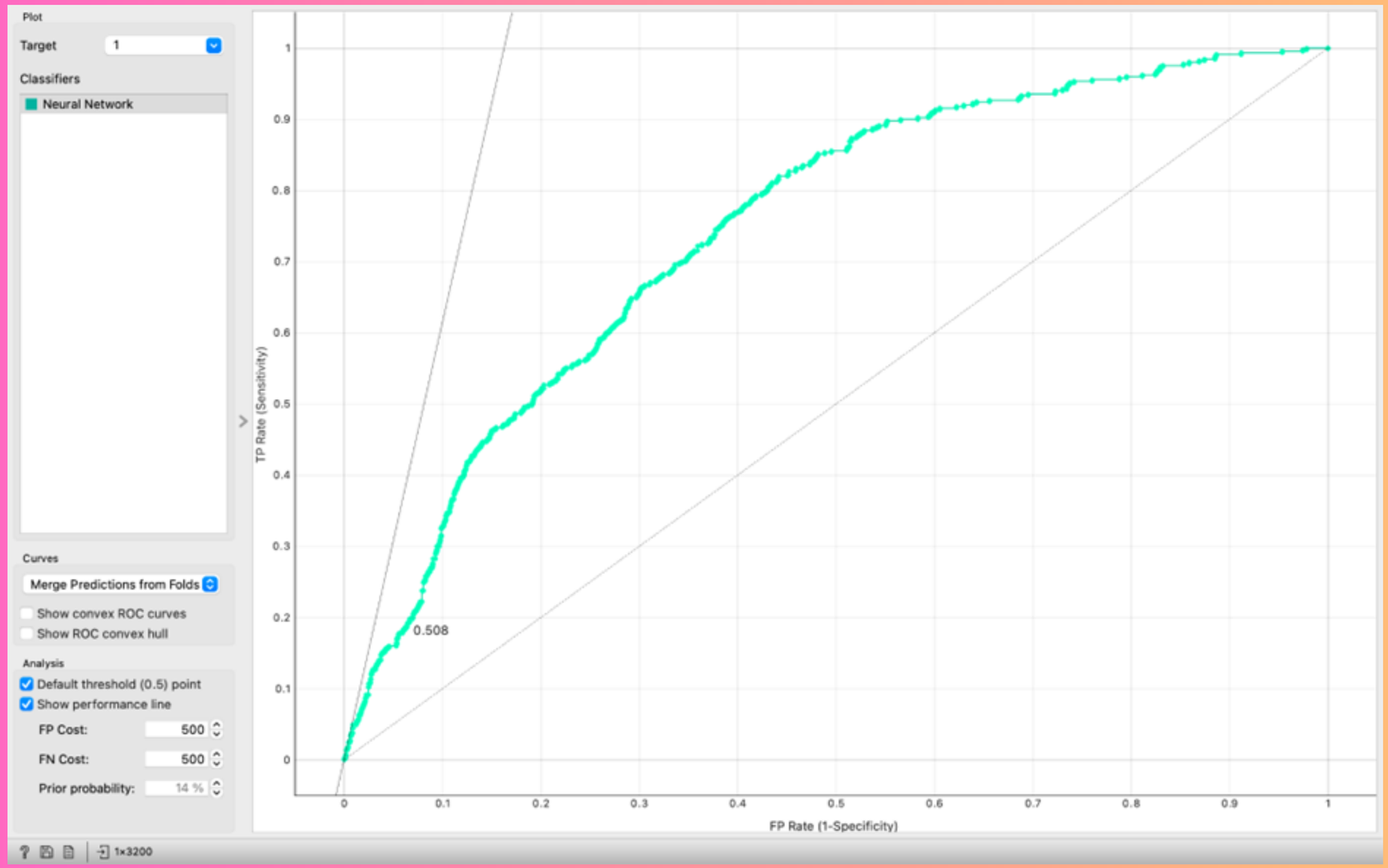
	Neural Network
Neural Network	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column.Small numbers show the probability that the difference is negligible.

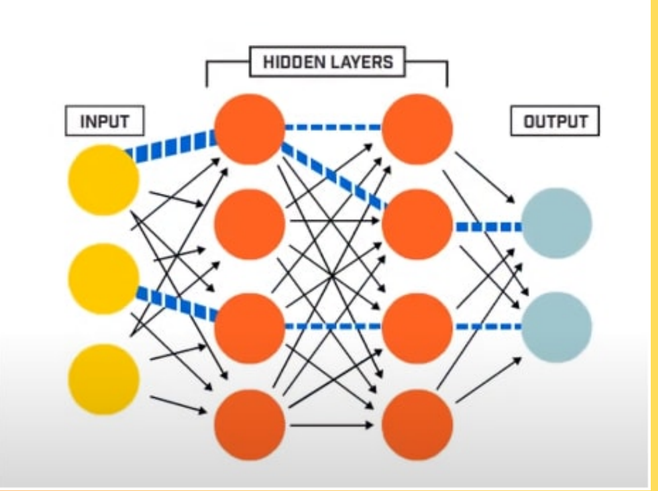
3200 | - | 0 | - 3200 | 1×3200

Neural Network:





Neural Network:



Curve

Target: 1

- Lift Curve
- Cumulative Gains
- Precision Recall

Models

- Neural Network

Settings

- Show thresholds
- Show points

Area under the curve

Neural Network: 0.767

Apply Automatically

Probability threshold(s):
0.025

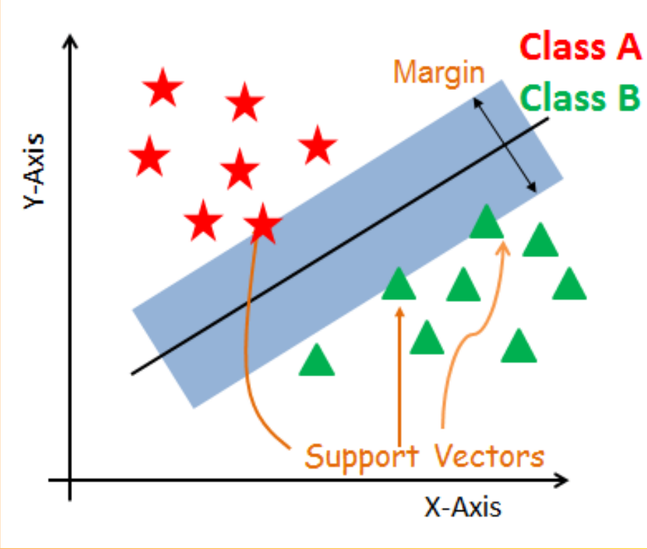
Precision

Recall

1x3200

Can't output a model:

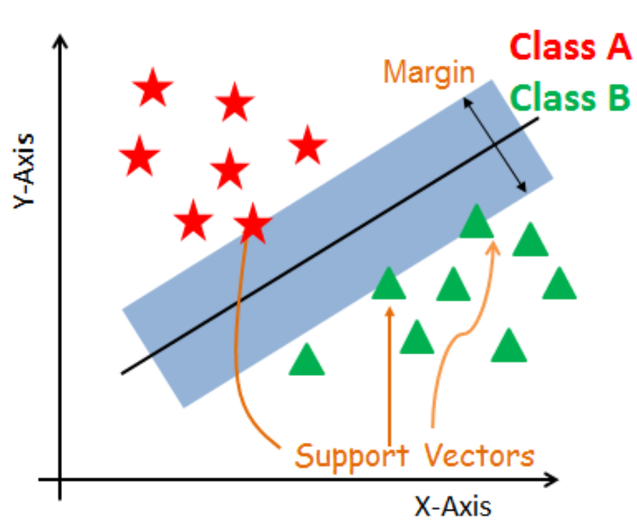
Support Vector Machine



SVM - w/ Feature Selection - Full Dataset

- Logistic Regression with Feature Selection : 0.706 (+/- 0.00) | AUC: 0.709 (+/- 0.01) |
- Logistic Regression with Feature Selection : 0.267 (+/- 0.00) | F1 score: 0.268 (+/- 0.01) |
- RF Selected Features: ['DIABETE3', '_RFHYPE5', 'TOLDHI2', '_CHOLCHK', '_BMI5', 'SMOKE100', 'CVDSTRK3', '_MICHD', '_TOTINDA', '_FRTLT1', '_VEGLT1', '_RFDRHV5', 'HLTHPLN1', 'MEDCOST', 'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK', 'SEX', '_AGEG5YR', 'EDUCA', 'INCOME2']

Support Vector Machine



Cross validation

Number of folds: 5

Stratified

Cross validation by feature

Random sampling

Repeat train/test: 20

Training set size: 66 %

Stratified

Leave one out

Test on train data

Test on test data

Evaluation results for target 1

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.706	0.801	0.267	0.276	0.258	0.152

Compare models by: Area under ROC curve

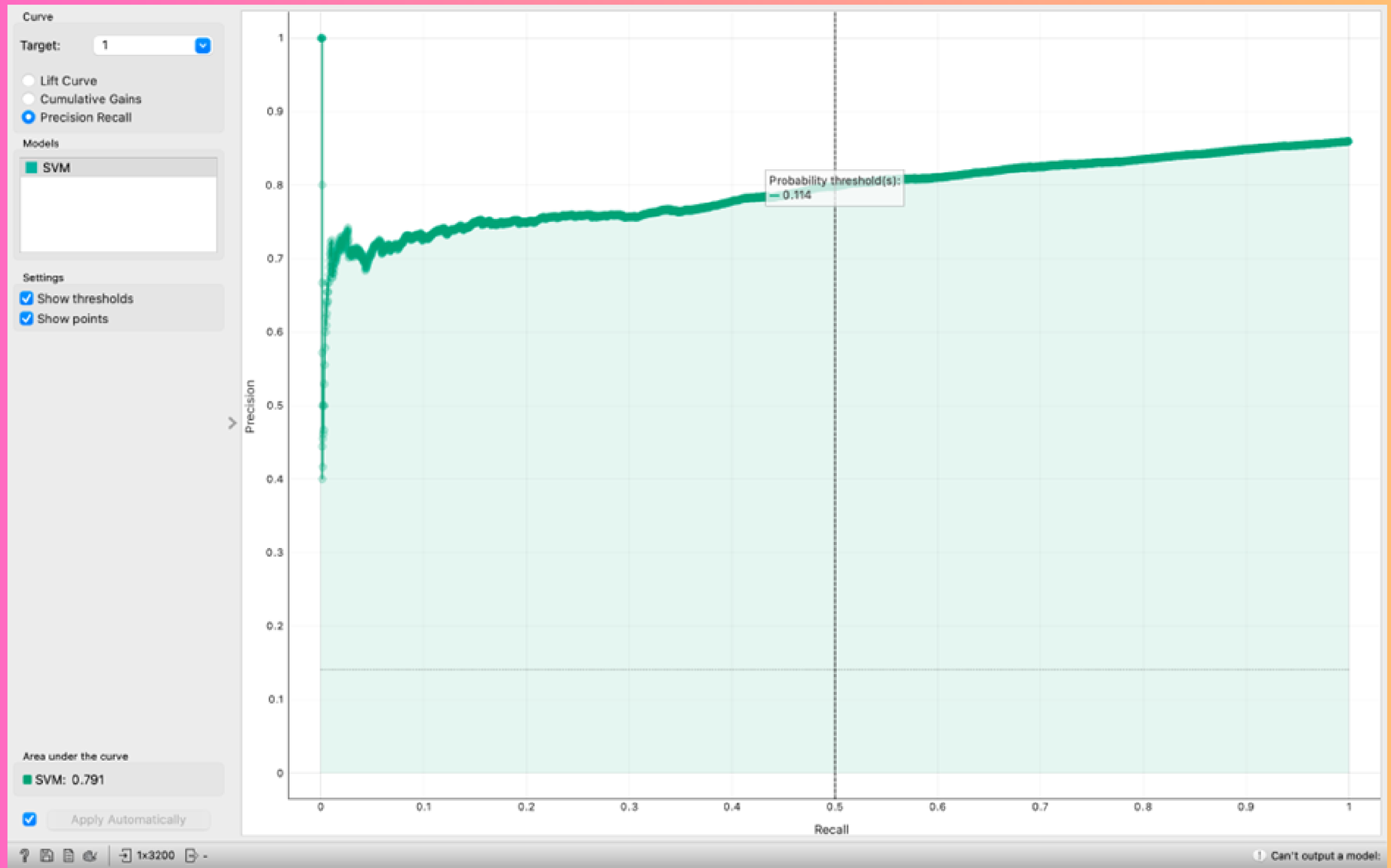
Negligible diff.: 0.1

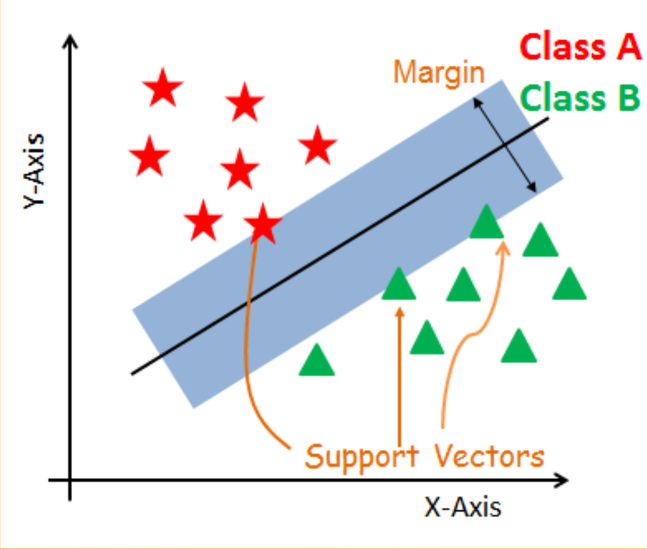
	SVM
SVM	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column.Small numbers show the probability that the difference is negligible.

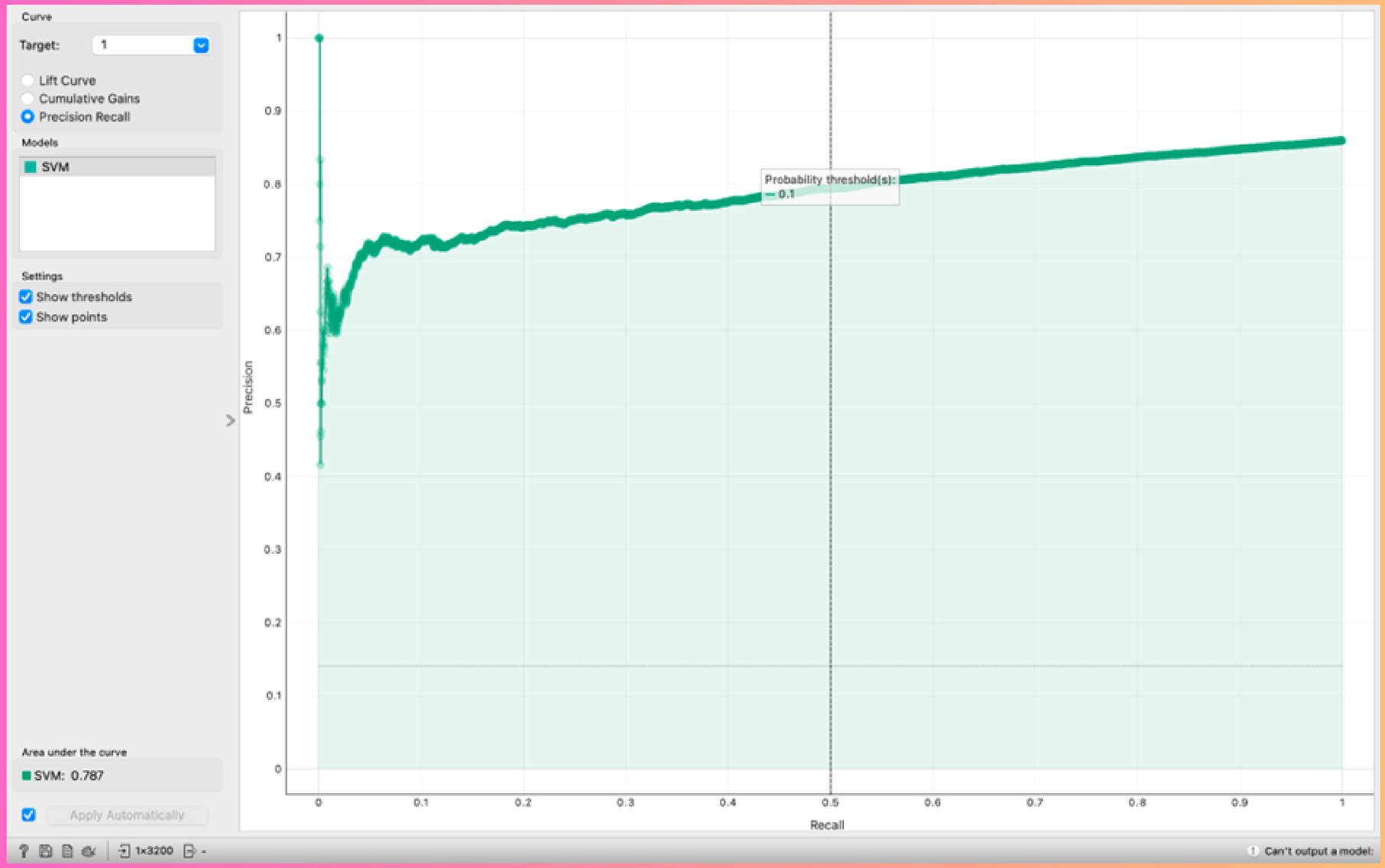
3200 | - | 0 | - 3200 | 1×3200

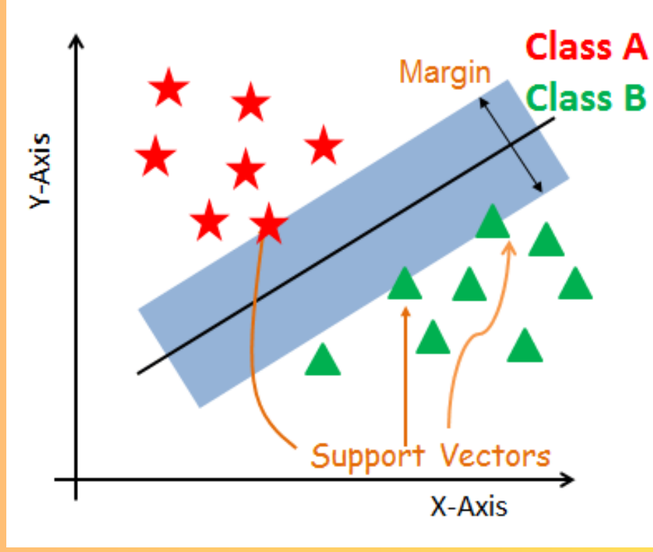
Support Vector Machine



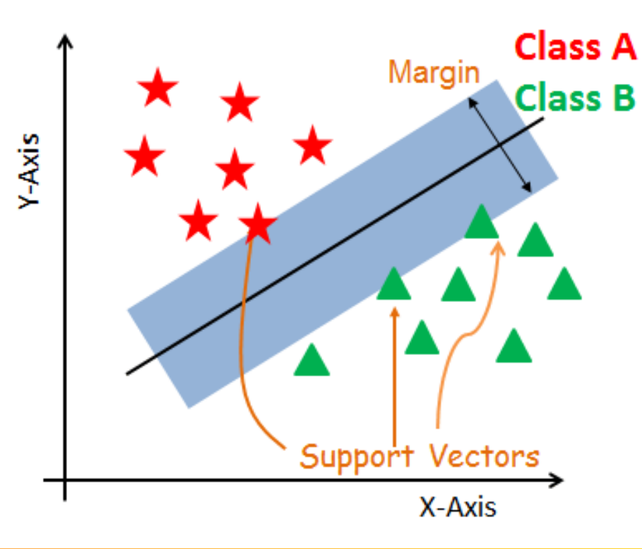


Support Vector Machine





Prediction(50-50)



Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
SVM	-2.751	0.630	0.686	0.802	0.630
Tree	-3.470	0.827	0.820	0.814	0.827

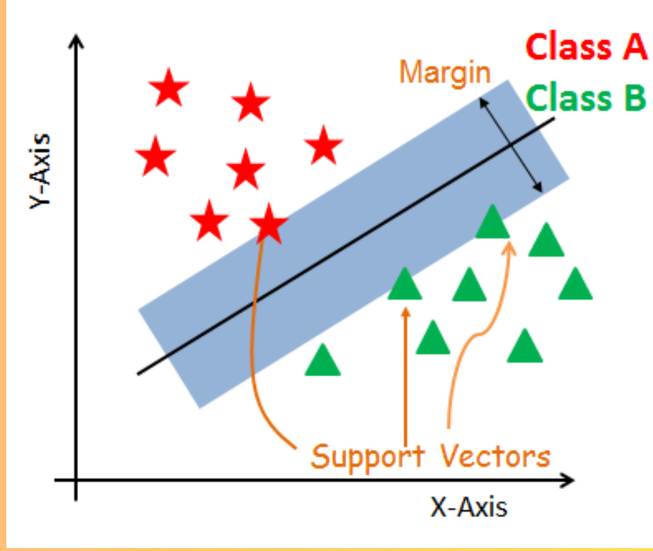
Test and Score (1) - Orange

Cross validation
Number of folds: 10
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 20
Training set size: 66 %
Stratified

Evaluation results for target 1

Model	AUC	CA	F1	Precision	Recall
Tree	0.929	0.971	0.884	0.998	0.794
SVM	0.371	0.631	0.294	0.201	0.549

Prediction(60-40)



Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
SVM	-0.754	0.213	0.223	0.656	0.213
Tree (1)	-0.596	0.828	0.821	0.816	0.828

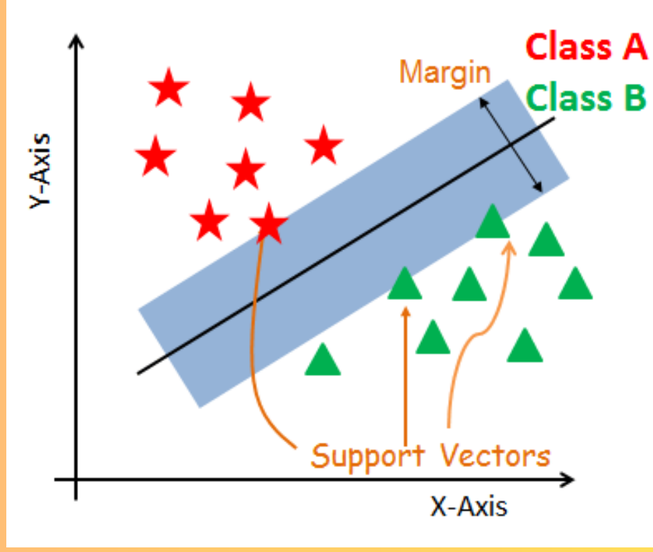
Test and Score - Orange

Cross validation
Number of folds: 10
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 20
Training set size: 66 %
Stratified

Evaluation results for target 1

Model	AUC	CA	F1	Precision	Recall
Tree (1)	0.930	0.972	0.887	0.999	0.798
SVM	0.598	0.352	0.255	0.151	0.802

Prediction(70-30)



Show perfomance scores Target class: (Average o

Model	AUC	CA	F1	Precision	Recall
SVM (1)	-1.571	0.215	0.225	0.655	0.215
Tree (1) (1)	-2.168	0.935	0.932	0.931	0.935

Test and Score (1) - Orange

Cross validation

Number of folds: 10

Stratified

Cross validation by feature

Random sampling

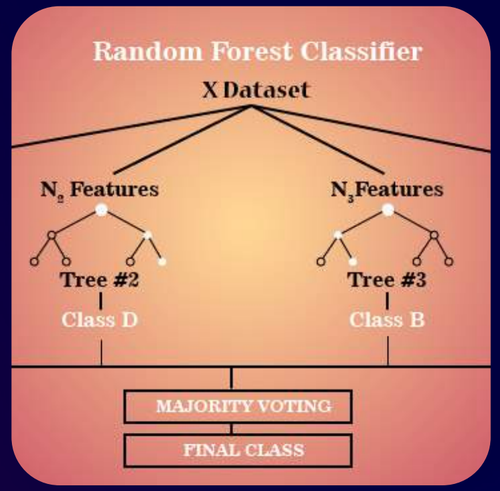
Repeat train/test: 20

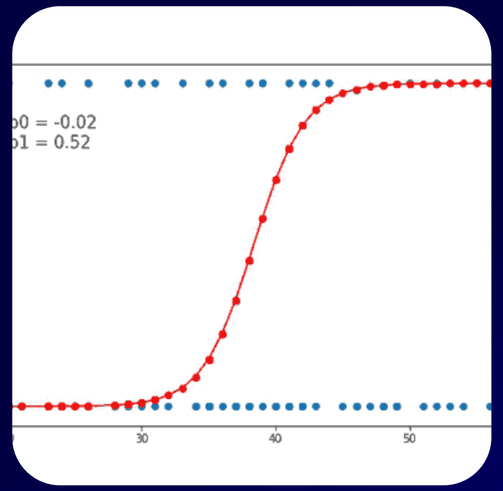
Evaluation results for target 1

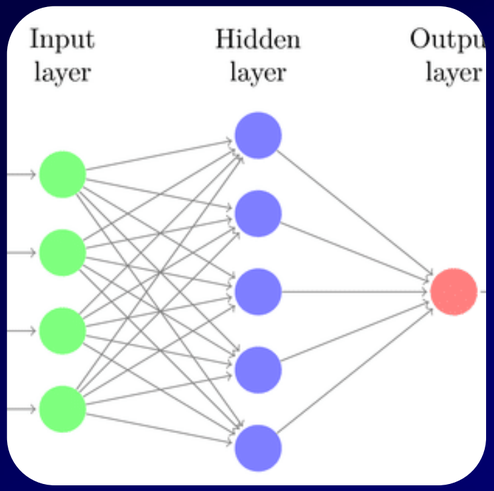
Model	AUC	CA	F1	Precision	Recall
Tree (1) (1)	0.930	0.972	0.887	0.999	0.798
SVM (1)	0.402	0.352	0.255	0.151	0.802

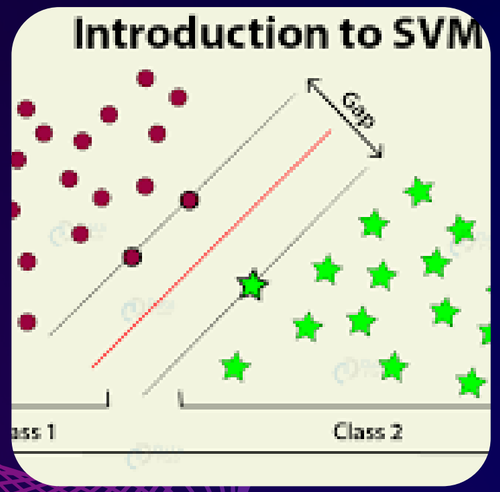
Final Model

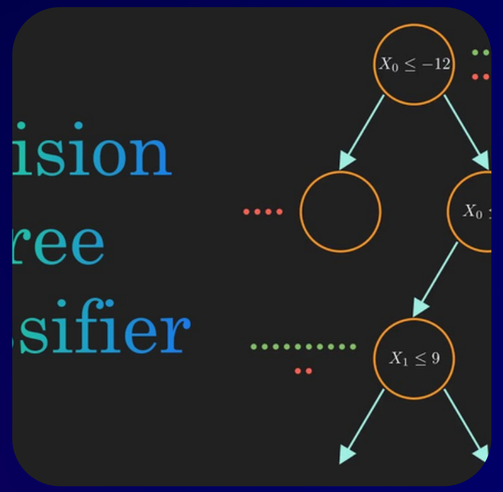
As we can see that Random Forest and Logestic regression having good accuracy , so we can choose any one of them for prediction.











Thank you!