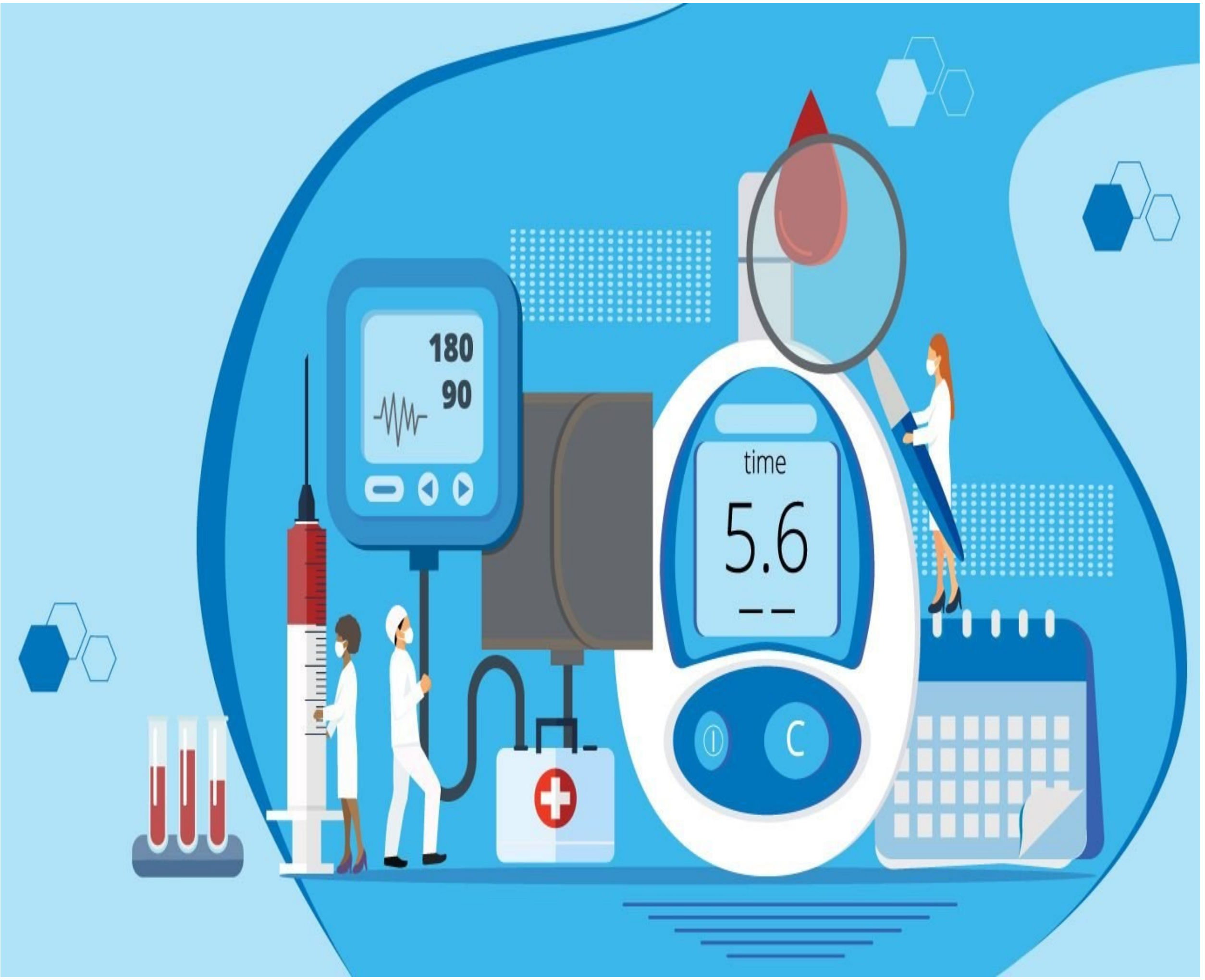


DIABETES PREDICTION MODEL USING WEKA TOOL

Presented by S. A. L. M. Mihiliya Jayasiri

Content

1. Purpose of Creating the Model
2. Dataset Description
3. Data Preprocessing
4. Choosing the Algorithm
5. Building and Evaluating the Model
6. Results

1.Purpose of Creating the Model

Diabetes is a chronic condition that affects how body processes blood sugar(glucose). By the use of data mining tools like WEKA developing models to predict diabetes using various health related attributes is helpful in early detection and management of diabetes, which is crucial for preventing complications.

2. Dataset Description

- The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- It includes medical data from 768 female patients of Pima Indian Heritage, who are at least 21 years old.
- The dependent variable is a binary indicator of diabetes, where ‘0’ indicates no diabetes and ‘1’ indicates diabetes.

2. Dataset Description (Continued)

The dataset contains 8 medical predictor variables: Pregnancies

Glucose

Blood pressure

Skin thickness

Insulin, BMI

Diabetes Pedigree Function

Age

	Normal	Prediabetes	Diabetes
Fasting Blood sugar level (FBS)	< 1 g/l	1 g/l ≤ BS ≤ 1.25 g/l	BS ≥ 1.26 g/l
Glycated haemoglobin (HbA1C)	< 5.7%	5.7% ≤ A1C ≤ 6.4%	A1C ≥ 6.5 %
Oral glucose Tolerance (OGTT)	< 1.4g/l	1.4g/l ≤ OGTT ≤ 1.99g/l	OGTT ≥ 2 g/l

3. Data Preprocessing

Data preprocessing takes place in the **preprocess** section of **Explorer** in the Weka tool.

a. Loading the Dataset

b. Scaling and Categorising the Data

c. Splitting the Dataset

3. Data Preprocessing (Continued)

b. Scaling and Categorising the Data

- Important to improve the performance of Machine Learning Algorithms.
- Helps in reducing the impact of outliers.
- Improves the accuracy of the model.

E.g. Normalize, Standardize, Discretize, etc…

b. Scaling and Categorising the Data (Continued)

The process of converting continuous numeric attributes into discrete nominal attributes or bins is called **Discretization**.

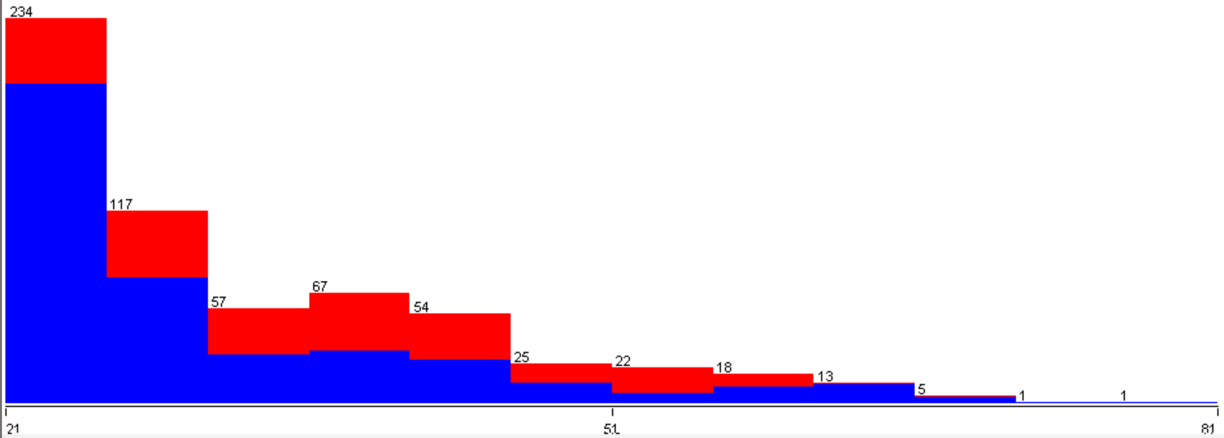
Algorithms: Naïve Bayes, Random Forest, J48

Selected attribute

Name: age | Type: Numeric
Missing: 0 (0%) | Distinct: 51 | Unique: 4 (1%)

Statistic	Value
Minimum	21
Maximum	81
Mean	33.357
StdDev	11.792

Class: class (Nom) | Visualize All



Before

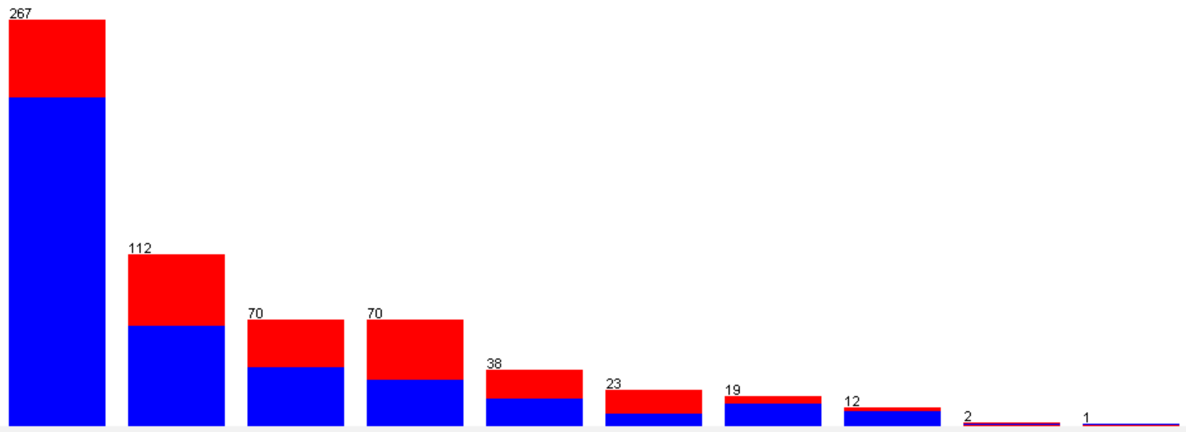
After

Selected attribute

Name: age | Type: Nominal
Missing: 0 (0%) | Distinct: 10 | Unique: 1 (0%)

No.	Label	Count	Weight
1	'(-inf-27]'	267	267
2	'(27-33]'	112	112
3	'(33-39]'	70	70
4	'(39-45]'	70	70
5	'(45-51]'	38	38
6	'(51-57]'	23	23
7	'(57-63]'	19	19
8	'(63-69]'	12	12
9	'(69-75]'	2	2
10	'(75-inf)'	1	1

Class: class (Nom) | Visualize All



b. Scaling and Categorising the Data (Continued)

The process of transforming the data so that it has a mean of 0 and a standard deviation of 1 is called **Standardization**.

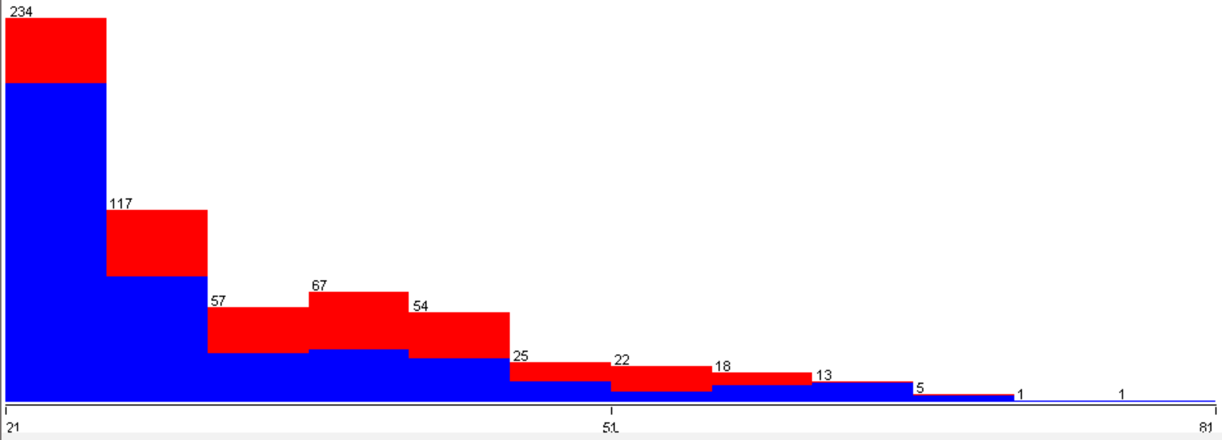
Algorithms: SMO, LMT

Selected attribute

Name: age | Type: Numeric
Missing: 0 (0%) | Distinct: 51 | Unique: 4 (1%)

Statistic	Value
Minimum	21
Maximum	81
Mean	33.357
StdDev	11.792

Class: class (Nom) | Visualize All



Before

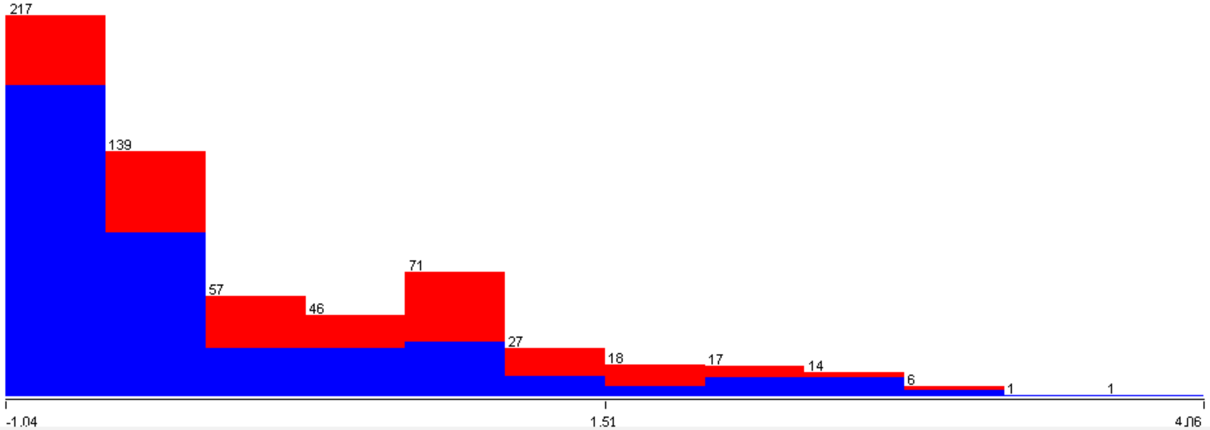
After

Selected attribute

Name: age | Type: Numeric
Missing: 0 (0%) | Distinct: 52 | Unique: 7 (1%)

Statistic	Value
Minimum	-1.041
Maximum	4.061
Mean	-0.005
StdDev	1.005

Class: class (Nom) | Visualize All



3. Data Preprocessing (Continued)

Algorithm	Evaluated test dataset of 154 instances (Accuracy)							
	No Technique		Discretize		Normalize		Standardize	
Naïve Bayes	114	74.03%	117	75.97%	114	74.03%	116	75.32%
SMO	115	74.68%	111	72.08%	115	74.68%	118	76.62%
Random Forest	112	72.73%	119	77.27%	113	73.38%	116	75.32%
LMT	114	74.03%	114	74.03%	114	74.03%	119	77.27%
J48	106	68.83%	108	70.13%	106	68.83%	102	66.23%

3. Data Preprocessing (Continued)

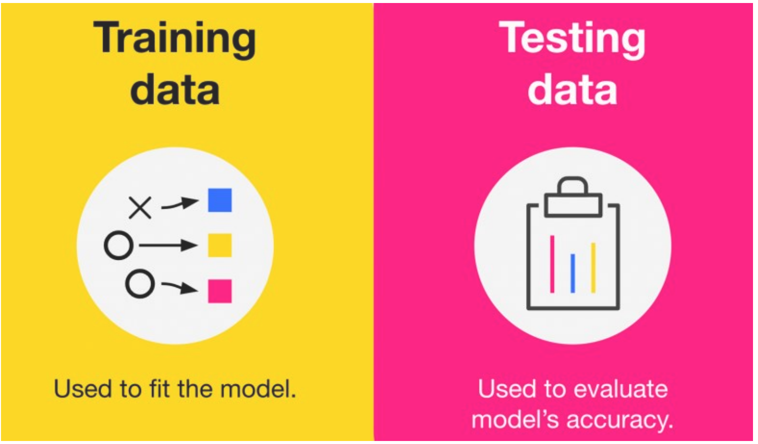
c. Splitting the Dataset

In a prediction model two datasets are used; 1. Training dataset

2. Testing dataset

Training dataset – 80% of the full dataset

Testing dataset – remaining 20% of the full dataset



4. Choosing the Algorithm

From 5 different types of algorithms; Naïve Bayes, SMO, Random Forest, LMT, J48.

Algorithm	Evaluated test dataset of 154 instances (Accuracy)	
Naïve Bayes	117	74.97%
SMO	118	76.62%
Random Forest	119	77.27%
LMT	119	77.27%
J48	108	70.13%

4. Choosing the Algorithm (Continued)

Algorithm	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Relative Absolute Error (RAM)	Root Relative Squared Error (RRSE)
Random Forest	0.3164	0.4082	69.55%	85.54%
LMT	0.314	0.4027	69.02%	84.39%

4. Choosing the Algorithm (Continued)

Lower MAE, RMSE values and lower RAE, RRSE percentages suggest that the model’s predictions are more accurate and model performance is better.

Therefore, the chosen algorithm is the <u>LMT algorithm</u>.

5. Building and Evaluating the Model

a. Creating the Model

By classifying the training dataset on the LMT algorithm the Model was created.

5. Building and Evaluating the Model

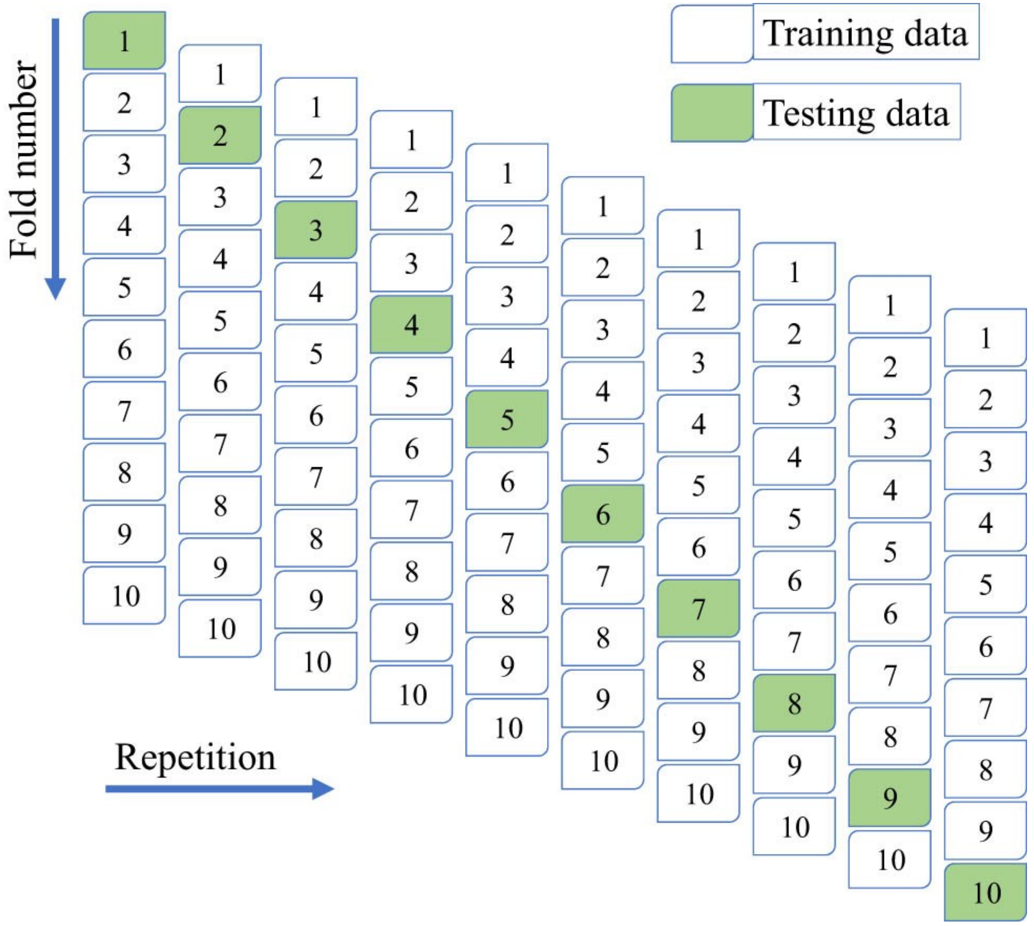
b. Creation Process

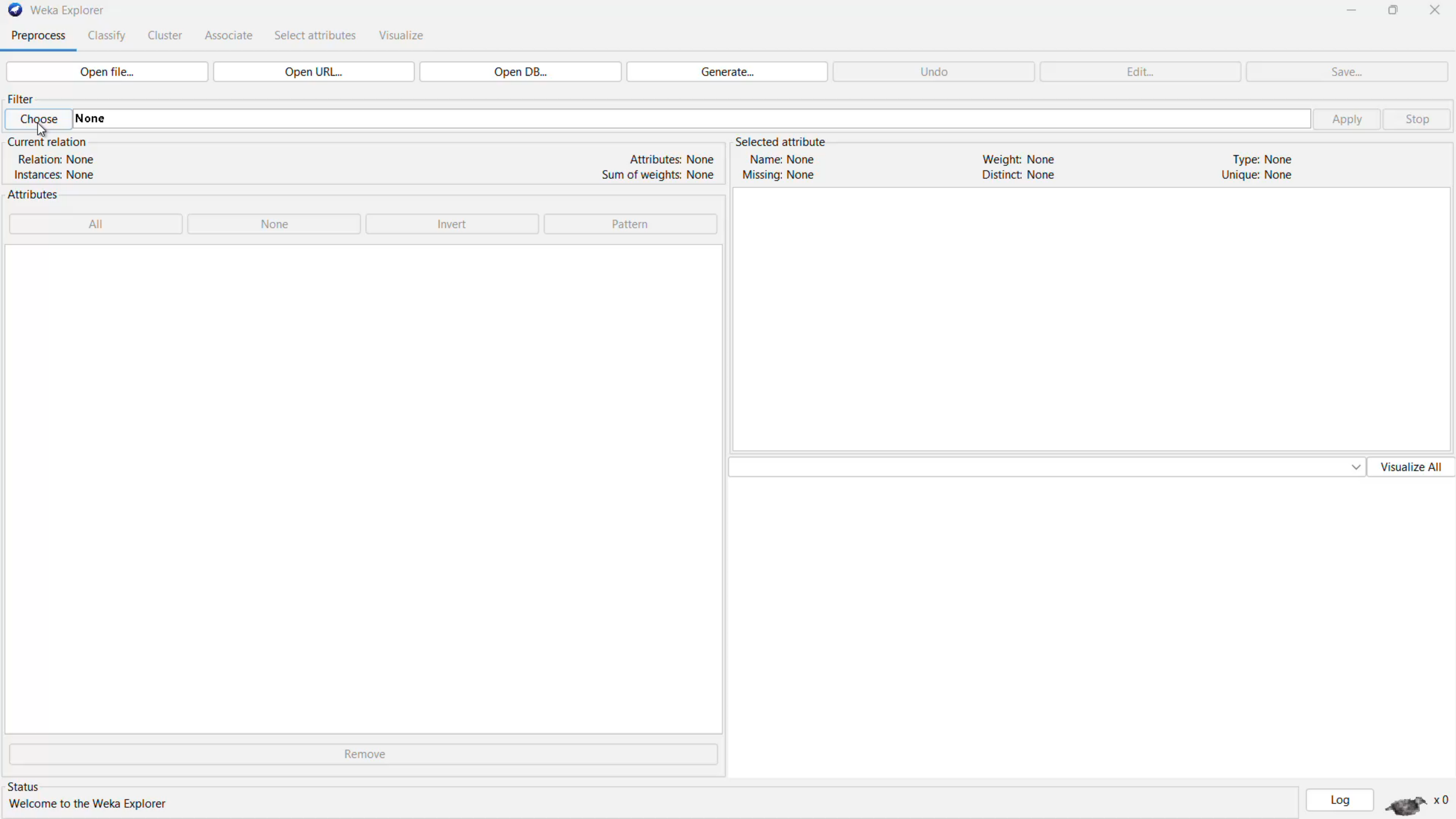
It is important to choose best testing option relevant to the dataset to evaluate its performance.

1. Use training set
2. Supplied test set
3. Cross-validation
4. Percentage Split

10-Fold Cross Validation

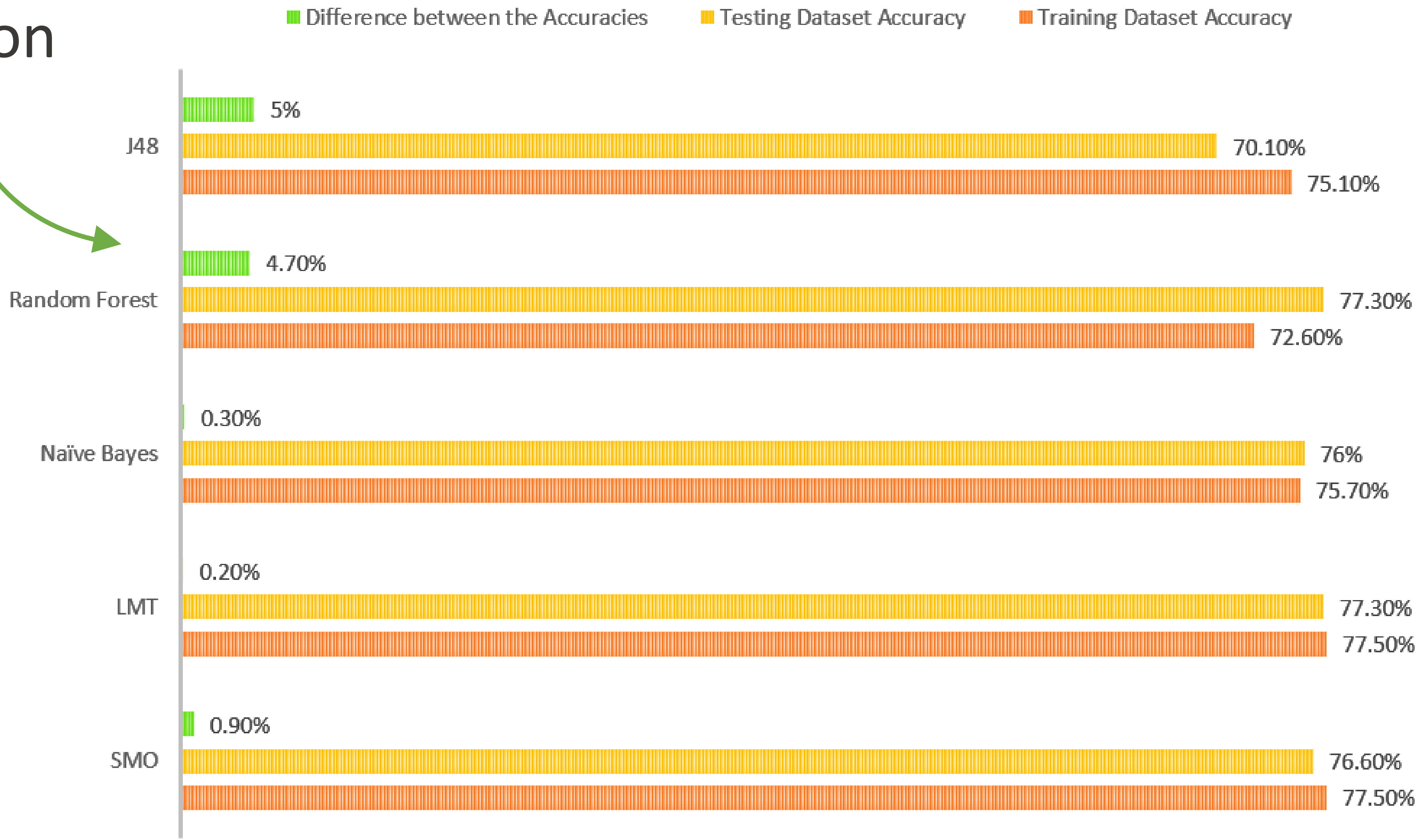
- Reduces overfitting
- Provides a more reliable estimate
- Efficiently uses the data



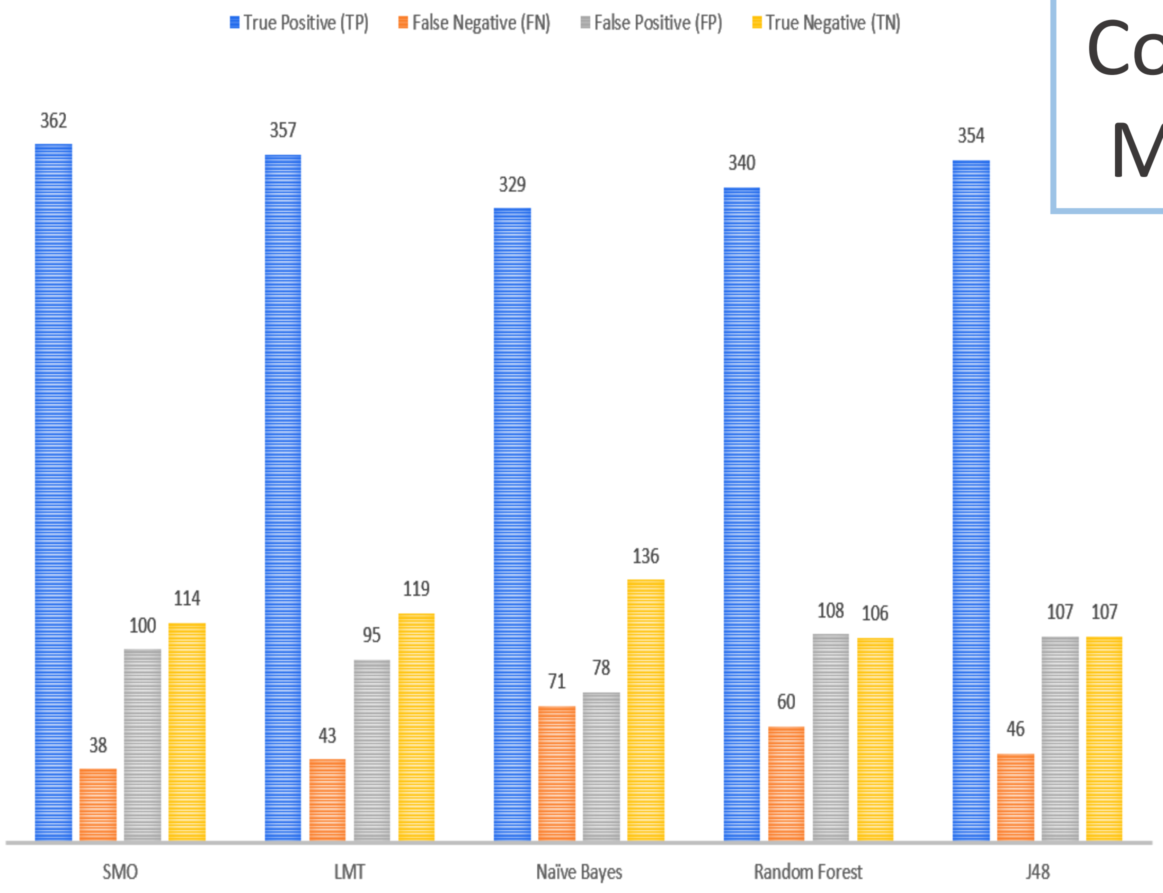

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: None
Instances: None
Attributes: None
Sum of weights: None
Selected attribute
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All
Remove
Status
Welcome to the Weka Explorer
Log
x 0

5. Results

Model Generalization

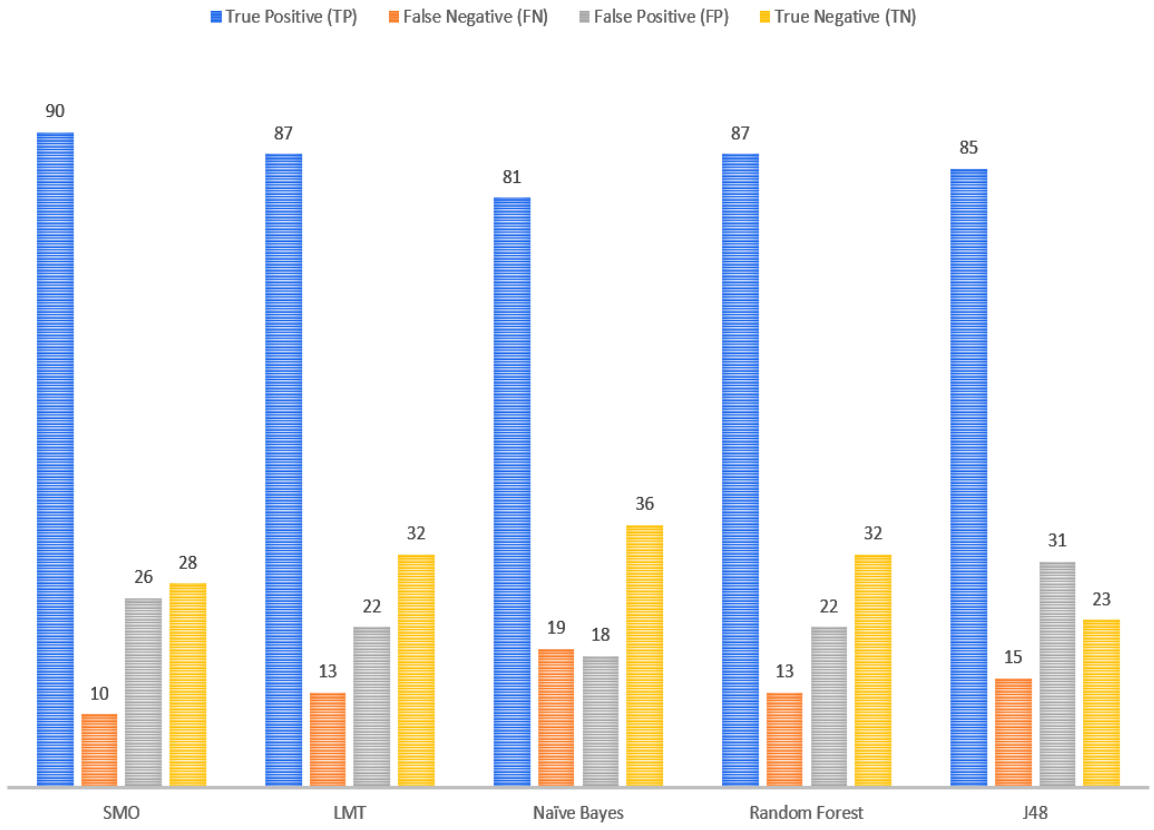


Confusion Matrices



Training Set

Testing Set



4. Choosing the Algorithm

From 5 different types of algorithms; Naïve Bayes, SMO, Random Forest, LMT, J48.

Algorithm	Evaluated test dataset of 154 instances (Accuracy)	
Naïve Bayes	117	74.97%
SMO	118	76.62%
Random Forest	119	77.27%
LMT	119	77.27%
J48	108	70.13%

Questions

THANK YOU