An ML Tool to Detect Heart Disease

CIND-820: Big Data Analytics Project
Robert M. Pineau
941-049-371

Supervisor: Dr. Ceni Babaoglu
December 8th, 2022

Toronto Metropolitan University

Why Heart Disease?

365 x 24 x 14

Every hour, 14 adults die in Canada due to heart disease[1]

Toronto Metropolitan University

1: Public-Health-Agency-of-Canada. (2022)

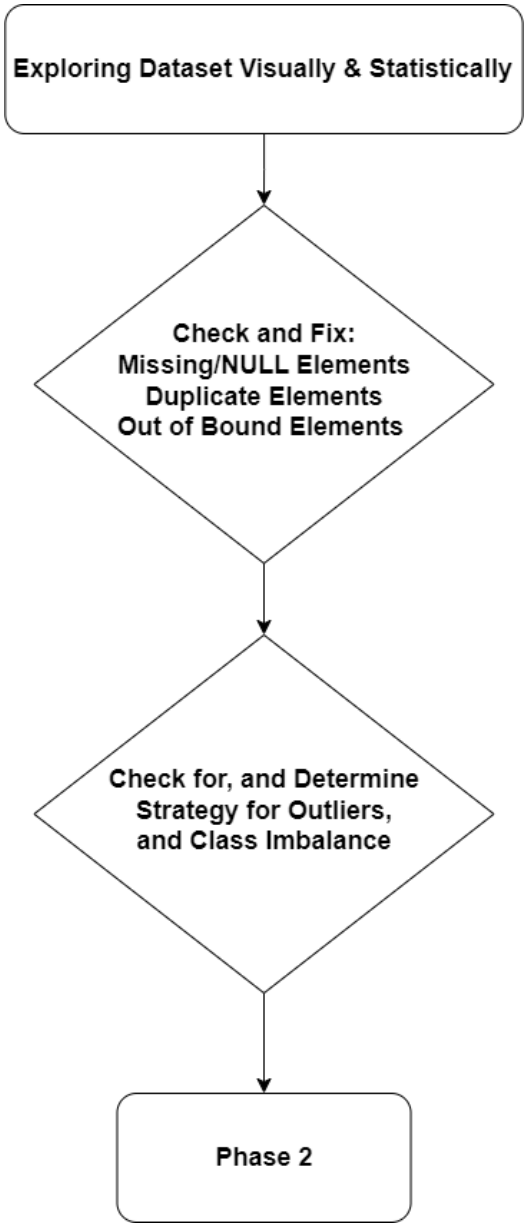
The Data:

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

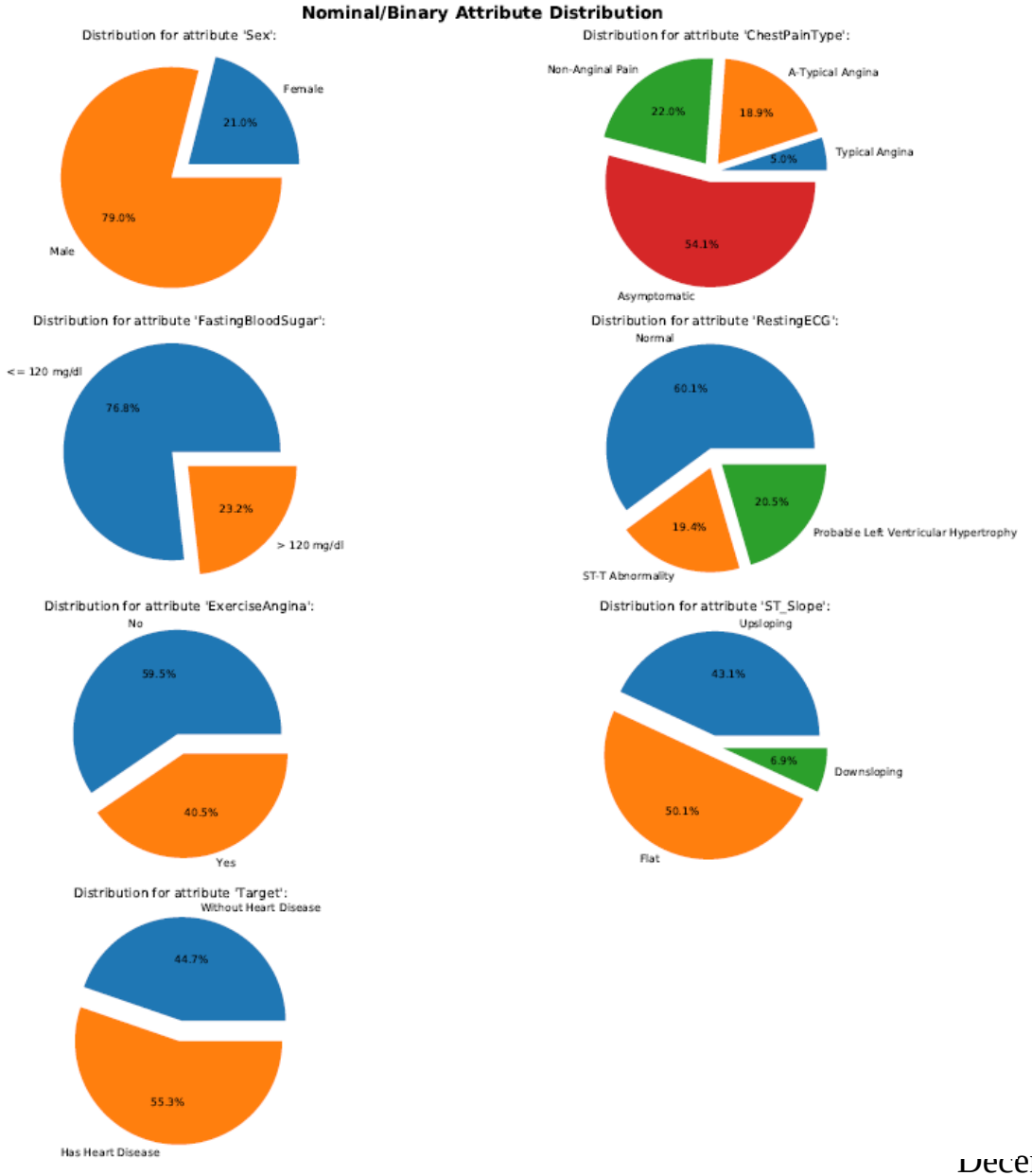
Heart Disease Dataset Attribute Description

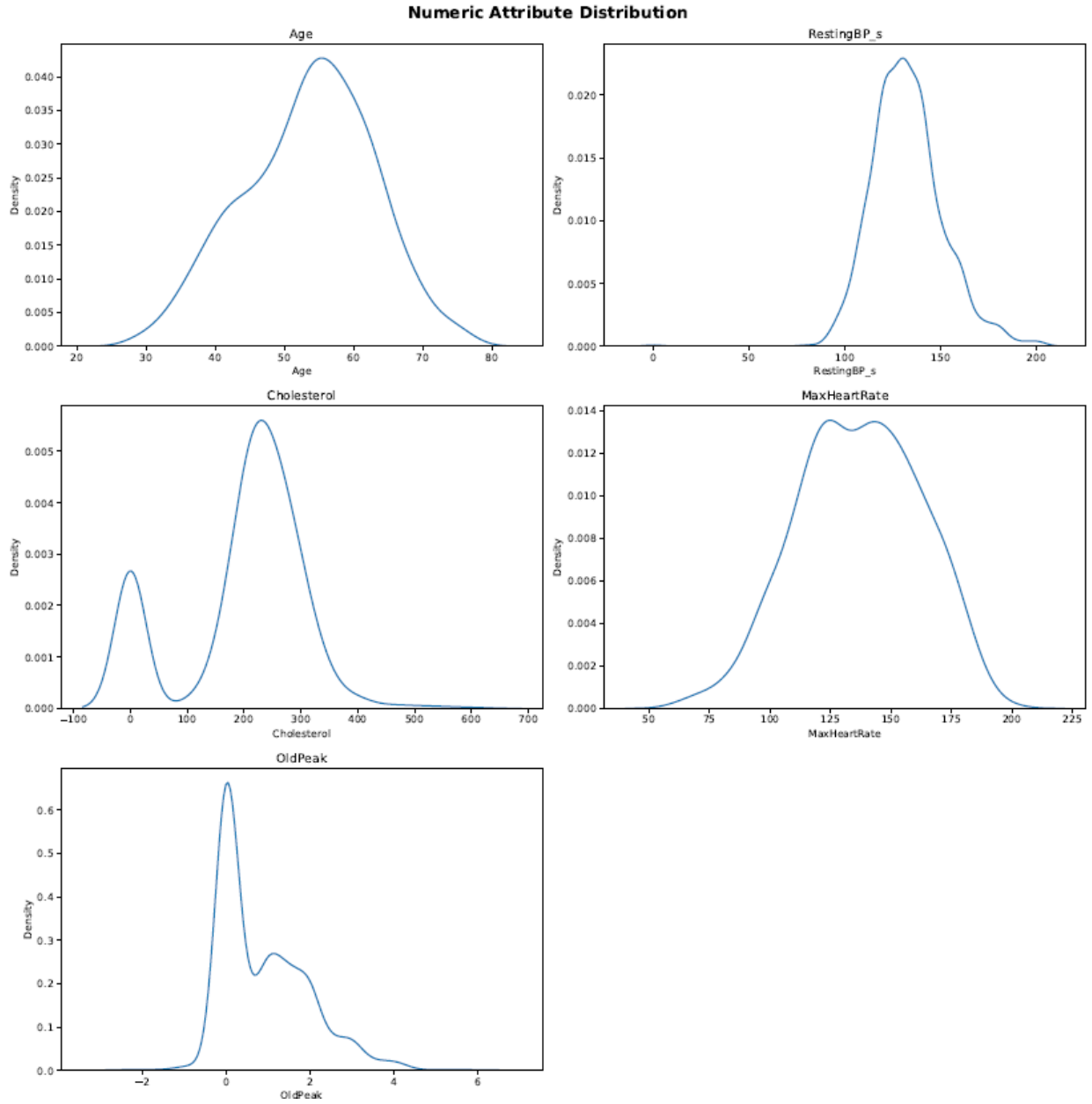
S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

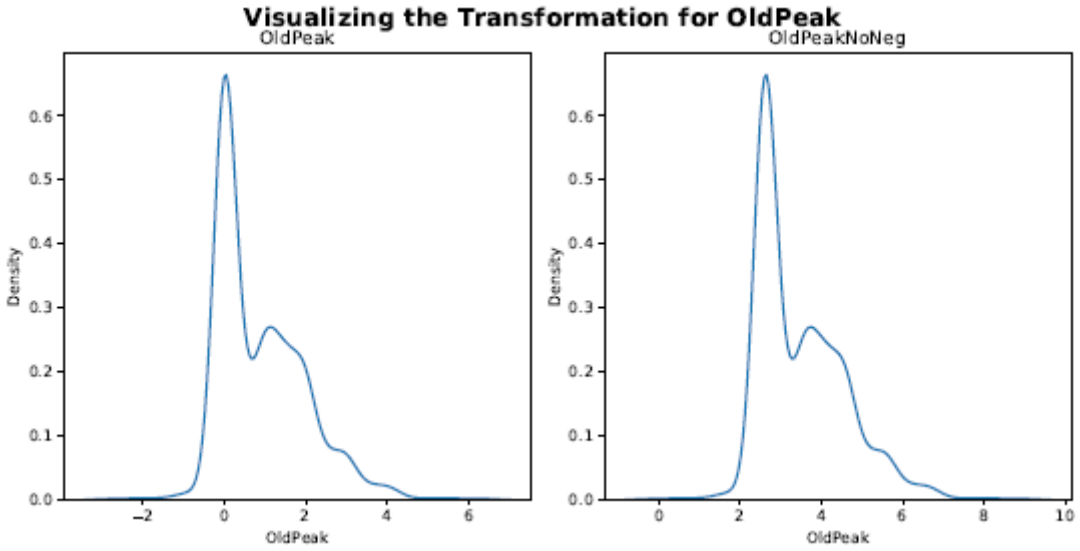
Phase-1:

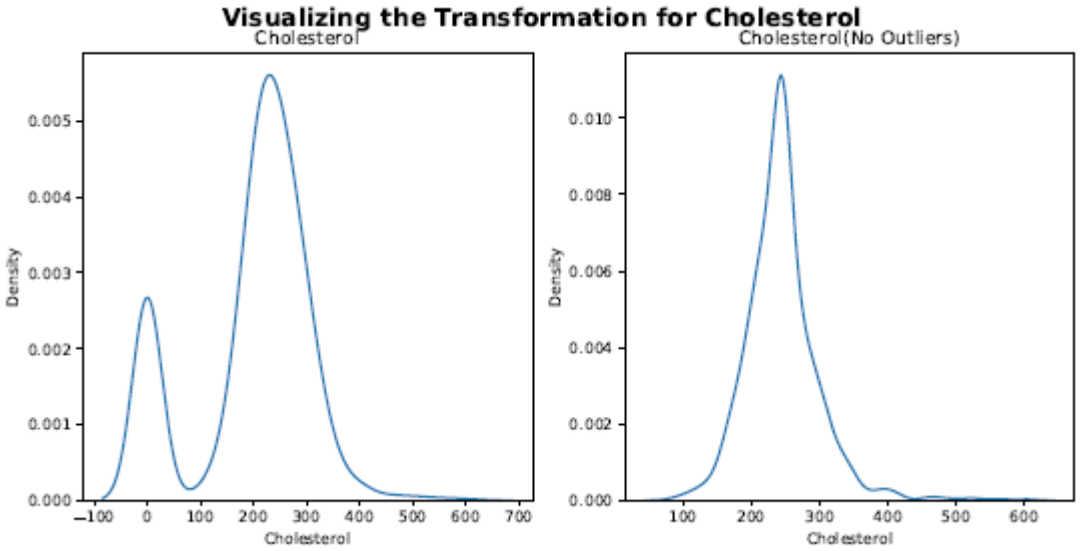


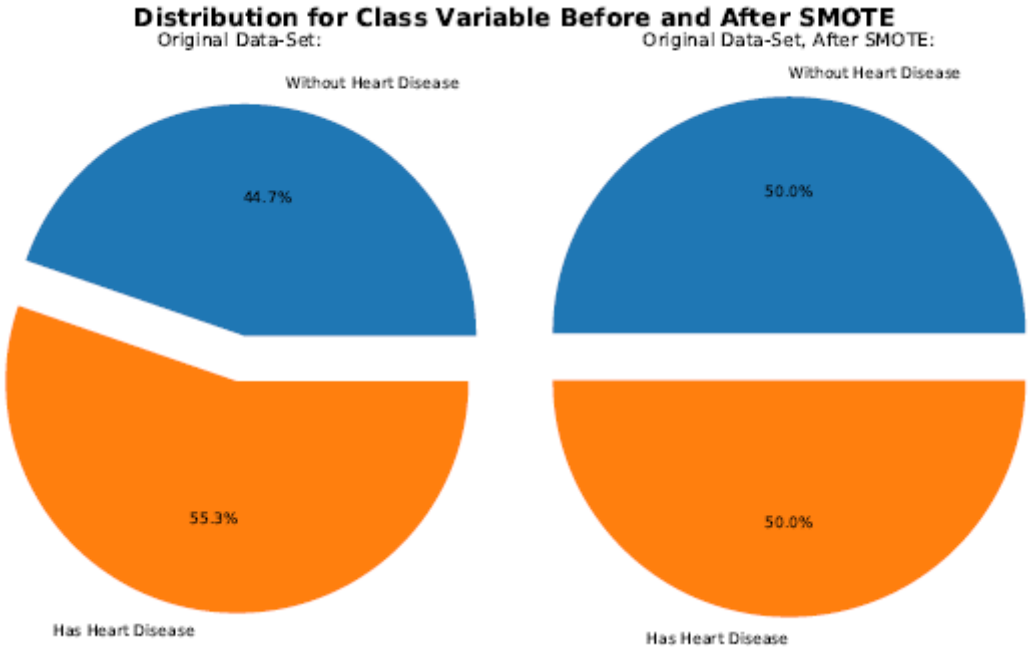
Toronto Metropolitan University


Nominal/Binary Attribute Distribution
Distribution for attribute 'Sex':
Female
21.0%
79.0%
Male
Distribution for attribute 'ChestPainType':
Non-Anginal Pain
22.0%
A-Typical Angina
18.9%
5.0%
Typical Angina
54.1%
Asymptomatic
Distribution for attribute 'FastingBloodSugar':
<= 120 mg/dl
76.8%
23.2%
> 120 mg/dl
Distribution for attribute 'RestingECG':
Normal
60.1%
20.5%
Probable Left Ventricular Hypertrophy
19.4%
ST-T Abnormality
Distribution for attribute 'ExerciseAngina':
No
59.5%
40.5%
Yes
Distribution for attribute 'ST_Slope':
Upsloping
43.1%
6.9%
Downsloping
50.1%
Flat
Distribution for attribute 'Target':
Without Heart Disease
44.7%
55.3%
Has Heart Disease

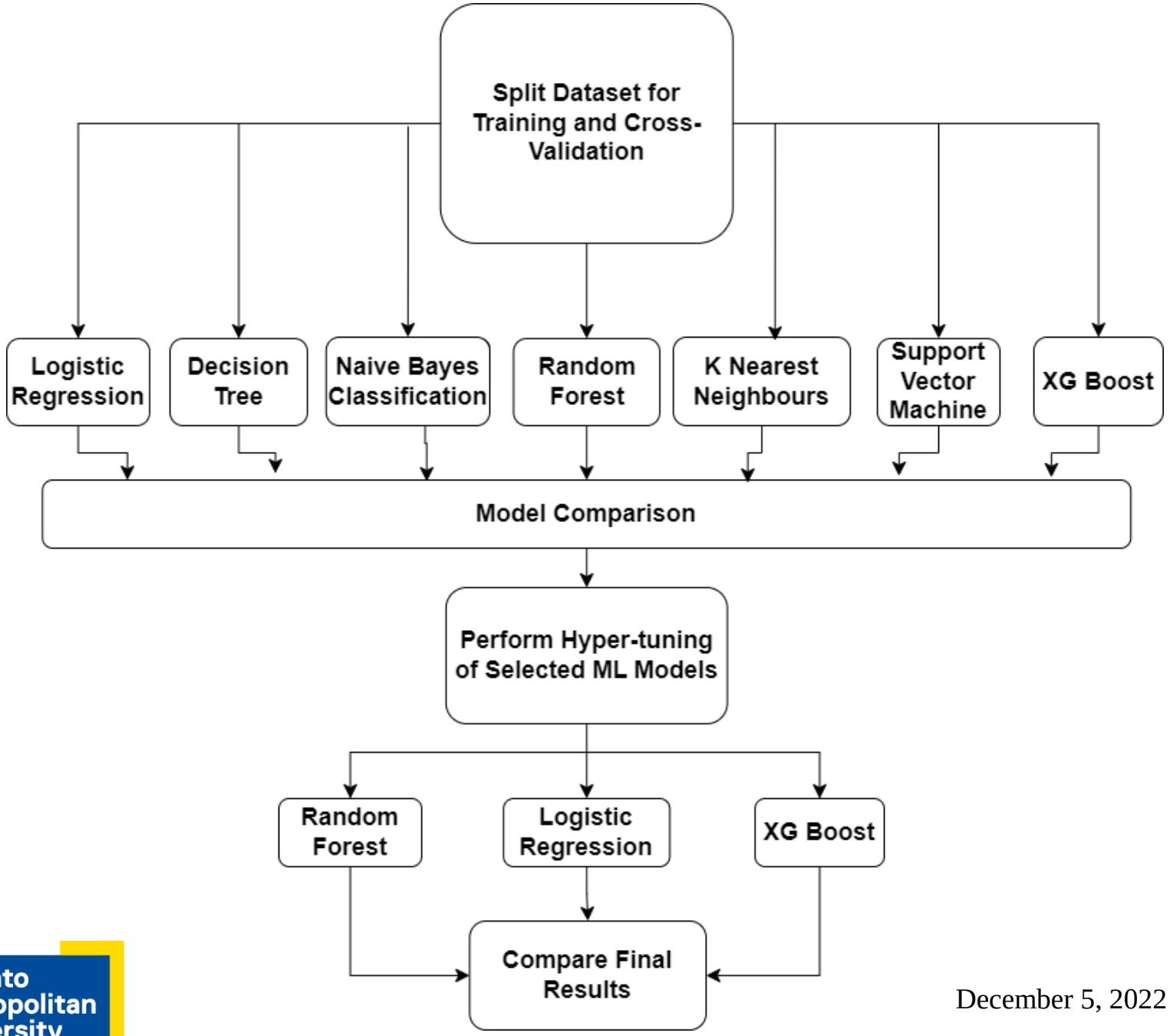

Numeric Attribute Distribution
Age
Density
Age
RestingBP_s
Density
RestingBP_s
Cholesterol
Density
Cholesterol
MaxHeartRate
Density
MaxHeartRate
OldPeak
Density
OldPeak


Visualizing the Transformation for OldPeak
OldPeak
OldPeakNoNeg
Density
OldPeak

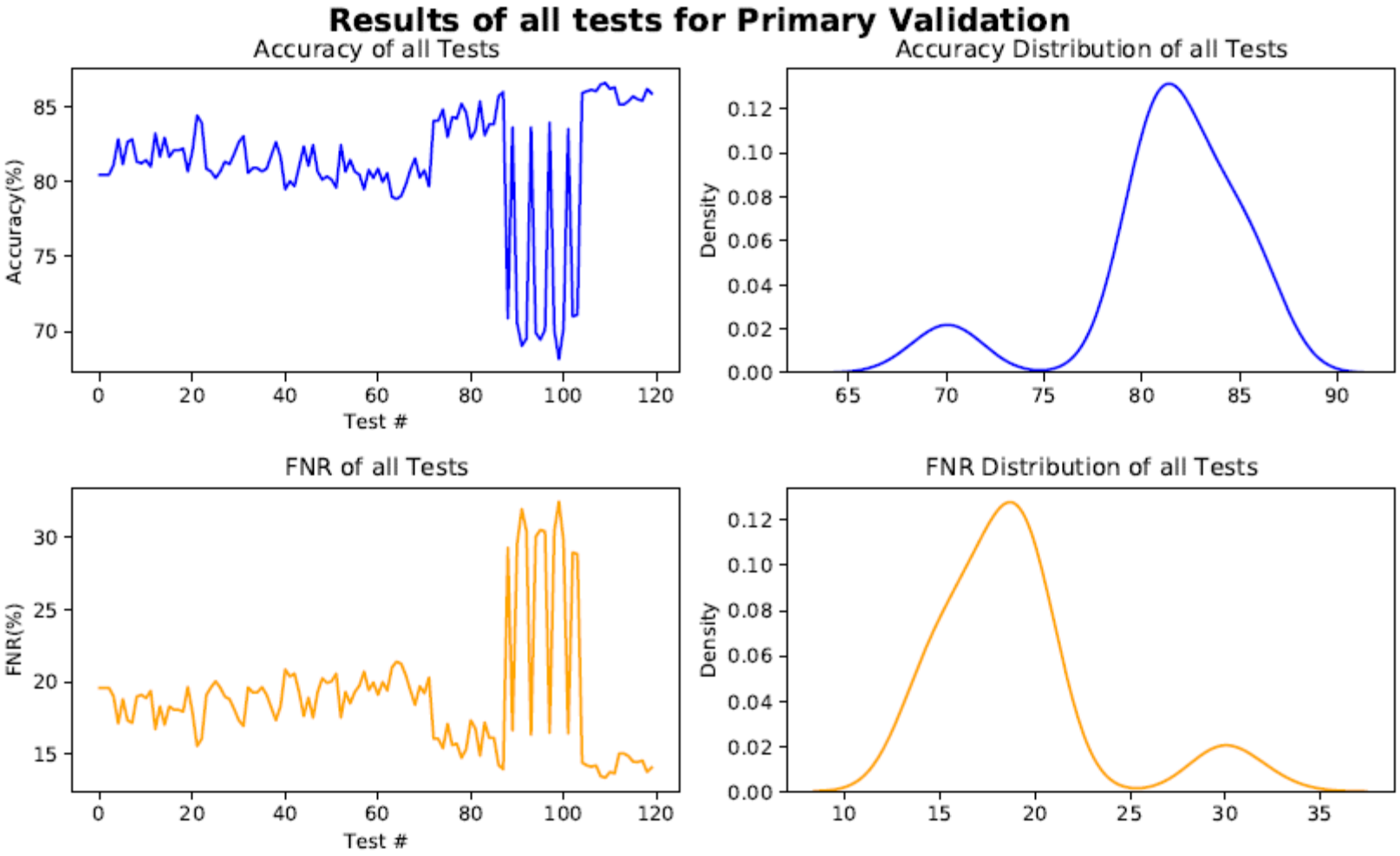

Visualizing the Transformation for Cholesterol
Cholesterol
Cholesterol(No Outliers)
Density
Cholesterol



Toronto Metropolitan University



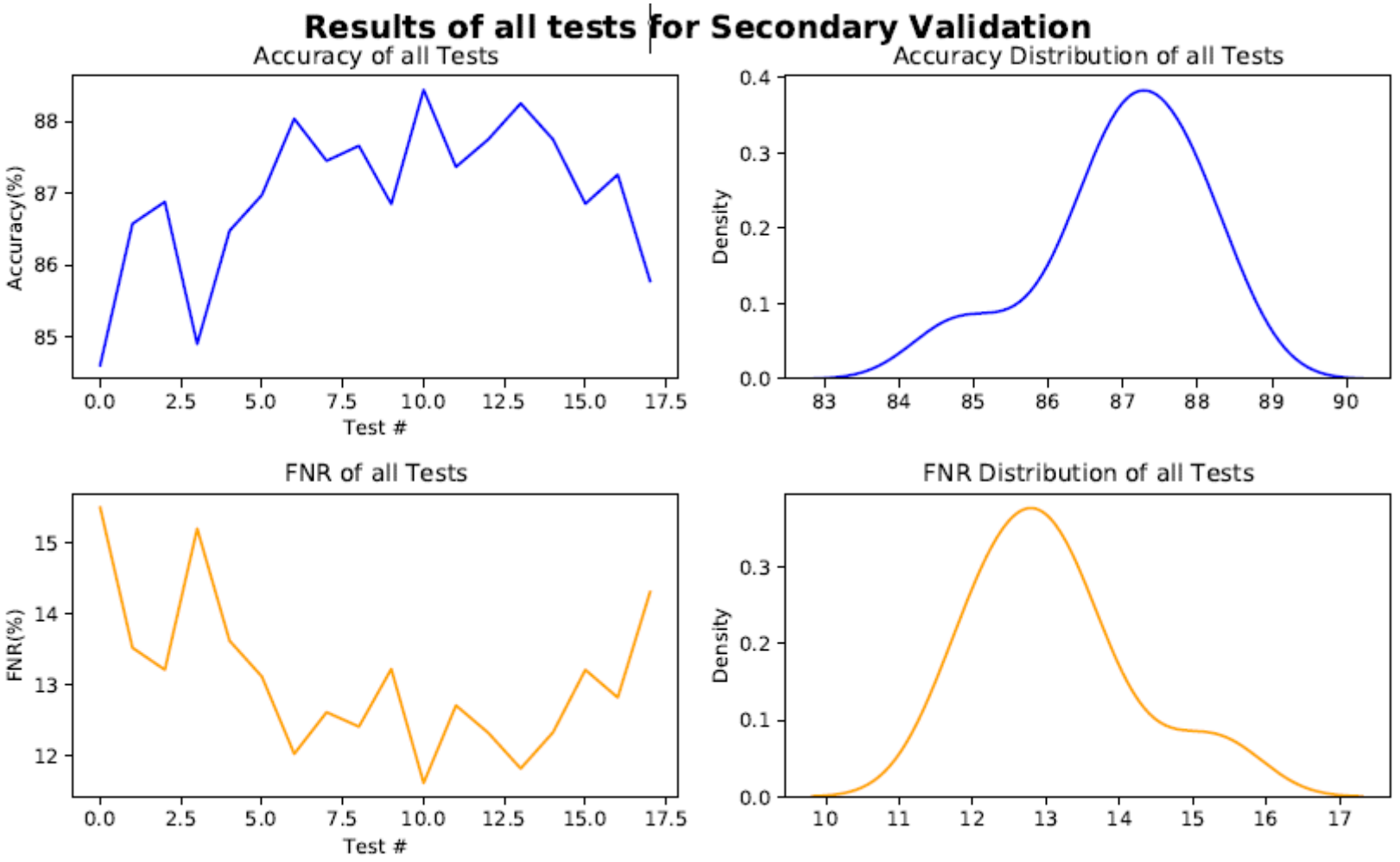
Toronto Metropolitan University



Toronto Metropolitan University

Results of Initial Tests sorted by Accuracy (Top 10)

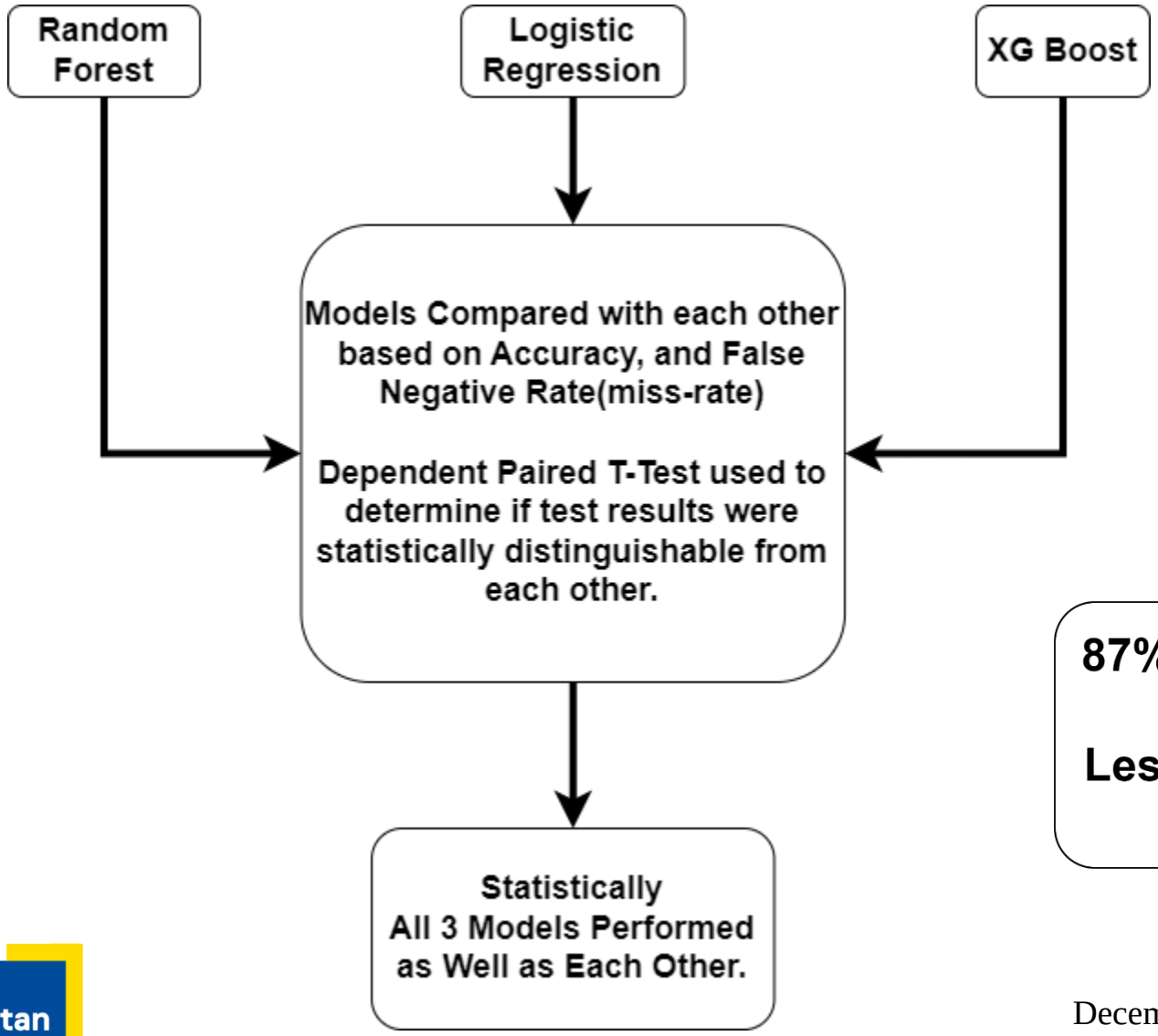
ML Model	Data-Set	Accuracy	FNR
Random Forest(RF)	Normalized, After SMOTE	86.58%	13.37%
Random Forest(RF)	Original, After SMOTE	86.48%	13.46%
Random Forest(RF)	Outliers Addressed, After ONEHOT, After SMOTE	86.29%	13.66%
Random Forest(RF)	Original, After ONEHOT, After SMOTE	86.19%	13.76%
XG Boost(XGB)	Original, After ONEHOT, After SMOTE	86.18%	13.76%
Random Forest(RF)	Original, After ONEHOT	86.13%	14.12%
Random Forest(RF)	Outliers Addressed	86.02%	14.21%
Random Forest(RF)	Normalized	86.02%	14.22%
Logistic Regression(LR)	Outliers Addressed, After ONEHOT, After SMOTE	85.99%	13.95%
Random Forest(RF)	Original	85.91%	14.41%


Results of all tests for Secondary Validation
Accuracy of all Tests
Accuracy(%)
Test #
Accuracy Distribution of all Tests
Density
FNR of all Tests
FNR(%)
Test #
FNR Distribution of all Tests
Density

Results of Secondary Validation sorted by Accuracy (Top 10)

Random Forest(RF)	Original, After ONEHOT, After SMOTE	89.03%	11.02%
Random Forest(RF) - Tuned	Original, After ONEHOT, After SMOTE	88.34%	11.72%
XG Boost(XGB)	Original, After ONEHOT, After SMOTE	88.24%	11.83%
Random Forest(RF)	Original, After SMOTE	88.03%	12.03%
XG Boost(XGB)	Original, After SMOTE	87.85%	12.22%
Random Forest(RF) - Tuned	Outliers Addressed, After ONEHOT, After SMOTE	87.56%	12.52%
XG Boost(XGB) - Tuned	Original, After ONEHOT, After SMOTE	87.55%	12.52%
Random Forest(RF)	Outliers Addressed, After ONEHOT, After SMOTE	87.26%	12.81%
XG Boost(XGB)	Outliers Addressed, After ONEHOT, After SMOTE	87.16%	12.92%
Logistic Regression(LR) - Tuned	Outliers Addressed, After ONEHOT, After SMOTE	87.07%	12.98%

The Best Model?



87%+ Accuracy & Less Than 12 % FNR

Deployment Considerations:

Cloud vs On-Premise?

Data Safety and Privacy Concerns?

Is 87% Accuracy Good Enough for a Medical based tool?

Next-Steps:

Obtain more Data?

Obtain higher quality Data?

More advanced ML Algorithms? Deep learning? Neural Networks?

Turning a trained Model into a usable tool?

Questions?

Toronto Metropolitan University