Characteristic	Dataset 1				Dataset 2				Dataset 3			
Number of variables	12				18				12			
Independent	11				17				11			
Dependent	1				1				1			
Number of records	3075				45211				32581			
Binary	4				4				2			
Nominal	3				6				3			
Categorical	6				6				4			
Textual	0				0				0			
Numerical	8				7				7			
Data Cleaning												
Irrelevant variables removed	1				2				0			
(Number & Proportion)	1/12				2/18				0			
Duplications removed	0				0				165			
(Number & Proportion)	0				0				165/32581			
Technique	NA				NA				Drop Rows			
Dimensionality reduction	Yes				Yes				Yes			
Technique	Dropped columns				Dropped columns				Dropped rows			
Missing values dealt with	3075				12				4011			
(Number & Proportion)	3075/36900				12/813798				4011/390972			
Technique	Drop columns				Filling with Mean values				Filling with Mean values			
Outliers filtered	2 variables				6 variables				3 variables			
Technique	IQR Method				IQR Method				IQR Method			
First four characteristics after cleaning	Vizualize Variables One hot encoding Normalized Split dataset to train and test data				Vizualize Variables One hot encoding Normalized Split dataset to train and test data				Vizualize Variables One hot encoding Normalized Split dataset to train and test data			
Data Normalization	All Variables				All Variables				All Variables			
Technique	MinMaxScaler				MinMaxScaler				MinMaxScaler			
Data balancing and splitting	70% Train 20% Test				70% Train 20% Test				70% Train 20% Test			
Training data size	2152				24471				21406			
Testing data size	923				10488				9174			
Evaluation	RFC 1	RFC 2	LR 1	LR 2	RFC 1	RFC 2	LR 1	LR 2	RFC 1	RFC 2	LR 1	LR 2
Accuracy	97.61%	98.04%	96.09%	98.37%	92.86%	92.93%	92.37%	92.41%	93.12%	93.39%	86.73%	86.83%
F1 Score	92.41%	93.95%	87.05%	94.98%	43.75%	49.21%	39.20%	40%	80.59%	81.76%	61.90%	62%

RFC 1	Random forest classifier - Scikit learn library
RFC 2	Random forest classifier - XGBoost library
LR 1	Logistic Regression - Scikit learn library
LR 2	Logistic Regression - Stats Module library