Machine Learning to Predict Cardiovascular Disease - Full Results

Name: Natalia Rubio
SUNet ID: 06130110
Department of Mechanical Engineering
Stanford University
nrubio@stanford.edu

1 Linear and Logisitic Regression

	Logistic Regression						**Linear Regression**					
Feature Set:	Basic			Technical			Basic			Technical		
Regularization Type:	none	L1	L2	none	L1	L2	none	L1	L2	none	L1	L2
(Un-normalized Data)												
Cleveland	0.69	0.69	0.69	0.75	0.77	0.75	0.64	0.67	0.64	0.83	0.78	0.83
Hungary	0.69	0.69	0.69	0.85	0.78	0.85	0.75	0.75	0.75	0.79	0.77	0.79
Long Beach	0.75	0.78	0.75	0.7	0.73	0.7	0.73	0.75	0.73	0.63	0.7	0.63
Switzerland	0.97	0.97	0.97	0.86	0.86	0.86	0.91	0.94	0.91	0.78	0.89	0.78
all	0.71	0.74	0.71	0.73	0.74	0.73	0.75	0.78	0.75	0.81	0.81	0.81
(Normalized Data)												
Cleveland	0.62	0.56	0.62	0.69	0.56	0.69	0.65	0.65	0.65	0.83	0.8	0.83
Hungary	0.47	0.29	0.47	0.71	0.29	0.71	0.75	0.75	0.75	0.79	0.76	0.79
Long Beach	0.8	0.78	0.8	0.8	0.78	0.8	0.75	0.75	0.75	0.7	0.7	0.7
Switzerland	0.97	0.97	0.97	0.86	0.97	0.86	0.97	0.97	0.97	0.89	0.89	0.89
all	0.60	0.57	0.60	0.67	0.57	0.67	0.78	0.78	0.78	0.81	0.81	0.81

Table 1: Validation accuracy of linear and logistic regression trained and evaluated on different data sets.

2 Support Vector Machine

Validation Set Accuracy (Correct Binary Prediction of Cardiovascular Disease)

	SV Regression						**SV Classification**					
Kernel:	Linear			RBF			Linear			RBF		
Regularization Coefficient:	0.1	1	10	0.1	1	10	0.1	1	10	0.1	1	10
(Un-normalized Data, Basic Features)												
Cleveland	0.65	0.64	0.65	0.43	0.43	0.50	0.68	0.68	0.65	0.44	0.44	0.49
Hungary	0.80	0.79	0.79	0.70	0.70	0.73	0.67	0.75	0.75	0.71	0.71	0.77
Long Beach	0.73	0.73	0.73	0.78	0.76	0.75	0.73	0.71	0.71	0.76	0.76	0.76
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
All	0.76	0.77	0.77	0.57	0.59	0.67	0.77	0.77	0.77	0.60	0.61	0.69
(Normalized Data, Basic Features)												
Cleveland	0.58	0.57	0.6	0.51	0.64	0.61	0.67	0.65	0.65	0.50	0.65	0.65
Hungary	0.80	0.79	0.79	0.80	0.77	0.80	0.73	0.75	0.75	0.75	0.80	0.82
Long Beach	0.71	0.71	0.71	0.78	0.75	0.7	0.71	0.71	0.71	0.76	0.76	0.7
Switzerland	0.97	0.97	0.97	0.97	0.97	1.0	0.97	0.97	0.97	0.97	0.97	0.97
All	0.76	0.76	0.76	0.74	0.74	0.73	0.76	0.76	0.76	0.78	0.77	0.77
(Un-normalized Data, Technical Features):												
Cleveland	0.8	0.75	0.78	0.43	0.55	0.70	0.82	0.82	0.83	0.44	0.47	0.74
Hungary	0.89	0.87	0.87	0.70	0.71	0.78	0.84	0.80	0.79	0.71	0.70	0.79
Long Beach	0.68	0.63	0.68	0.78	0.78	0.78	0.71	0.68	0.68	0.76	0.76	0.76
Switzerland	0.78	0.78	0.81	0.97	0.97	0.83	0.89	0.89	0.86	0.97	0.97	0.97
all	0.75	0.75	0.67	0.61	0.68	0.70	0.81	0.81	0.79	0.59	0.72	0.74
(Normalized Data, Technical Features):												
Cleveland	0.76	0.75	0.75	0.47	0.76	0.72	0.8	0.8	0.77	0.48	0.8	0.69
Hungary	0.87	0.86	0.86	0.83	0.84	0.75	0.82	0.78	0.76	0.83	0.77	0.73
Long Beach	0.75	0.71	0.66	0.78	0.8	0.73	0.71	0.66	0.66	0.76	0.78	0.61
Switzerland	0.83	0.83	0.83	0.97	0.86	0.75	0.97	0.86	0.86	0.97	0.97	0.86
all	0.74	0.74	0.73	0.78	0.78	0.78	0.81	0.80	0.81	0.78	0.81	0.81

Table 2: Preliminary Results: Validation accuracy of SV regression and classification trained and evaluated on different data sets.

3 Neural Network

Activation Function:	None			ReLU			Logistic		
Regularization Coefficient:	10^1	10^{-2}	10^{-5}	10^1	10^{-2}	10^{-5}	10^1	10^{-2}	10^{-5}
(Regression, Basic Features)									
Cleveland	0.69	0.69	0.69	0.67	0.69	0.69	0.70	0.69	0.69
Hungary	0.70	0.67	0.67	0.75	0.74	0.74	0.70	0.68	0.68
Long Beach	0.78	0.78	0.78	0.73	0.73	0.73	0.76	0.75	0.75
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
all	0.72	0.72	0.72	0.75	0.74	0.75	0.67	0.74	0.74
(Classification, Basic Features)									
Cleveland	0.65	0.65	0.65	0.67	0.69	0.67	0.44	0.65	0.65
Hungary	0.71	0.75	0.75	0.71	0.74	0.74	0.71	0.74	0.74
Long Beach	0.73	0.75	0.75	0.73	0.66	0.65	0.76	0.71	0.71
Switzerland	0.97	0.94	0.94	0.97	0.94	0.94	0.97	0.97	0.97
all	0.78	0.77	0.77	0.79	0.77	0.77	0.55	0.78	0.78
(Regression, Technical Features)									
Cleveland	0.77	0.77	0.77	0.76	0.68	0.68	0.84	0.65	0.70
Hungary	0.86	0.80	0.80	0.85	0.77	0.78	0.29	0.70	0.70
Long Beach	0.75	0.75	0.75	0.8	0.63	0.65	0.76	0.63	0.63
Switzerland	0.86	0.86	0.86	0.86	0.70	0.70	0.97	0.75	0.75
all	0.74	0.72	0.72	0.75	0.7	0.70	0.75	0.73	0.73
(Classification, Technical Features)									
Cleveland	0.81	0.78	0.78	0.78	0.72	0.72	0.72	0.75	0.77
Hungary	0.71	0.82	0.82	0.71	0.76	0.80	0.71	0.82	0.82
Long Beach	0.71	0.73	0.73	0.73	0.63	0.63	0.76	0.7	0.68
Switzerland	0.97	0.83	0.83	0.97	0.86	0.86	0.97	0.86	0.83
all	0.81	0.79	0.79	0.80	0.73	0.74	0.55	0.80	0.80

Table 3: Preliminary Results: Validation accuracy of neural network regression and classification trained and evaluated on different data sets.

4 Gaussian Mixture

Number of Gaussians:	1	2	3	4	5	6
(Un-normalized, Basic Features)						
Cleveland	0.44	0.44	0.56	0.51	0.56	0.58
Hungary	0.71	0.71	0.60	0.70	0.70	0.71
Long Beach	0.76	0.76	0.76	0.76	0.76	0.71
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97
all	0.55	0.48	0.55	0.54	0.56	0.63
(Normalized, Basic Features)						
Cleveland	0.44	0.44	0.58	0.57	0.57	0.58
Hungary	0.71	0.68	0.73	0.75	0.73	0.75
Long Beach	0.76	0.76	0.71	0.71	0.65	0.66
Switzerland	0.97	0.97	0.97	0.97	0.91	0.97
all	0.55	0.47	0.71	0.73	0.72	0.71
(Un-normalized, Technical Features)						
Cleveland	0.44	0.44	0.6	0.61	0.6	0.50
Hungary	0.71	0.69	0.70	0.79	0.64	0.68
Long Beach	0.76	0.76	0.76	0.76	0.7	0.75
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97
all	0.55	0.57	0.62	0.70	0.70	0.7
(Normalized, Technical Features)						
Cleveland	0.44	0.44	0.58	0.57	0.57	0.58
Hungary	0.71	0.68	0.73	0.75	0.73	0.75
Long Beach	0.76	0.76	0.71	0.71	0.65	0.66
Switzerland	0.97	0.97	0.97	0.97	0.91	0.97
all	0.55	0.47	0.71	0.73	0.72	0.71

Table 4: Preliminary Results: Validation accuracy of Gaussian mixture classifiers trained and evaluated on different data sets.

5 Principal Components Analysis

	Regression						Classification					
Number of Principal Components:	1	2	3	4	5	6	1	2	3	4	5	6
(Un-normalized, Basic Features)												
Cleveland	0.51	0.50	0.63	0.67	0.71	0.69	0.44	0.51	0.48	0.72	0.72	0.64
Hungary	0.52	0.52	0.69	0.69	0.68	0.69	0.67	0.67	0.69	0.69	0.75	0.75
Long Beach	0.78	0.78	0.8	0.8	0.78	0.75	0.76	0.76	0.78	0.76	0.73	0.73
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.94	0.97
all	0.57	0.59	0.62	0.61	0.62	0.74	0.57	0.59	0.63	0.62	0.65	0.77
(Normalized, Basic Features)												
Cleveland	0.64	0.71	0.67	0.72	0.69	0.69	0.48	0.61	0.61	0.67	0.67	0.67
Hungary	0.69	0.69	0.69	0.69	0.68	0.69	0.70	0.71	0.71	0.71	0.74	0.75
Long Beach	0.78	0.76	0.76	0.75	0.78	0.78	0.76	0.75	0.71	0.71	0.71	0.75
Switzerland	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.97	0.94	0.94
all	0.57	0.68	0.74	0.74	0.73	0.73	0.64	0.75	0.76	0.77	0.77	0.78
(Un-normalized, Technical Features)												
Cleveland	0.54	0.58	0.58	0.61	0.68	0.67	0.44	0.44	0.48	0.6	0.74	0.71
Hungary	0.47	0.51	0.66	0.67	0.64	0.76	0.68	0.69	0.70	0.69	0.68	0.78
Long Beach	0.78	0.78	0.78	0.78	0.76	0.76	0.76	0.76	0.75	0.75	0.75	0.75
Switzerland	0.97	0.91	0.86	0.86	0.89	0.89	0.97	0.97	0.97	0.97	0.97	0.97
all	0.57	0.57	0.57	0.57	0.64	0.64	0.50	0.55	0.54	0.58	0.63	0.62
(Normalized, Technical Features)												
Cleveland	0.8	0.8	0.82	0.82	0.89	0.85	0.72	0.74	0.78	0.81	0.8	0.81
Hungary	0.75	0.74	0.73	0.70	0.71	0.75	0.83	0.82	0.79	0.82	0.82	0.83
Long Beach	0.78	0.78	0.8	0.8	0.78	0.8	0.76	0.8	0.8	0.78	0.76	0.75
Switzerland	0.97	0.94	0.94	0.94	0.94	0.94	0.97	0.97	0.97	0.94	0.97	0.97
all	0.57	0.57	0.74	0.77	0.76	0.77	0.55	0.55	0.79	0.79	0.79	0.79

Table 5: Preliminary Results: Validation accuracy of PCA-transformed linear and logistic regression models trained and evaluated on different data sets.

UCI ML Repository Dataset Attributes

Attribute	Description	Type	Class	Commonly Used	Comments
id	patient identification number		unused	no	irrelevant
ccf	social security number		unused	no	irrelevant
age	age in years	continuous	basic	yes	
sex	sex (1 = male; 0 = female)	binary	basic	yes	
painloc	chest pain location (1 = substernal; 0 = otherwise)	binary	basic	no	
painexer	(1 = provoked by exertion; 0 = otherwise)	binary	basic	no	
relrest	(1 = relieved after rest; 0 = otherwise)	binary	basic	no	
pncaden	pncaden (sum of 5, 6, and 7)...unclear	continuous	unused	no	unclear
cp	chest pain type	categorical	basic	yes	
trestbps	resting blood pressure	continuous	technical	yes	
htn	hypertension	binary	technical	no	
chol	serum cholestoral	continuous	technical	yes	
smoke	is or is not a smoker	binary	basic	no	
cigs	cigarettes per day	continuous	basic	no	
years	number of years as a smoker	continuous	basic	no	
fbs	fasting blood sugar	continuous	technical	yes	
dm	history of diabetes	binary	basic	no	
famhist	family history of coronary artery disease	binary	basic	no	
restecg	resting electrocardiographic results	categorical	technical	no	
ekgmo	month of exercise ECG reading	continuous	unused	no	irrelevant
ekgday	day of exercise ECG reading	continuous	unused	no	irrelevant
ekgyr	year of exercise ECG reading	continuous	unused	no	irrelevant
dig	digitalis used furing exercise	binary	technical	no	
prop	Beta blocker used during exercise ECG	binary	technical	no	
nitr	nitrates used during exercise ECG	binary	technical	no	
pro	calcium channel blocker used during exercise ECG	binary	technical	no	
diuretic	diuretic used used during exercise ECG	binary	technical	no	
proto	exercise protocol	categorical	technical	no	

thaldur	duration of exercise test	continuous	technical	no	
thaltime	time when ST measure depression was noted	continuous	technical	no	
met	mets achieved	continuous	technical	no	
thalach	maximum heart rate achieved	continuous	technical	yes	
thalrest	resting heart rate	continuous	technical	no	
tpeakbps	peak exercise blood pressure (systolic)	continuous	technical	no	
tpeakbpd	peak exercise blood pressure (diastolic)	continuous	technical	no	
dummy			unused	no	
trestbpd	resting blood pressure	continuous	technical	no	
exang	exang: exercise induced angina	binary	technical	yes	
xhypo	hypoxia?	binary	technical	no	
oldpeak	ST depression induced by exercise relative to rest	continuous	technical	yes	
slope	slope of the peak exercise ST segment	categorical	technical	yes	
rldv5	height at rest	continuous	technical	no	
rldv5e	height at peak exercise	continuous	technical	no	
ca	number of major vessels (0-3) colored by flourosopy	continuous	technical	yes	
restckm	irrelevant		unused	no	missing data
exerckm	irrelevant		unused	no	missing data
restef	rest raidonuclid ejection fraction		unused	no	missing data
restwm	rest wall motion abnormality	categorical	unused	no	missing data
exeref	exeref: exercise radinalid ejection fraction		unused	no	missing data
exerwm	exerwm: exercise wall (sp?) motion		unused	no	missing data
thal	unclear	categorical	technical	yes	
thalsev	not used		unused	no	missing data
thalpul	not used		unused	no	missing data
earlobe	not used		unused	no	missing data
cmo	month of cardiac catheterization	continuous	unused	no	irrelevant
cday	day of cardiac catheterization	continuous	unused	no	irrelevant
cyr	year of cardiac catheterization	continuous	unused	no	irrelevant
num	diagnosis of heart disease	binary	technical	yes	

lmt	narrowing of left main trunk artery	binary	technical	no	
ladprox	narrowing of left anterior descending artery (proximal)	binary	technical	no	
laddist	narrowing of left anterior descending artery (distal)	binary	technical	no	
diag	narrowing of left anterior descending artery (diagonal branch)	binary	technical	no	
cxmain	narrowing of main circumflex artery	binary	technical	no	
ramus	narrowing of ramus coronary artery	binary	technical	no	
om1	narrowing of obtuse marginal coronary artery 1	binary	technical	no	
om2	narrowing of obtuse marginal coronary artery 2	binary	technical	no	
rcaprox	narrowing of right coronary artery (proximal)	binary	technical	no	
rcadist	narrowing of right coronary artery (distal)	binary	technical	no	
lvx1	narrowing of left circumflex artery 1	binary	unused	no	labeled unused
lvx2	narrowing of left circumflex artery 4	binary	unused	no	labeled unused
lvx3	narrowing of left circumflex artery 3	binary	unused	no	labeled unused
lvx4	narrowing of left circumflex artery 4	binary	unused	no	labeled unused
lvf	left ventricular failure?	binary	unused	no	labeled unused
cathef	not used		unused	no	missing data
junk	not used		unused	no	missing data

Table 6: UCI ML Repository Dataset Attributes. Attributes are as labeled in the data. Descriptions are taken from dataset or inferred from context.