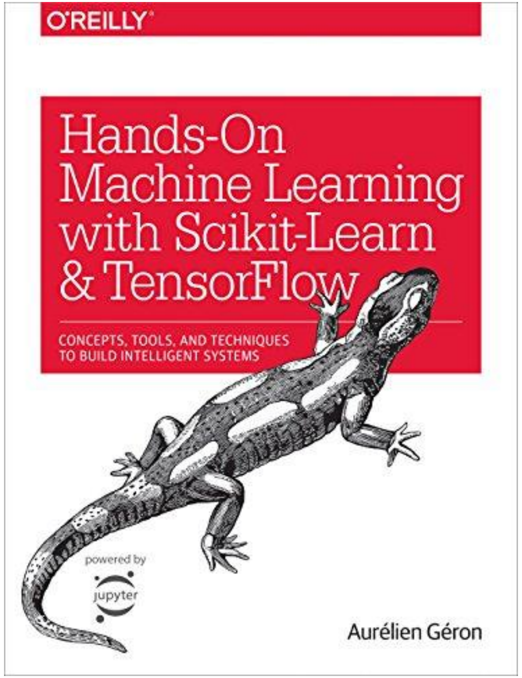
OSA Case Study

Gaining Intuition from Data

Sergio Pérez Morillo

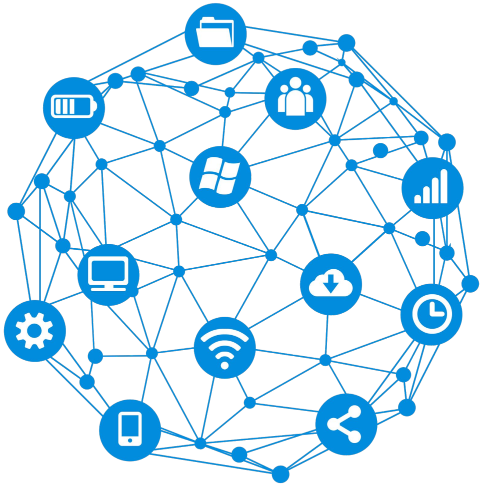
Methodology

1. Framing the Problem
2. Wrangling the Data
3. Exploratory Data Analysis
4. Data Preparation
5. Model Testing & Fine-Tuning
6. Results & Model Comparison



Framing the Problem

1. Framing the Problem



Google Scholar

1. Framing the Problem

Definition

Symptoms

Diagnosis

Treatment

Forgetfulness

Snoring

Somnolence

Discontent

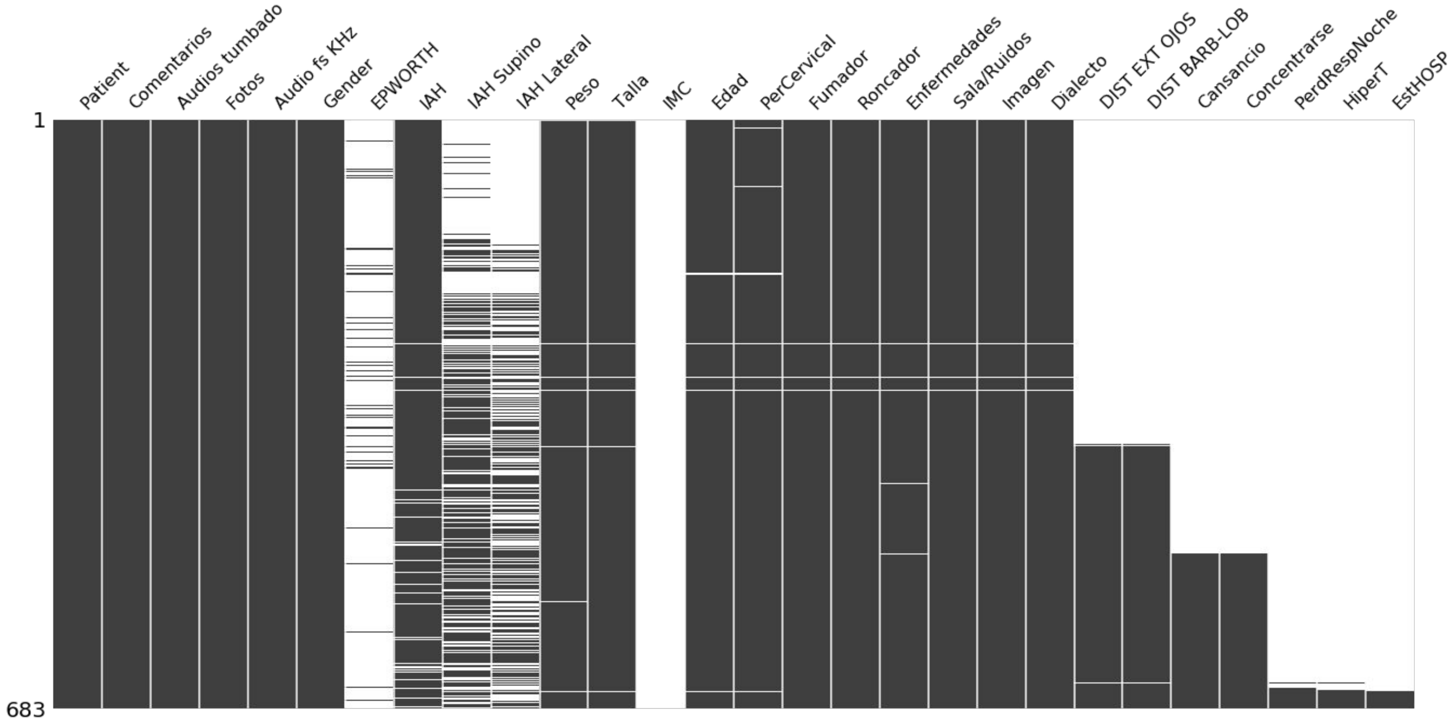
Collar Size

Age

Weight

Wrangling the Data

2. Wrangling the Data



Python Module: Missingno

2. Wrangling the Data

Patient	IAH	Weight
Gender	Age	Height
Snorer	Illness	Smoker

2. Wrangling the Data

Column

Patient	0
Gender	0
IAH	34
Peso	8
Talla	7
Edad	8
PerCervical	12
Fumador	3
Roncador	3
Enfermedades	5

Row

260	8
299	8
314	8
663	4
657	4
178	2
179	2
180	2
2	2
379	2

2. Wrangling the Data

Numerical Features

	Age	IAH	Cervical	Weight	Height	BMI
mean	49.50	20.39	40.64	87.73	171.28	29.86
std	12.39	18.60	3.96	18.36	9.56	5.62
min	20.00	0.00	30.00	45.00	144.00	18.29
25%	40.00	6.40	38.00	75.00	165.00	26.04
50%	49.00	14.40	41.00	86.00	171.00	28.73
75%	59.00	30.00	43.00	98.00	178.00	32.77
max	88.00	108.60	53.00	165.00	197.00	63.65

2. Wrangling the Data

Categorical Features

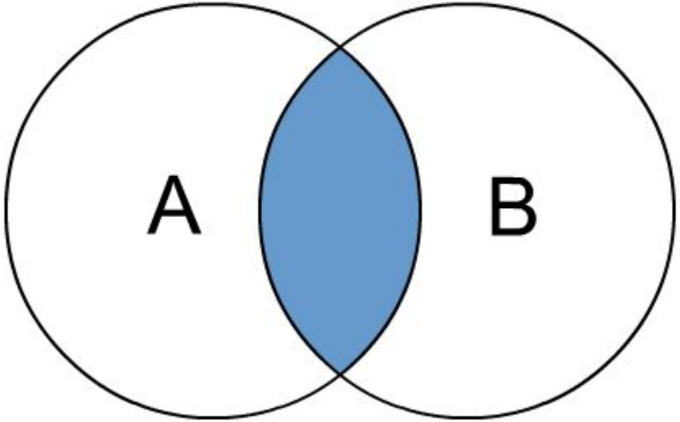
Features	Number of Categories	
	Original	Updated
Gender	2	2
Smoker	6	4
Snorer	8	4
Illness	249	3

2. Wrangling the Data

Classification Dataset

Category	Condition	Count
Severe	IAH >= 30	83
Healthy	IAH <= 10	91

Target Feature

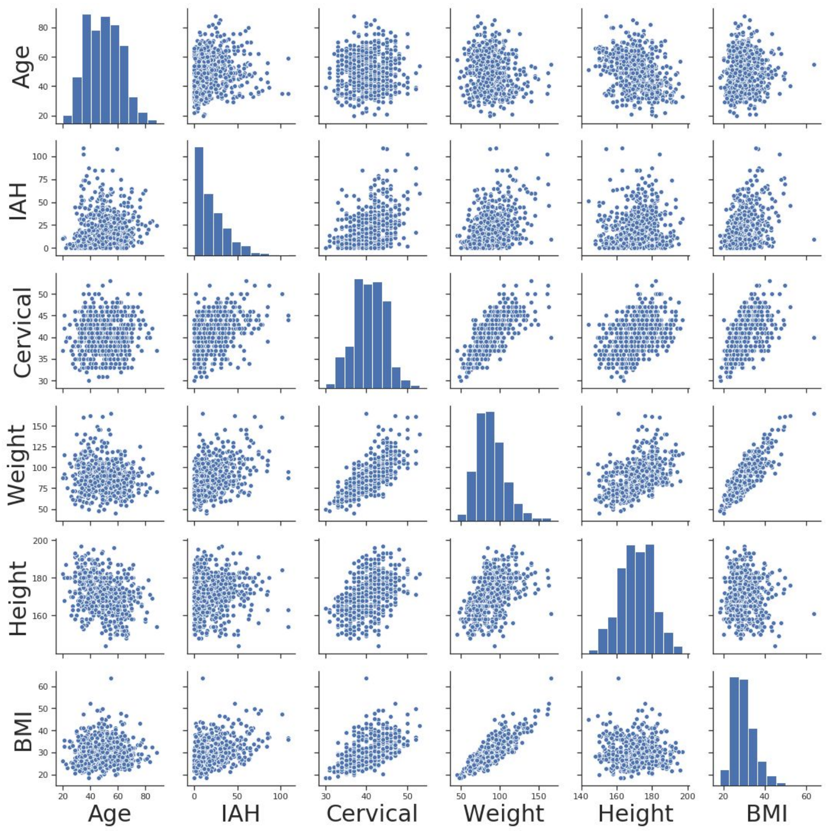


Inner Join

Exploratory Data Analysis

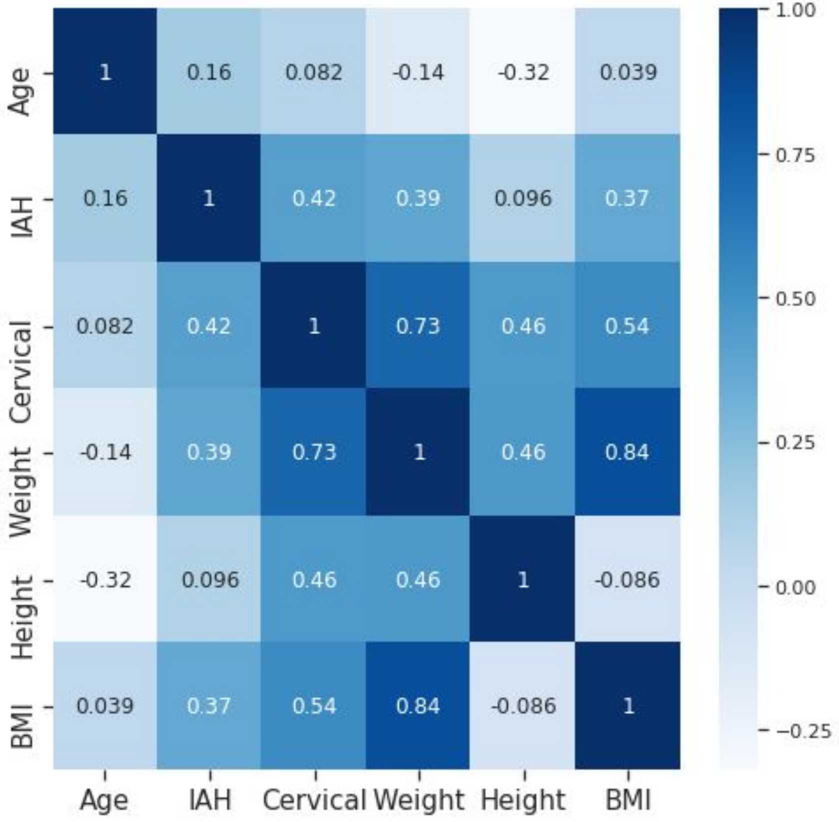
3. Exploratory Data Analysis

Numerical Features



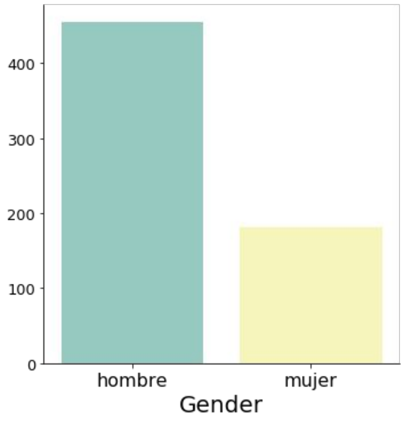
3. Exploratory Data Analysis

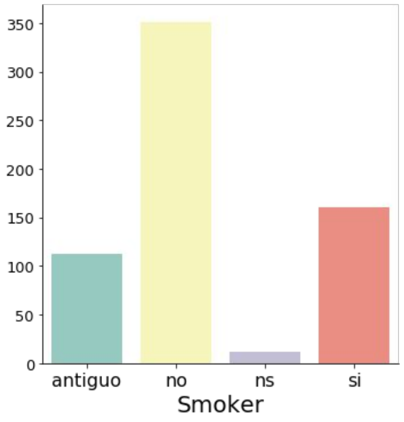
Numerical Features

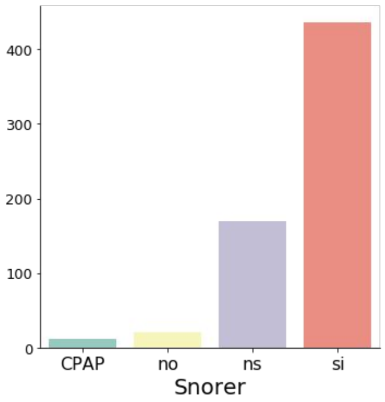


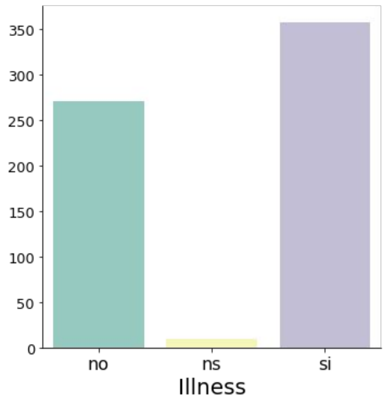
3. Exploratory Data Analysis

Categorical Features

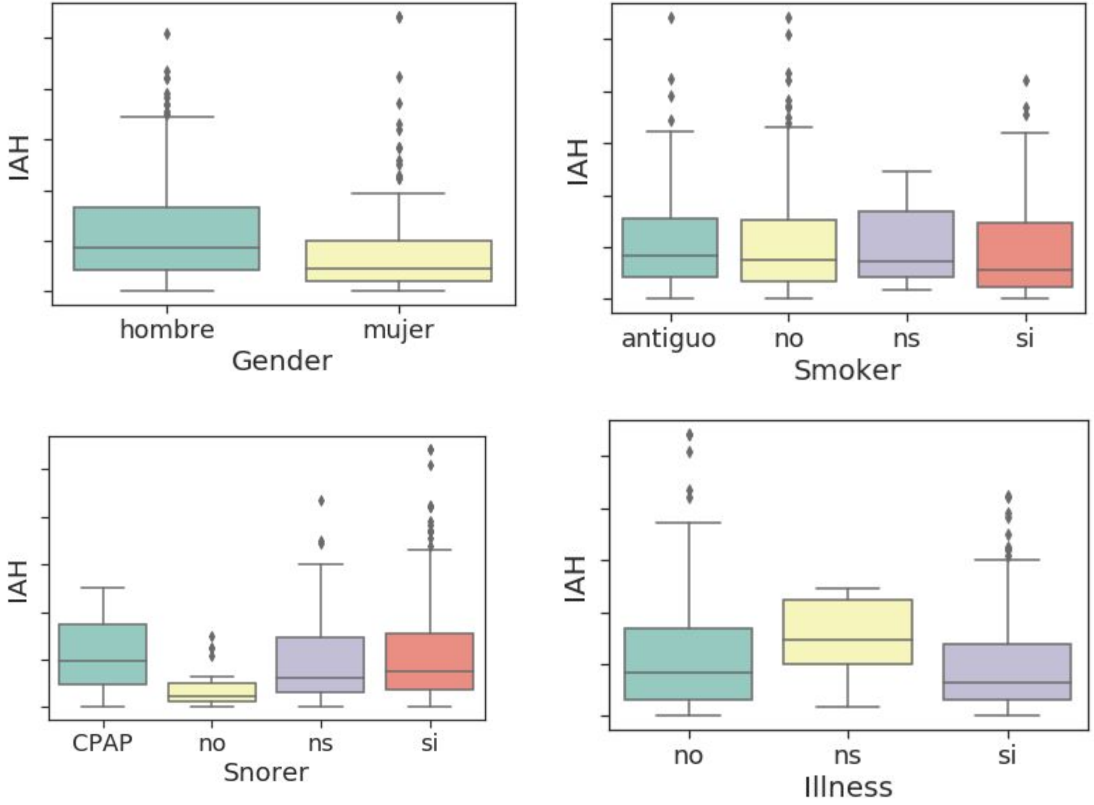




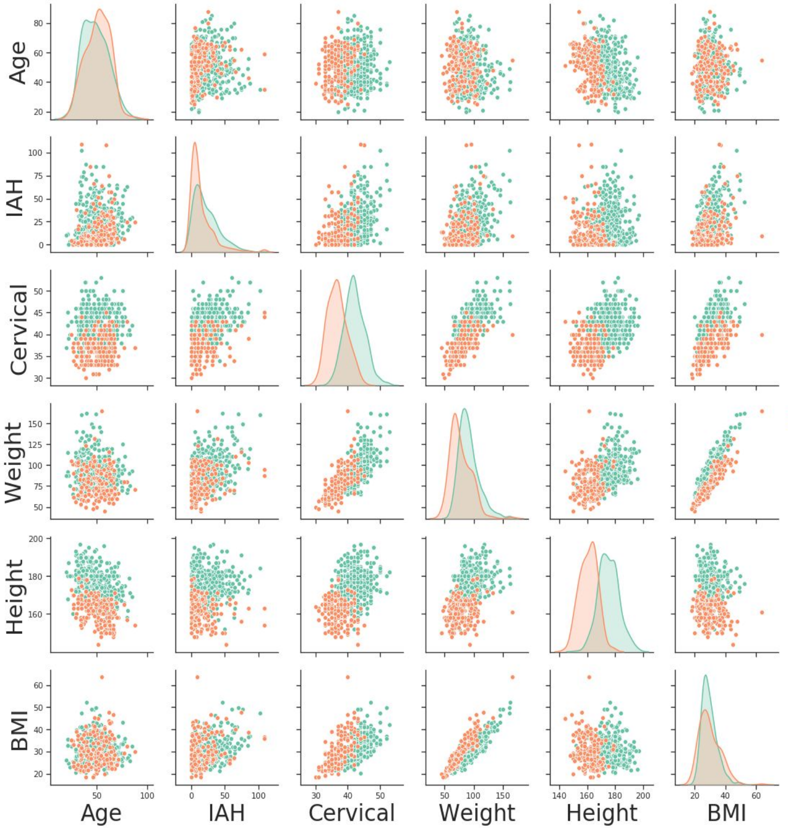


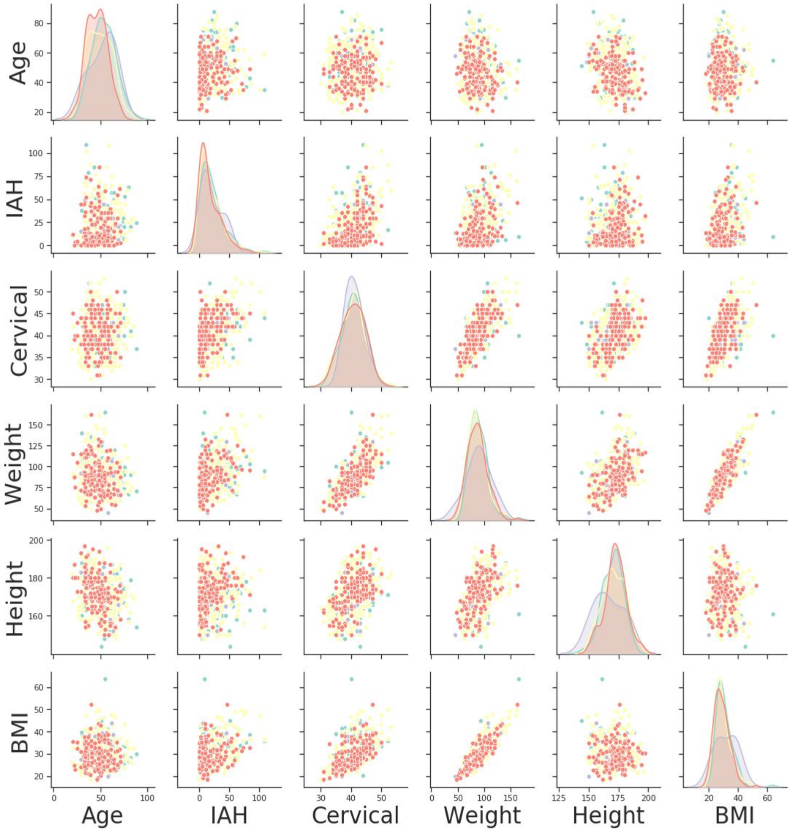


3. Exploratory Data Analysis



3. Exploratory Data Analysis





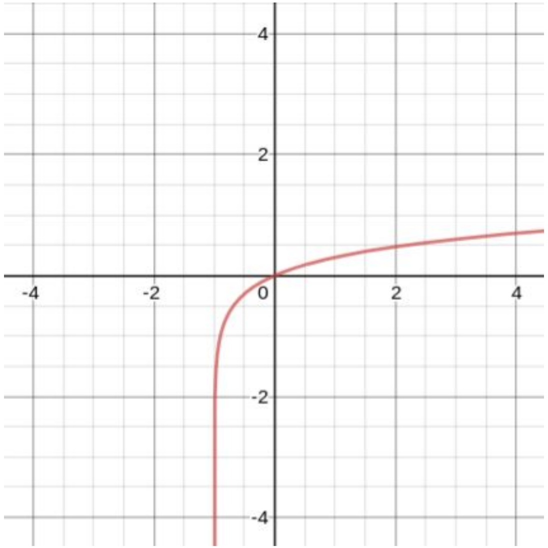
Data Preparation

4. Data Preparation

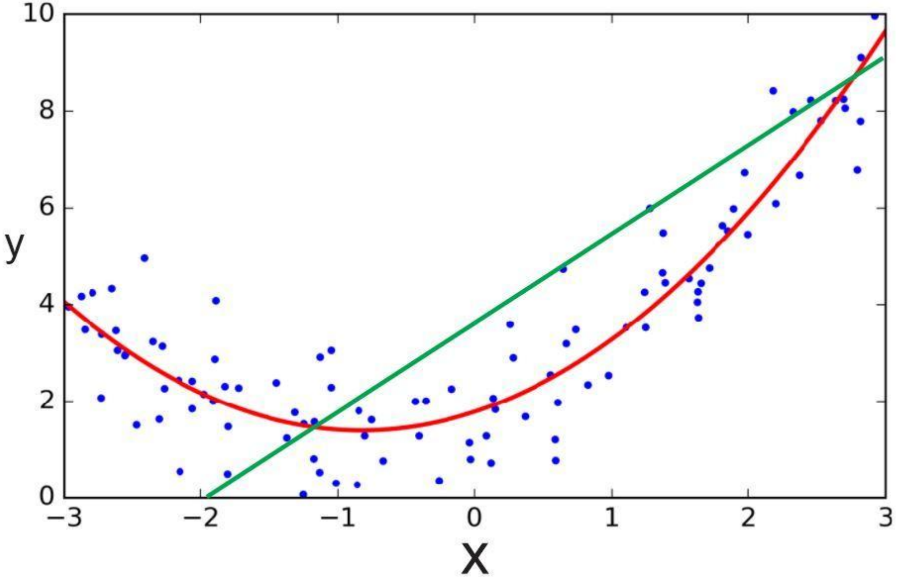
Data Transformation log(x+1), polynomials	**Data Scaling** normalization, standardization
Dimensionality Reduction PCA, t-SNE	**Feature Selection** filtering, wrapping, embedded

4.1. Data Transformation

log(x+1)



Polynomial Features



4.2. Data Scaling

Normalization

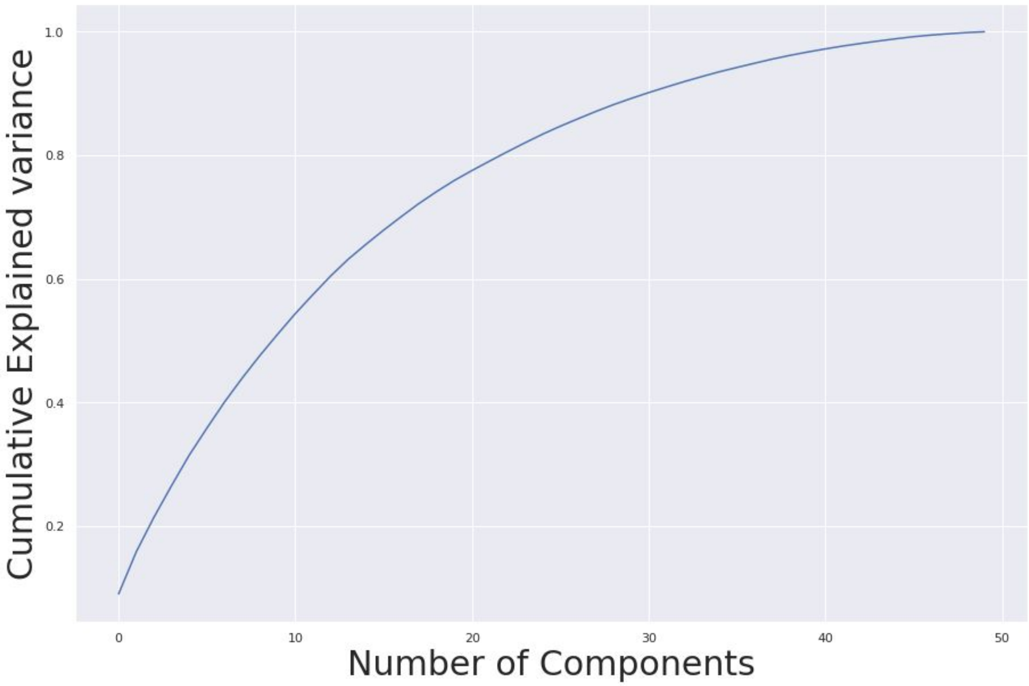
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization

$$z = \frac{x - \mu}{\sigma}$$

4.3. Dimensionality Reduction

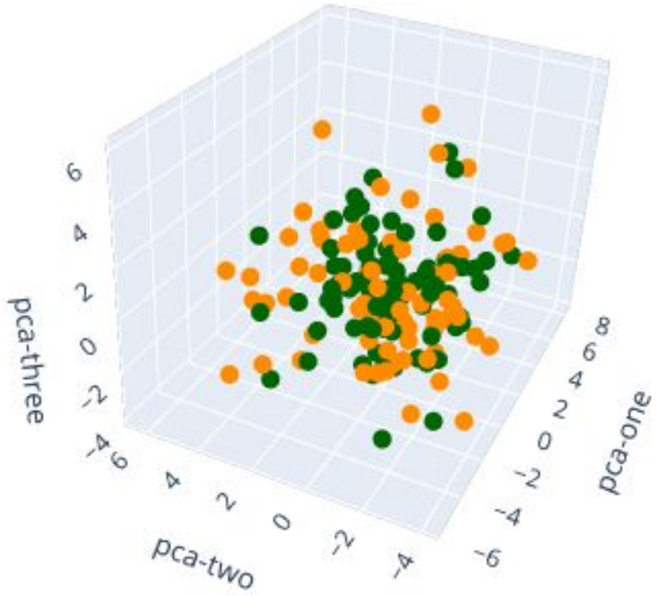
PCA



4.3. Dimensionality Reduction

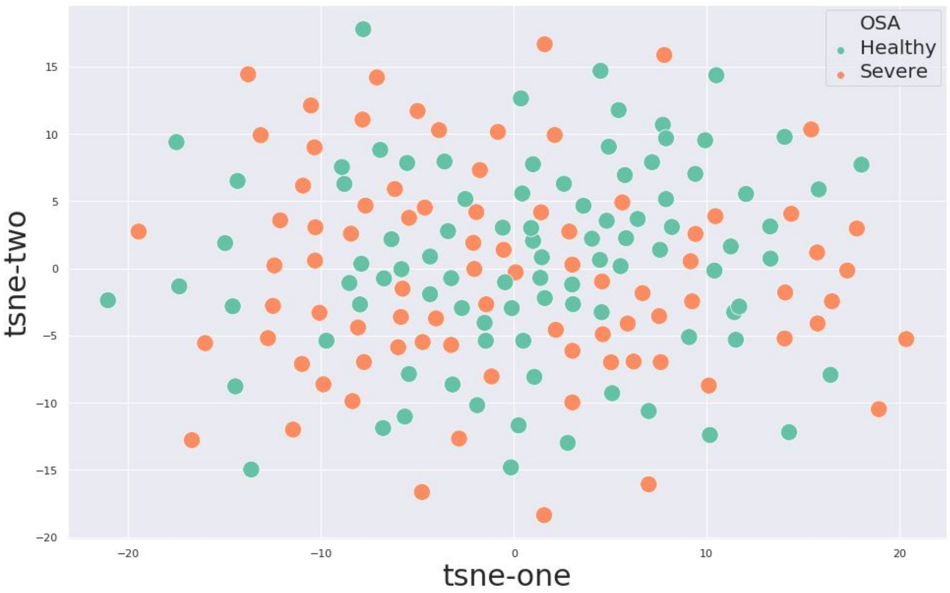
2-Component PCA

3-Component PCA

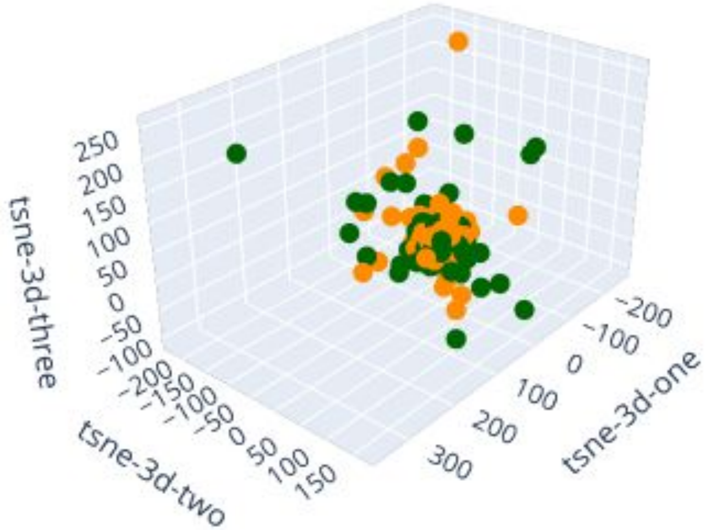


4.3. Dimensionality Reduction

2-Component t-SNE



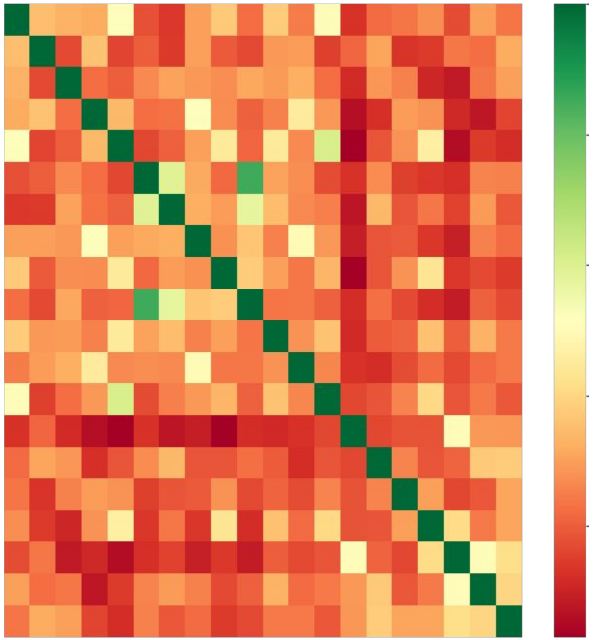
3-Component t-SNE



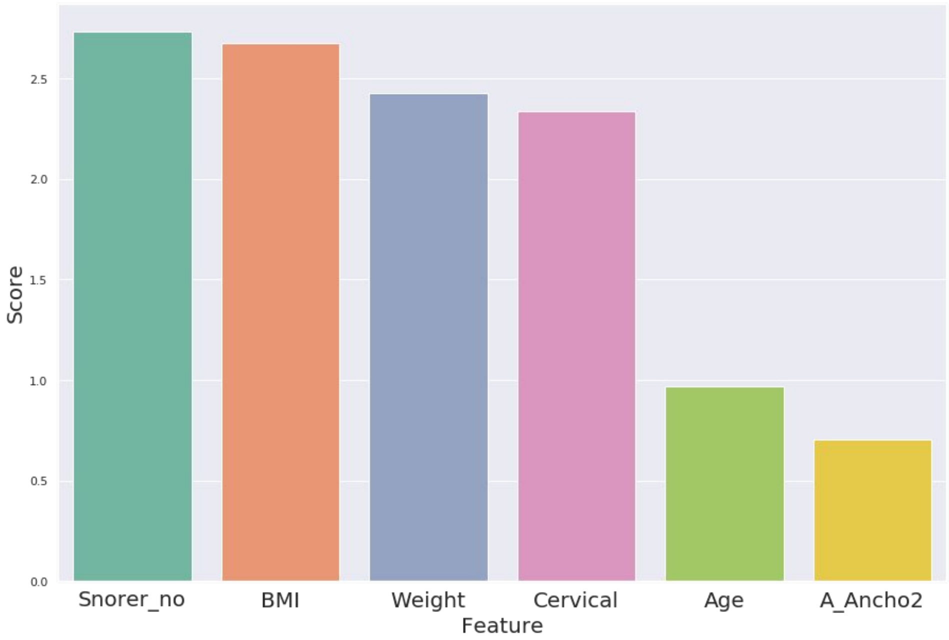
4.4. Feature Selection

Filtering

Pearson Correlation



Univariate Selection



4.4. Feature Selection

Wrapping

Recursive Feature Elimination (RFE)

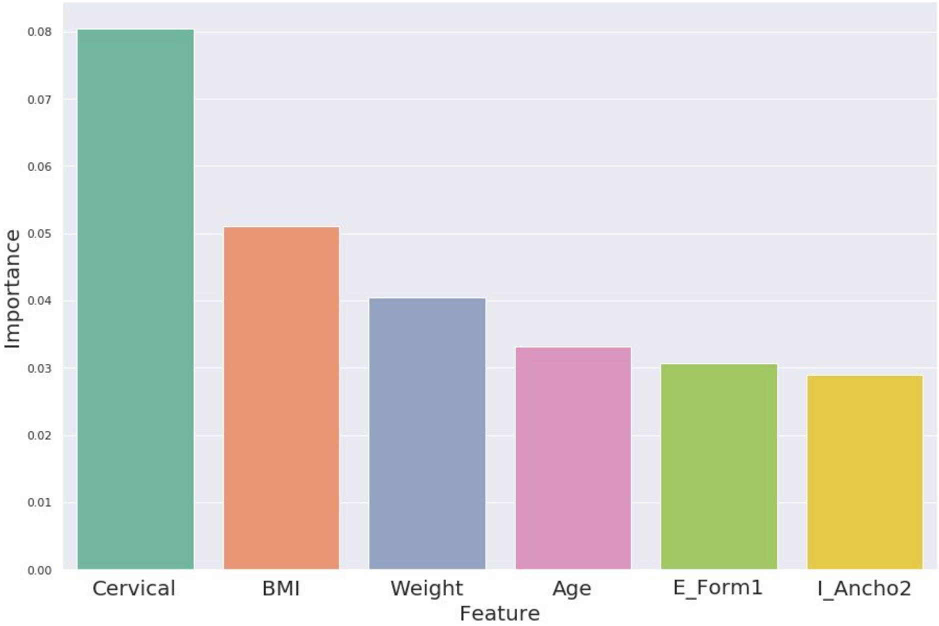
TOP FEATURES	
Importance	**Feature**
Tier 1	A_Ancho2, Age, BMI, Cervical, Weight
Tier 2	Snorer_no
Tier 3	A_Form4, E_Form2...

CV-RFE

4.4. Feature Selection

Embedded

Feature Importance



Select From Model

Top Features

Age

Cervical

Weight

A_Ancho2

Snorer_no

Model Testing
&
Fine-Tuning

5. Machine Learning Modeling

Cross Validation

Stratified 10-Fold

Hyperparameter Tuning

Grid Search

5. Machine Learning Modeling

Evaluation Metrics

Regression

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

Classification

Precision

Recall

f1-score

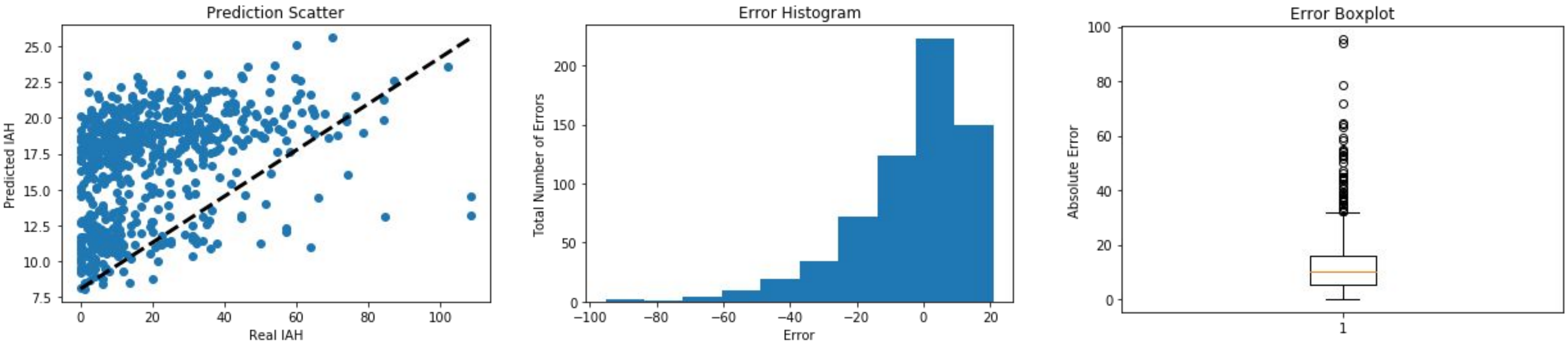
Accuracy

roc-auc score

5. Machine Learning Modeling

Evaluation Metrics

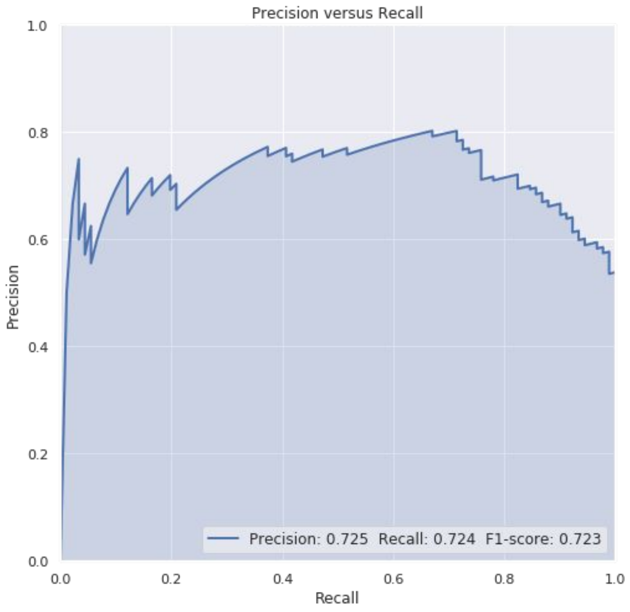
Regression

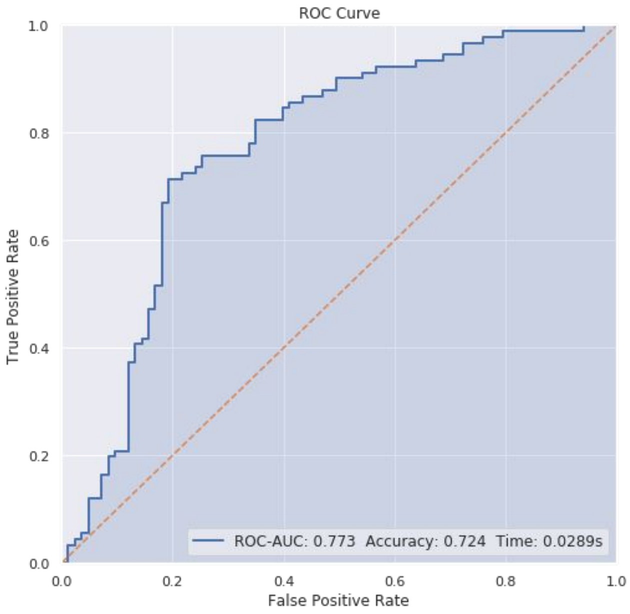


5. Machine Learning Modeling

Evaluation Metrics

Classification





5. Machine Learning Modeling

Modeling - Classical Models

Regression

Model	Hyperparameters
Linear Regression	-
Stochastic Gradient Descent (SGD)	Loss Function, Penalty.

Classification

Model	Hyperparameters
Logistic Regression	C, Penalty, Solver.
Stochastic Gradient Descent (SGD)	Loss Function, Penalty.

5. Machine Learning Modeling

Modeling - Regularizers

Regression

Model	Hyperparameters
Lasso	Alpha
Ridge	Alpha, Solver.
ElasticNet	Alpha, L1-ratio.

Classification

Model	Hyperparameters
Ridge	Alpha, Solver.

5. Machine Learning Modeling

Modeling - K-Nearest Neighbors

Regression

Model	Hyperparameters
Nearest Neighbors	Number of neighbors
Radius Neighbors	Radius

Classification

Model	Hyperparameters
Nearest Neighbors	Number of neighbors

5. Machine Learning Modeling

Modeling - Naive Bayes

Regression

Model	Hyperparameters
-	-

Classification

Model	Hyperparameters
Bernoulli	Alpha
Gaussian	-

5. Machine Learning Modeling

Modeling - Tree-based Models

Regression

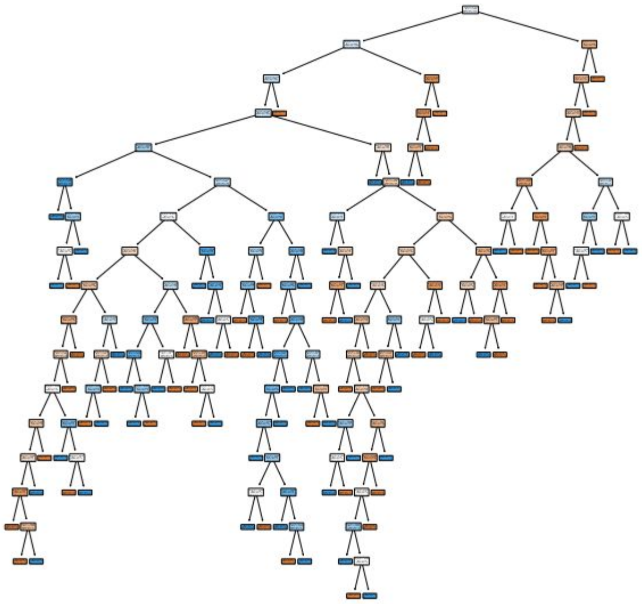
Model	Hyperparameters
Decision Trees Extra Trees	Max depth, max leaf nodes, min samples leaf, min samples split.

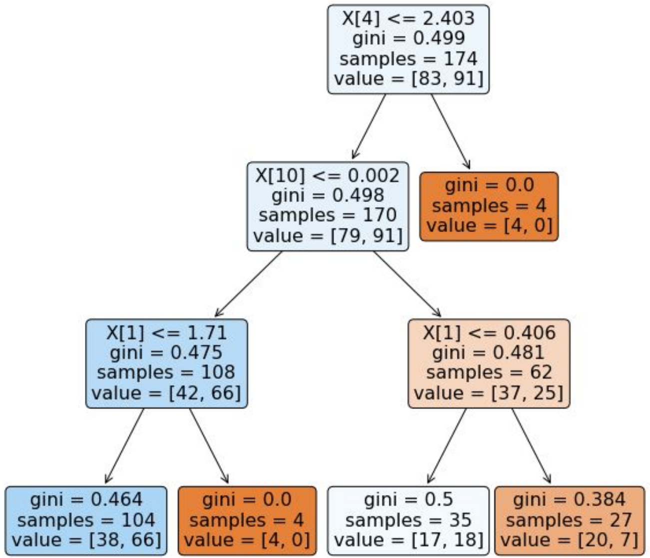
Classification

Model	Hyperparameters
Decision Trees Extra Trees	Max depth, max leaf nodes, min samples leaf, min samples split.

5. Machine Learning Modeling

Modeling - Tree-based Models





5. Machine Learning Modeling

Modeling - Ensemble Models

Regression

Model	Hyperparameters
Bagging **Random Forest** **Adaboost** **GradientBoosting** **XGBoost**	Number of estimators, max samples, learning rate (Adaboost & XGBoost).

Classification

Model	Hyperparameters
Bagging **Random Forest** **Adaboost** **GradientBoosting** **XGBoost**	Number of estimators, max samples, learning rate (Adaboost & XGBoost).

5. Machine Learning Modeling

Modeling - Support Vector Machines

Regression

Model	Hyperparameters
SVR Linear	C, Epsilon.
SVR Nonlinear	Kernel, C, Epsilon.

Classification

Model	Hyperparameters
SVC Linear	C
SVC Nonlinear	Kernel, C.

5. Machine Learning Modeling

Modeling - Neural Networks

Regression

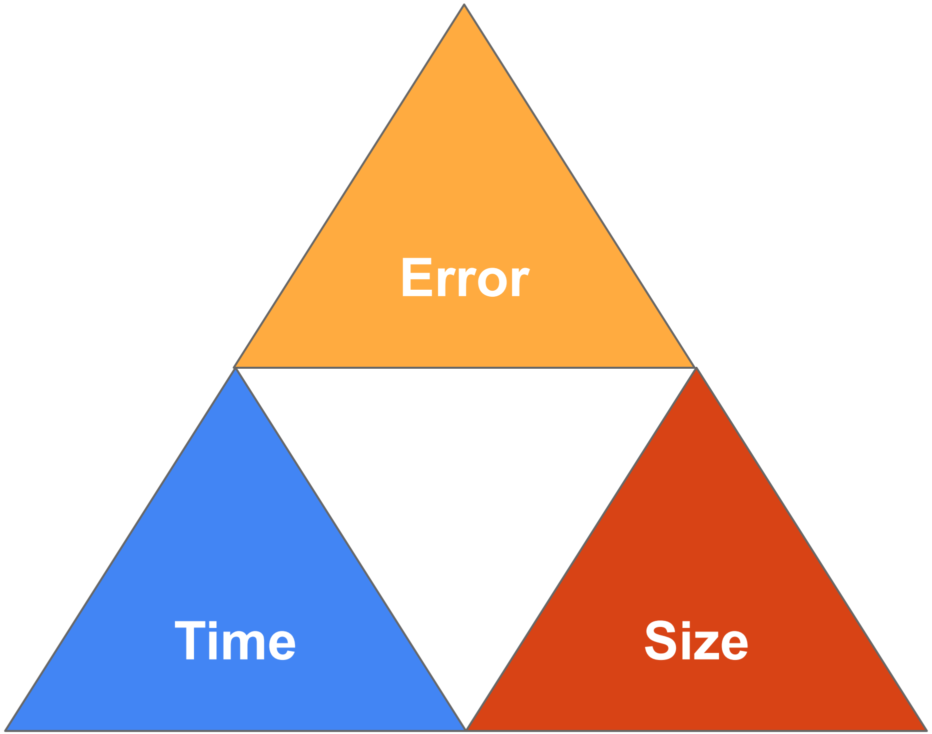
Model	Hyperparameters
MLP	Neurons, layers, activation function, solver function.

Classification

Model	Hyperparameters
Perceptron	Penalty, early stopping.
MLP	Neurons, layers, activation function, solver function.

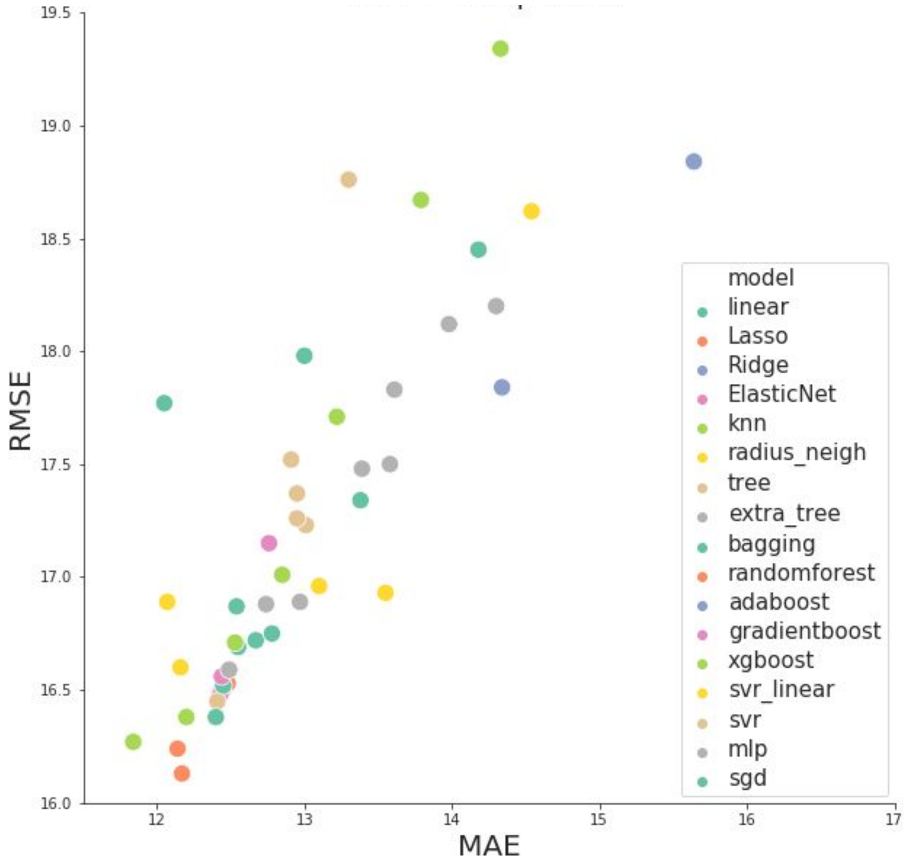
Results
&
Model Comparison

5. Results & Model Comparison



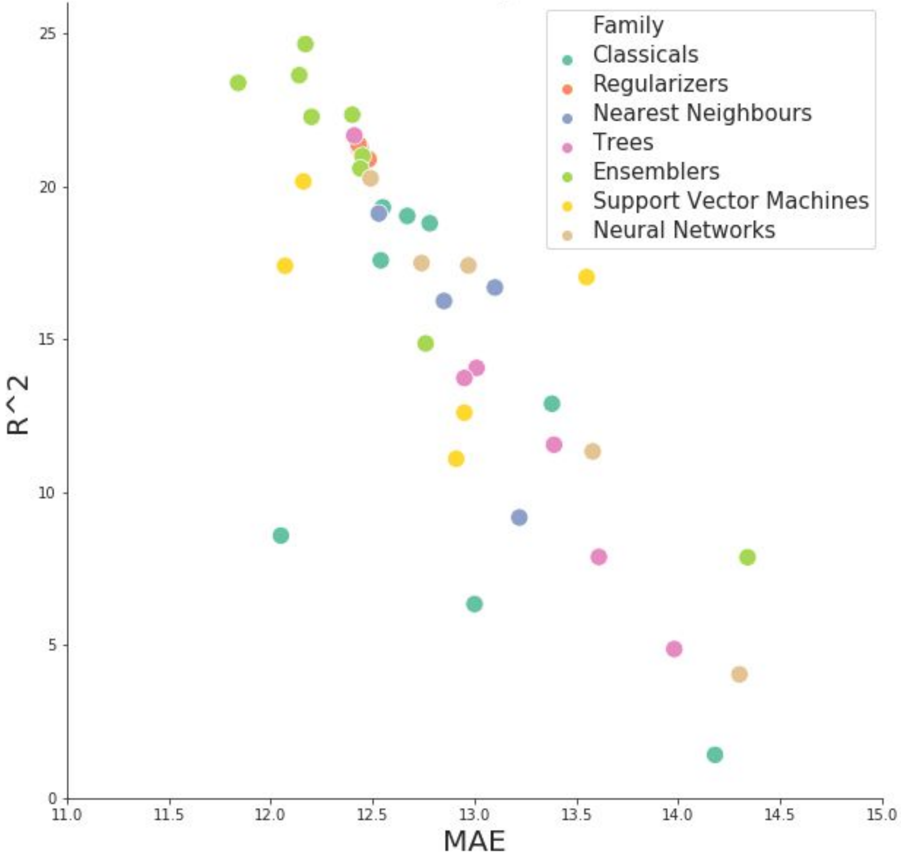
5.1 Regression Results

MAE vs. RMSE



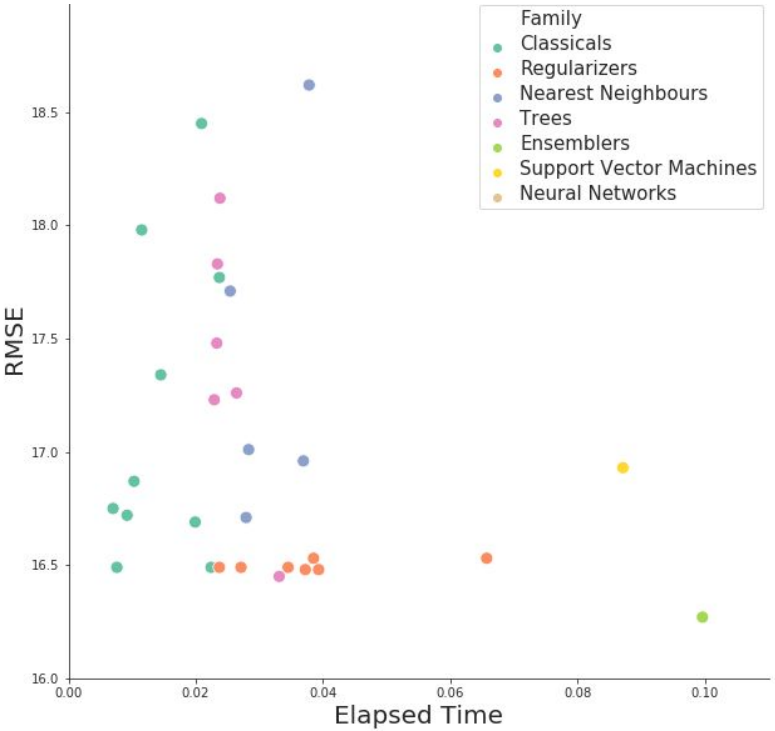
5.1 Regression Results

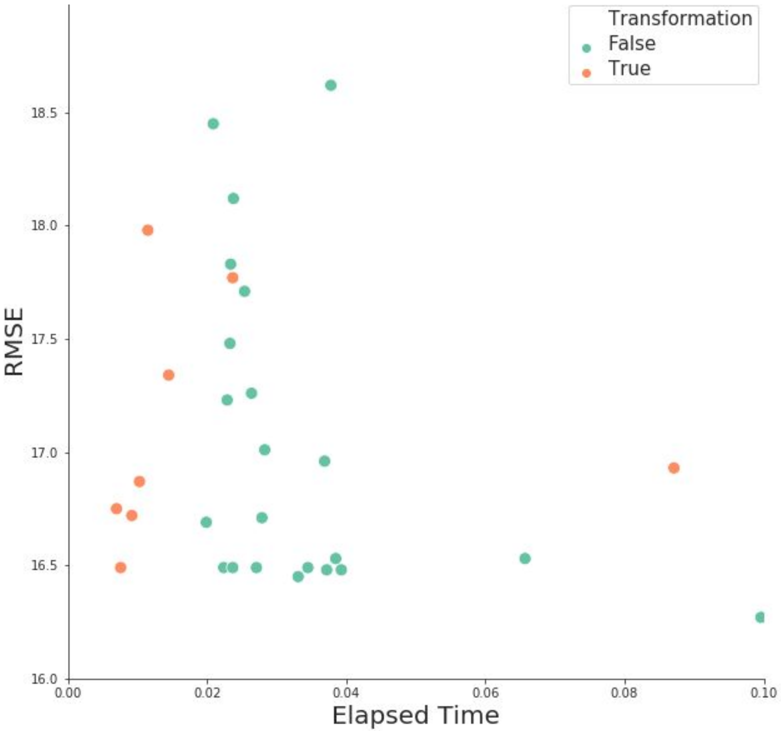
MAE vs. R^2



5.1 Regression Results

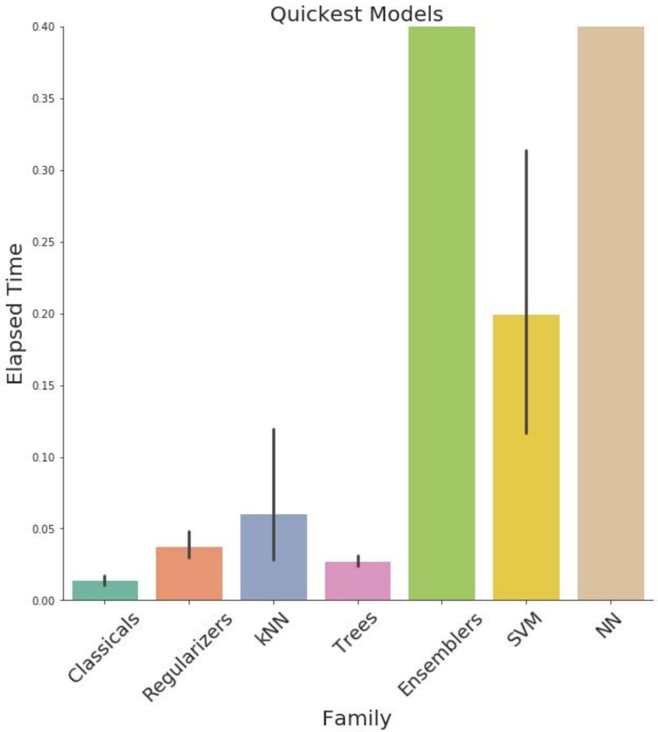
Elapsed Time vs. RMSE

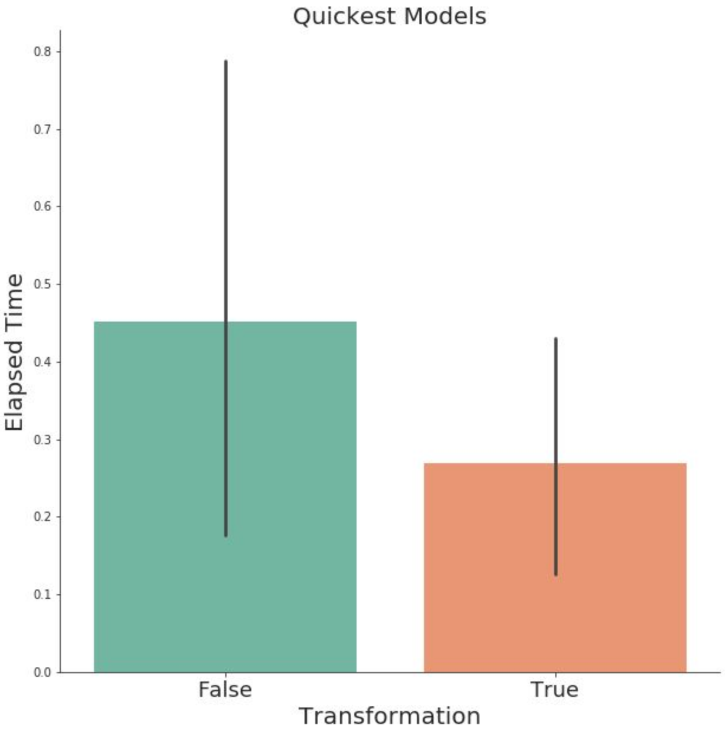




5.1 Regression Results

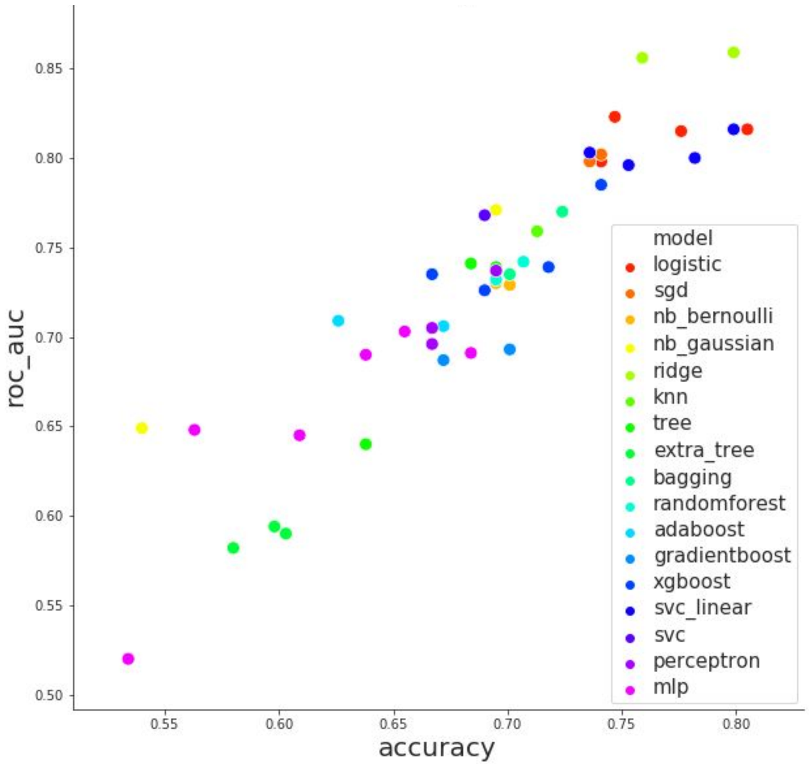
Elapsed Time

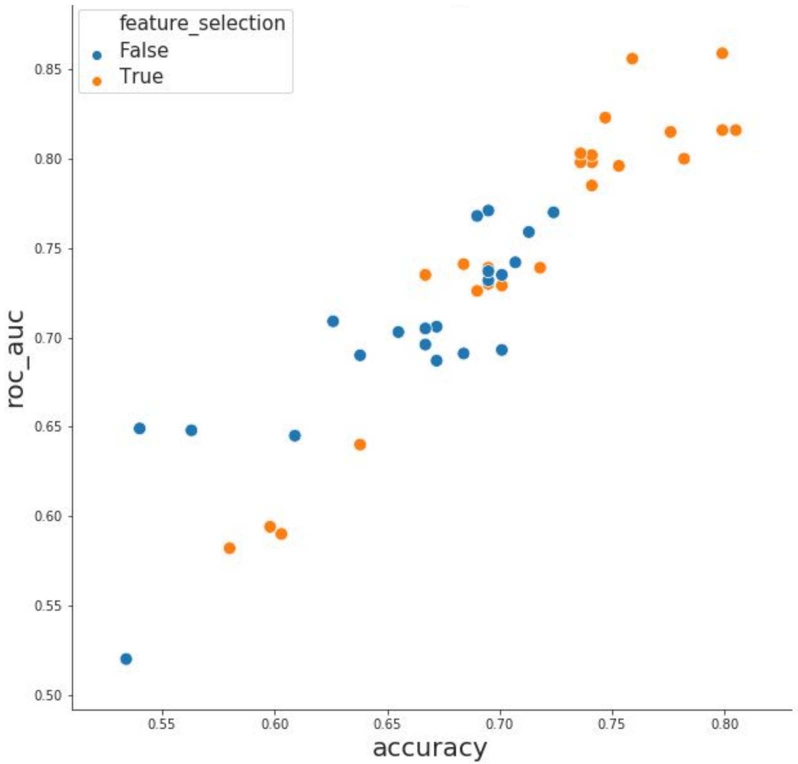




5.2. Classification Results

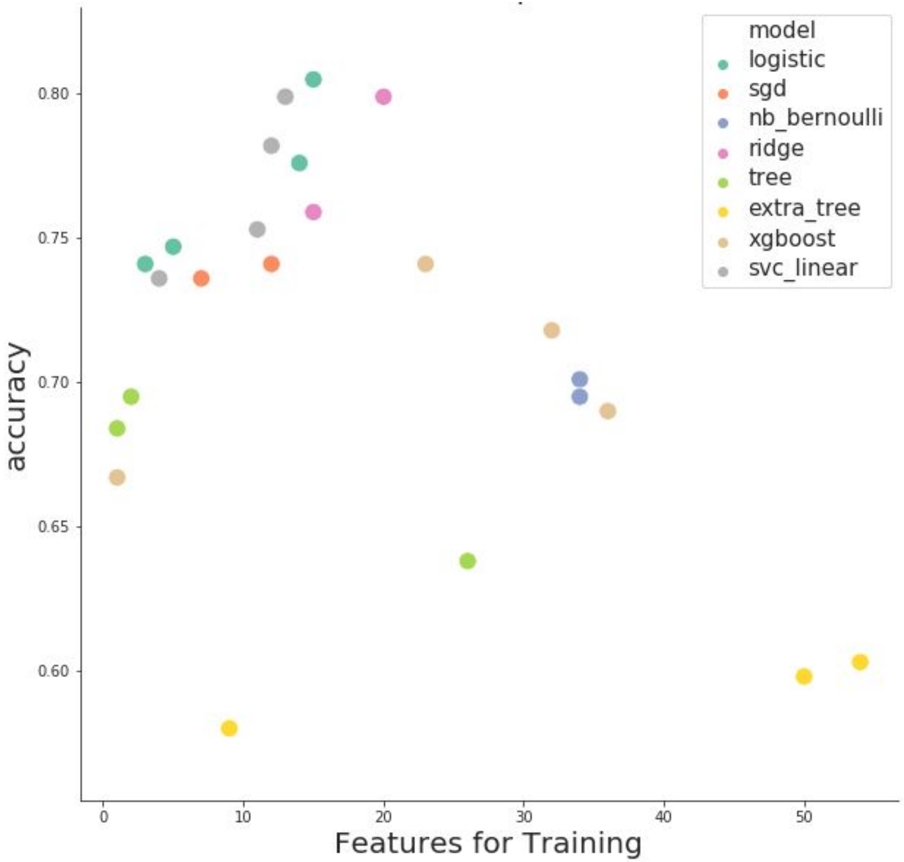
Accuracy vs. ROC-AUC





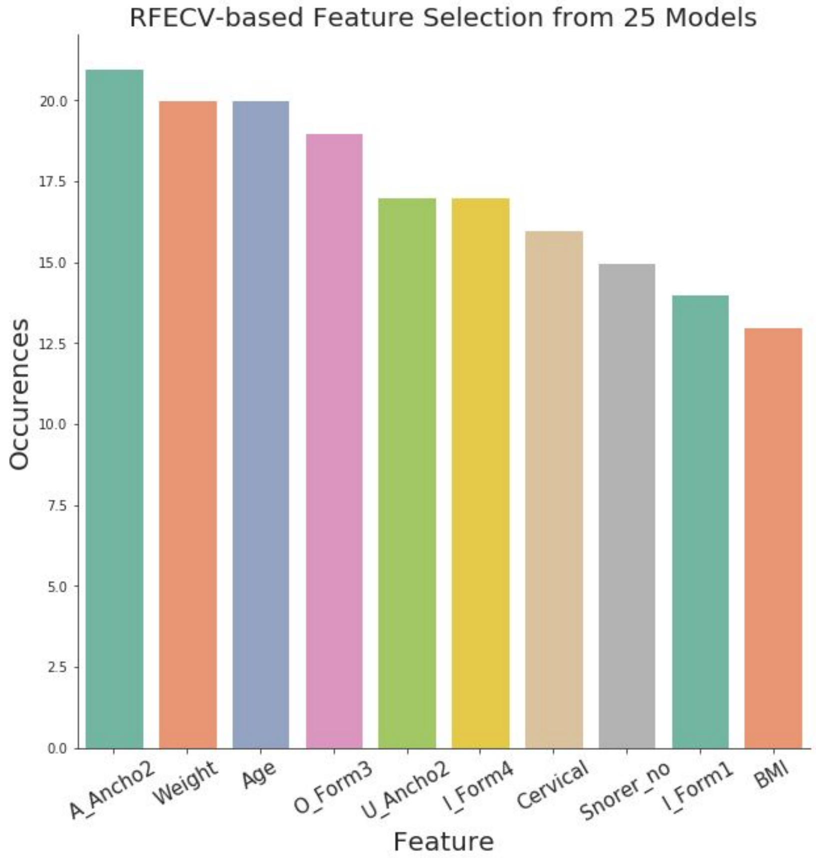
5.2. Classification Results

Accuracy
vs.
Features used
in training



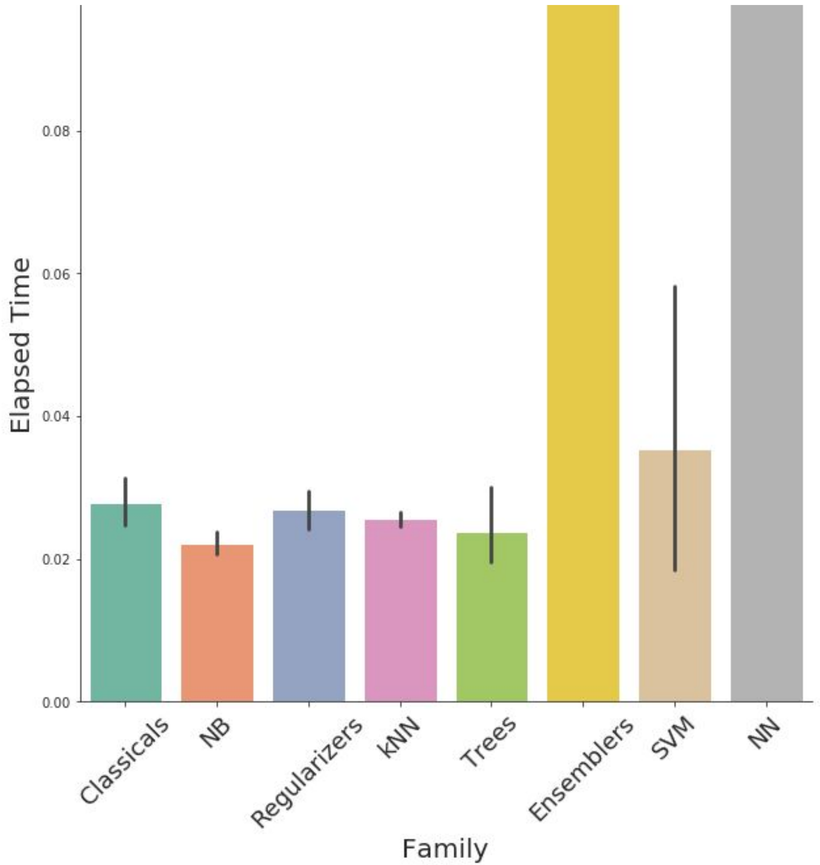
5.2. Classification Results

Most Important Features



5.2. Classification Results

Elapsed Time



Conclusions

6. Conclusions

The data were not sufficient for a real scenario, hence all models are underfitted. More data is necessary.

The most important features were highlighted in several steps of the methodology using different approaches.

Small models were almost as good as large models but much quicker. Feature Selection (CV-RFE) played a big role.

Questions?

You can find me:

linkedin.com/in/spmorillo

sergipm11@gmail.com