Machine Learning for Kawasaki Disease Diagnosis

April 30th, 2018

Problem Definition

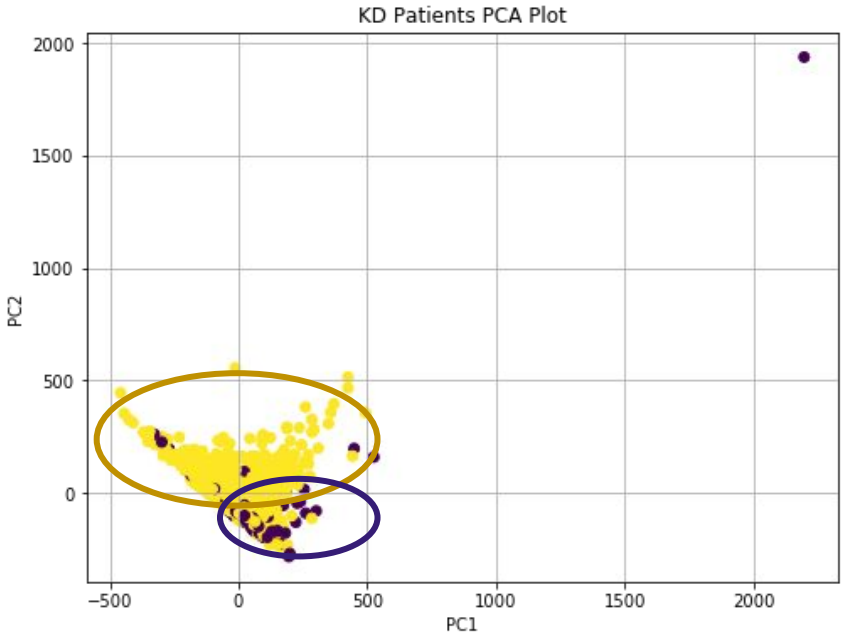
Exploratory Data Analysis

Feature Distributions

Orange: Kawasaki Disease
Blue: Febrile Control

Principal Component Analysis (PCA)

Yellow: Kawasaki Disease
Purple: Febrile Control



Classification Models

Models Evaluated

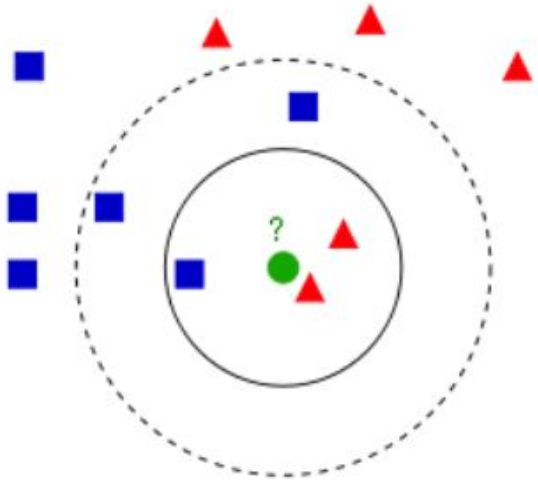
1. K-Nearest Neighbors
2. Logistic Regression
3. Support Vector Machine
4. Random Forest
5. XGBoost
6. Deep Neural Network (Multilayer Perceptron)

Evaluation Methodology

- 5-Fold Cross Validation
- Metrics:
  - True Positives, True Negatives
  - False Positives, **False Negatives**
  - Precision, Recall
  - **F-Beta Score**

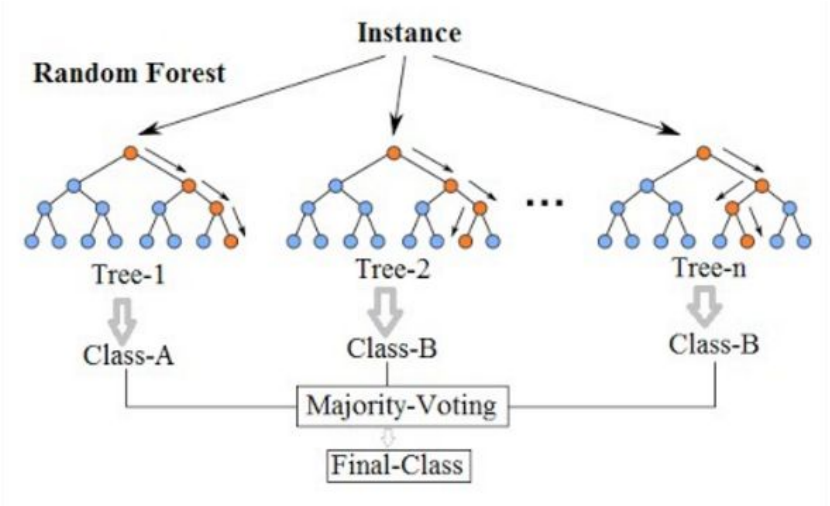
K-Nearest Neighbors (K-NN)

	Pred KD	Pred FC
Actual KD	97.73%	2.27%
Actual FC	36.49%	63.51%



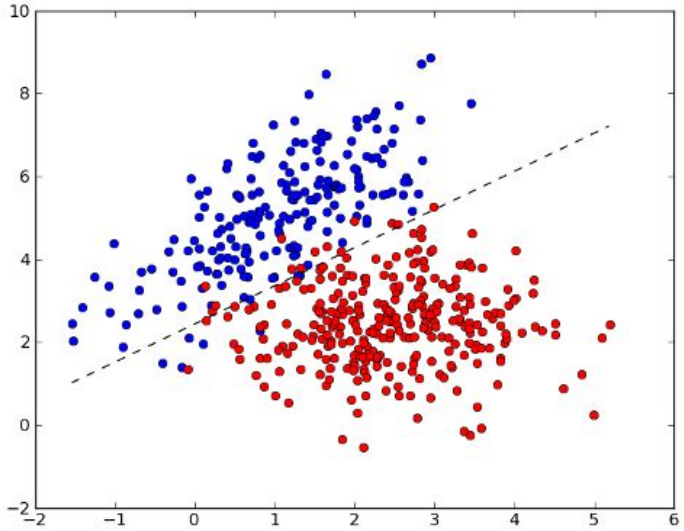
Random Forest

	Pred KD	Pred FC
Actual KD	97.48%	2.52%
Actual FC	27.01%	72.99%



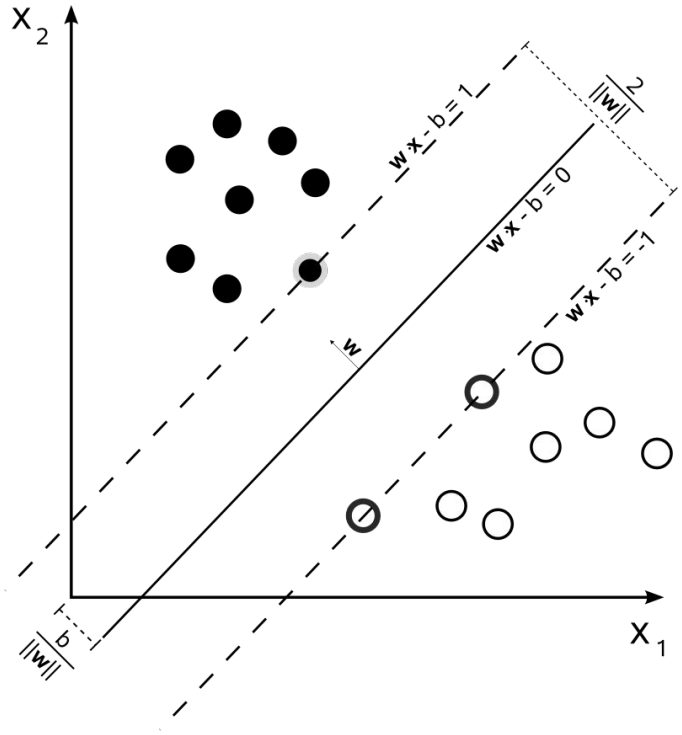
Logistic Regression

	Pred KD	Pred FC
Actual KD	96.85%	3.15%
Actual FC	18.01%	81.99%



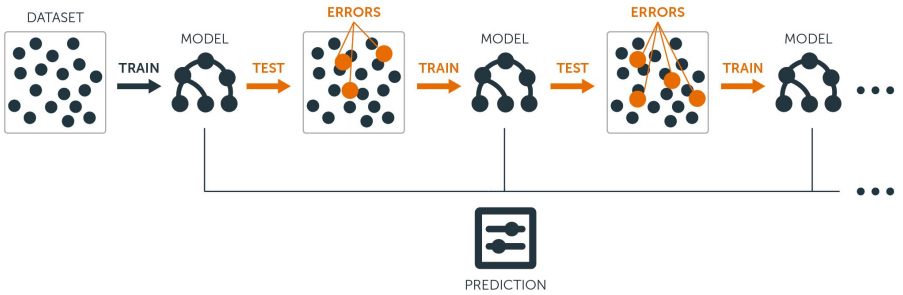
Support Vector Machine (SVM)

	Pred KD	Pred FC
Actual KD	96.35%	3.65%
Actual FC	15.17%	84.83%



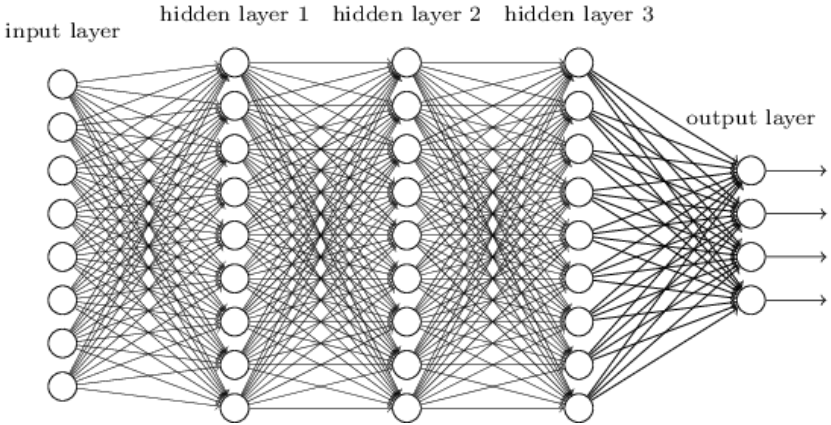
Gradient Boosted Trees (XGBoost)

	Pred KD	Pred FC
Actual KD	97.5%	2.5%
Actual FC	25.1%	74.9%



Deep Neural Network

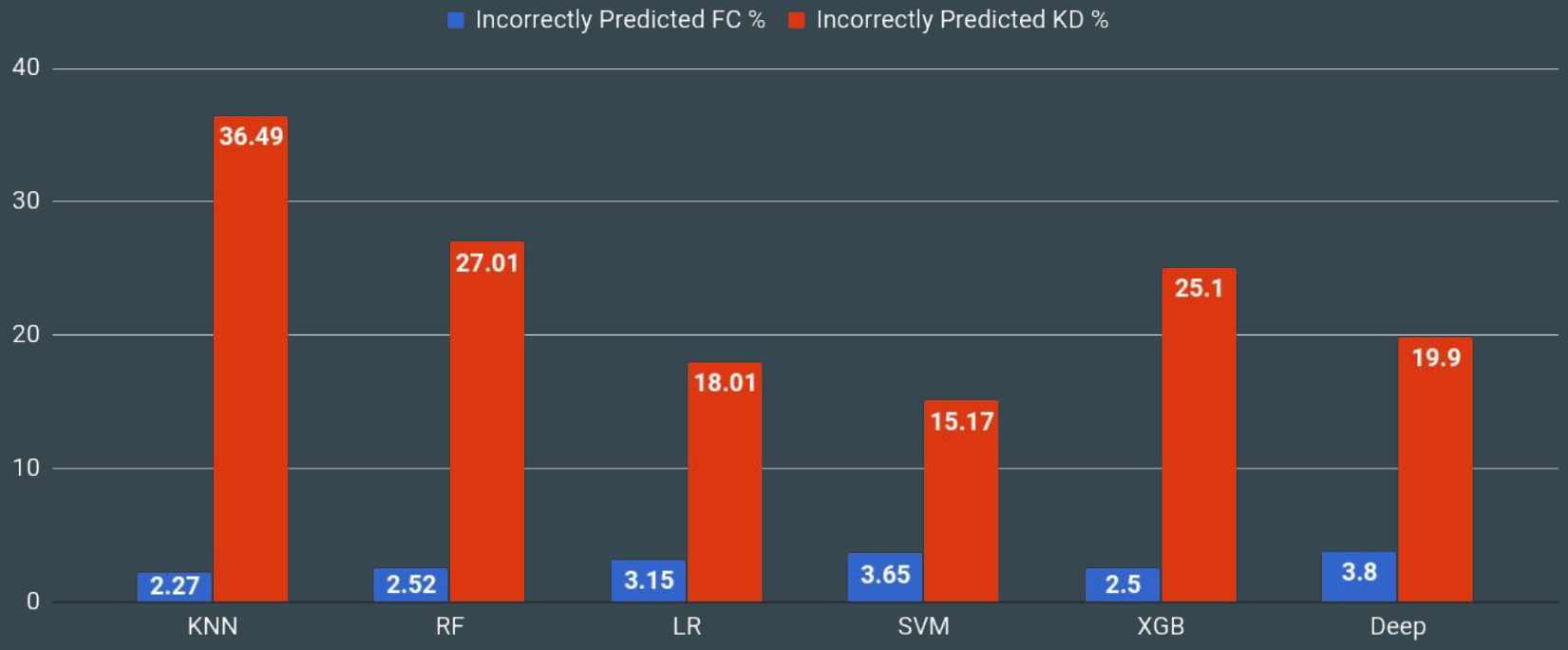
	Pred KD	Pred FC
Actual KD	96.2%	3.8%
Actual FC	19.9%	80.1%



Tunable Model Selection (F-Beta)

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Results Summary



Next Steps

1. Ensemble Methods
2. Data Augmentation
3. Model Interpretation
4. Deployable Diagnosis Application

Thank You!