Springboard Data Science Course

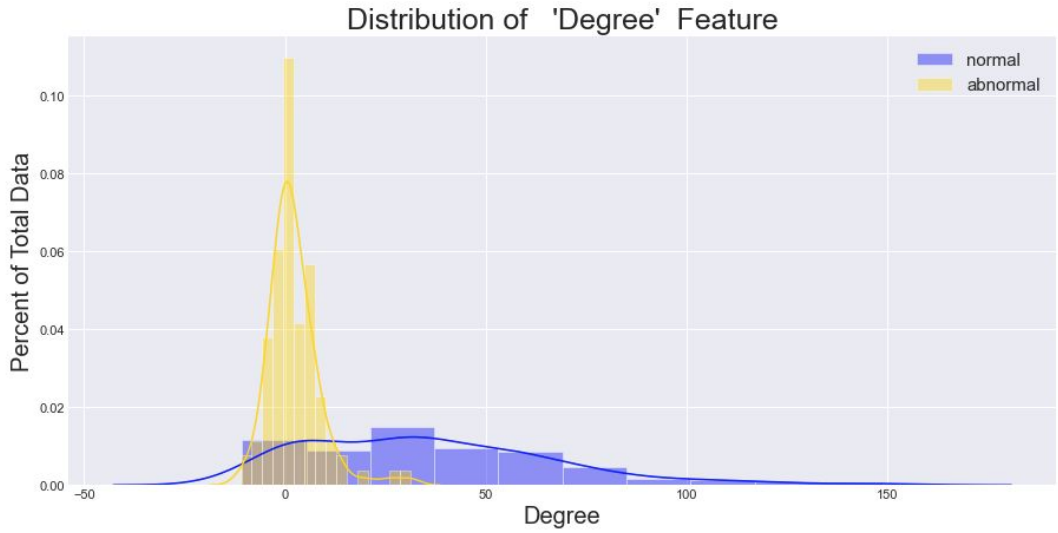
Data Science Capstone Project 1

Orthopedic Biomechanical Features

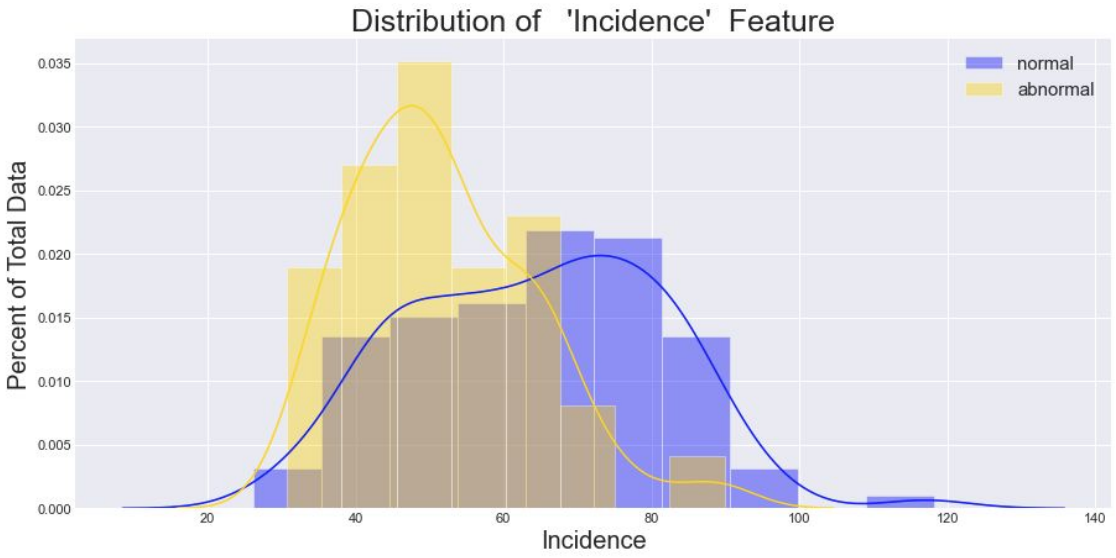
Michelle Ide - 6/3/2020

As we all know, every medical test result determines if we get the proper care and treatment needed. Accurate reports from our individual perspective, requires 100% accuracy. The determination of Abnormal radiological results of the lumbar spine takes several feature angles into account simultaneously.

Below we see the feature data for normal vs abnormal degree results, note the overlap for individual features. Near the zero angle, data for both normal and abnormal has complete overlap of ~2% of data.



But this is the best separation found, most features have much more overlap, for example, the 'incidence' feature below shows >50% of the measured angles are both normal and abnormal.



From the data alone we see the complexity of determining if a patient has an abnormal or normal x-ray result. Proper classification takes years of training and requires digesting multiple pieces of information simultaneously. Validation of results is therefore necessary and conducted by additional physicians.

Machine learning is particularly good at digesting multiple pieces of information simultaneously and can provide the validation in a timely and accurate manner. This project will create the best possible machine classification model that accepts 6 features to predict abnormal vs normal classification of results based on these features. The goal will be to provide a machine model to provide validation support, reducing physician time spent and improved accuracy.

Data

Data for this project was obtained from UCI Machine Learning Repository

Link: http://archive.ics.uci.edu/ml/datasets/Vertebral+Column#.

(Dr. Henrique da Mota during medical residence period in the Group of Applied Research in Orthopaedics (GARA) of the Centre Medicao-Chirurgical de Radaptation des Massues, Lyon, France).

Included is a CSV file containing 310 quantitative biometric records for each unique patient (Features) and a single binomial target (Target)

Features:

1) Pelvic Incidence (310)
2) Pelvic tilt (310)
3) Lumbar lordosis angle (310)
4) Sacral slope (310)
5) Pelvic radius (310)
6) Grade of spondylolisthesis (310)

Target:

1) Abnormal (210)
2) Normal (100)

Approach

- Data Cleansing:
    1. Remove records with empty values - do not impute
    2. Determine and remove full record of outliers
    3. Ensure proper format - encode target variable
    4. Store variables for modeling purposes -
        1) X (features-only)
        2) Y (target-only)
        3) df (full data in original formate
        4) data (full data in encoded format)

- EDA & Statistical Analysis:
    1. over-correlations, >90% - remove a feature with too much correlation
    2. best separations between features - add weight to features with high separation
    3. data distribution curves - Normal? bi-nomial?, guassian?
    4. Statistical distribution - is there a difference that confirms the hypothesis test?

- Modeling:

In search of the best model various classification algorithms will be tested including KNN classification, Logistic Regression, and Discriminant Analysis to provide improved accuracy on reported results.

Deliverables

The project reports, code, and slides will be provided on GitHub to include data, code that takes data, cleans, analyses, and runs through various machine learning models with accuracy scores using cross-validation to determine the best model choice for this problem.