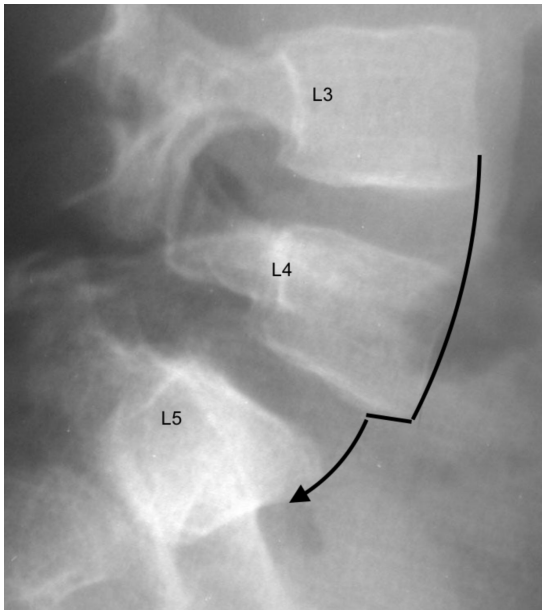
Springboard Data Science Course

Data Science Capstone Project 1

Orthopedic Biomechanical Features

Michelle Ide - 2/1/2021



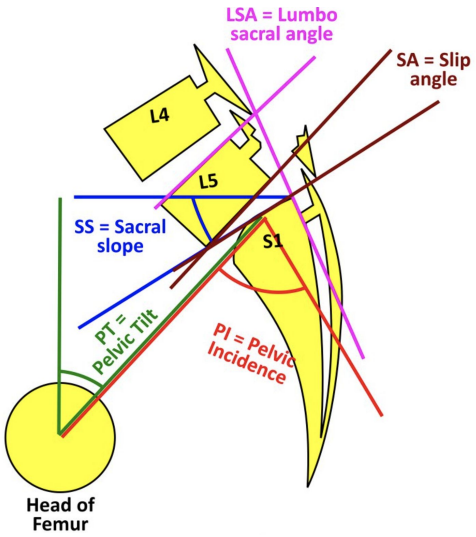
- Muscle spasms in the hamstring.
- Back stiffness.
- Difficulty walking or standing.
- Pain when bending over.
- Numbness or tingling in the foot.

Have you ever had these symptoms? Yes, they are very common, yet these are some of the common symptoms of a serious spinal injury called Spondylolisthesis, which also is referred to as a slipped vertebra.

Not the same as a slipped disc, a slipped vertebra has many possible causes and requires a simultaneous comparison of several angles in order to properly diagnose. These complex comparisons require specialized training for physicians, expertise that is costly and in high demand, reducing the likelihood of early or quick diagnosis. Yet, early treatment can reduce further damage, improve outcomes for patients, and reduce expensive late-stage treatments such as surgery.

Machine learning models are the perfect tool for simultaneous processing and classification, useful in assisting physicians to streamline the process and providing a valuable validation tool for improving accuracy.

The combined contribution of multiple human experiences as input into the models with the simultaneous processing power of a machine would create a powerful hub of combined experience and



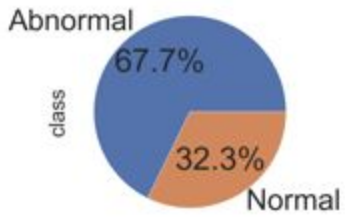
knowledge to be used in a single tool. This project will create and review various models to combine human knowledge with processing power with the goal of improving speed, accuracy, and patient outcomes in identification of spondylolisthesis, vertebral slippage.

Project

This is a supervised classification project using quantitative biomechanical data taken from patient x-rays to predict results as either "Normal" or "Abnormal". In this project abnormal indicates spondylolisthesis (slipped vertebra) of the lumbar spine specifically.

Data Source & Details

Data obtained from UCI Machine Learning Repository at the following 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)



The data is unbalanced, a stratified train-test split will be necessary. Resampling during hyper-parameter tuning is also suggested.

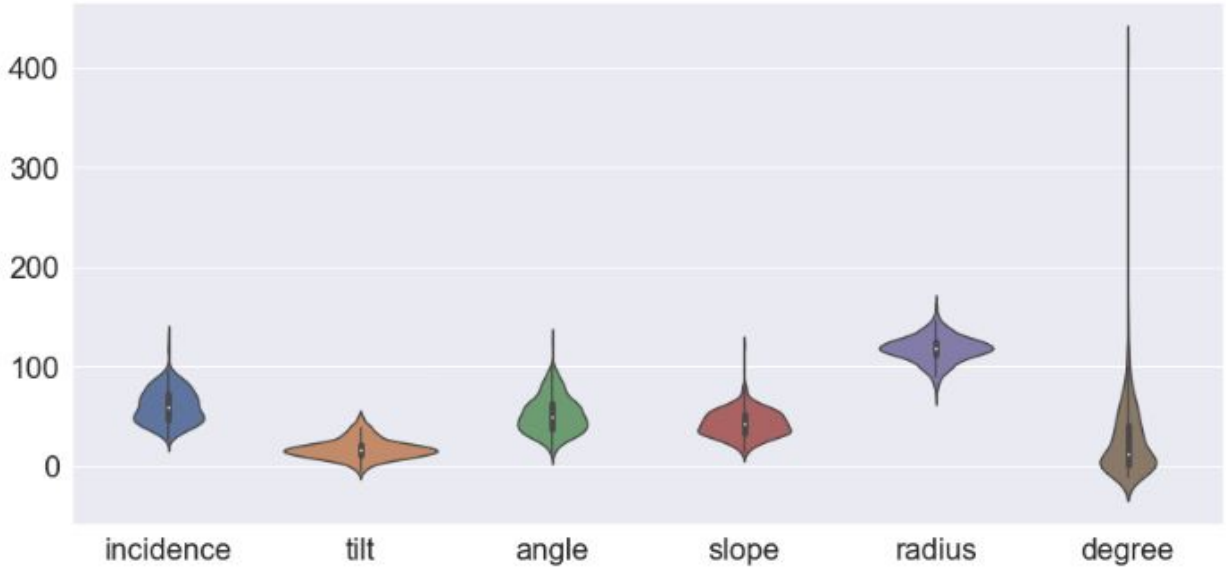
- Total 310 records: 210 Abnormal, 100 Normal
- 1 Target: binomial string: Abnormal/Normal
- 6 Features: Quantitative (integer) below:

Feature	Column (new name)
1) Pelvic Incidence	INCIDENCE
2) Pelvic tilt	TILT
3) Lumbar lordosis angle	ANGLE
4) Sacral slope	SLOPE
5) Pelvic radius	RADIUS
6) Grade of spondylolisthesis	DEGREE

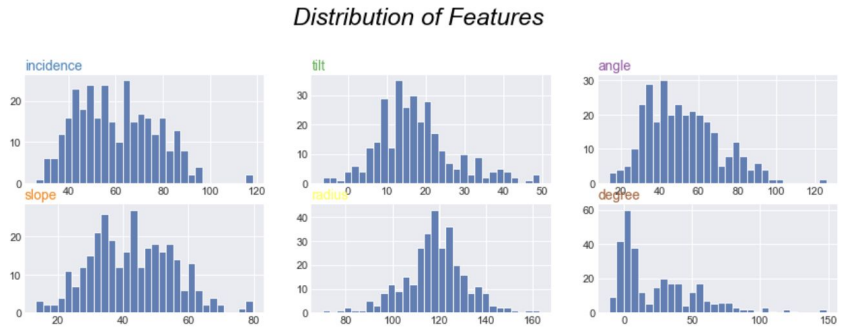
Data Cleaning

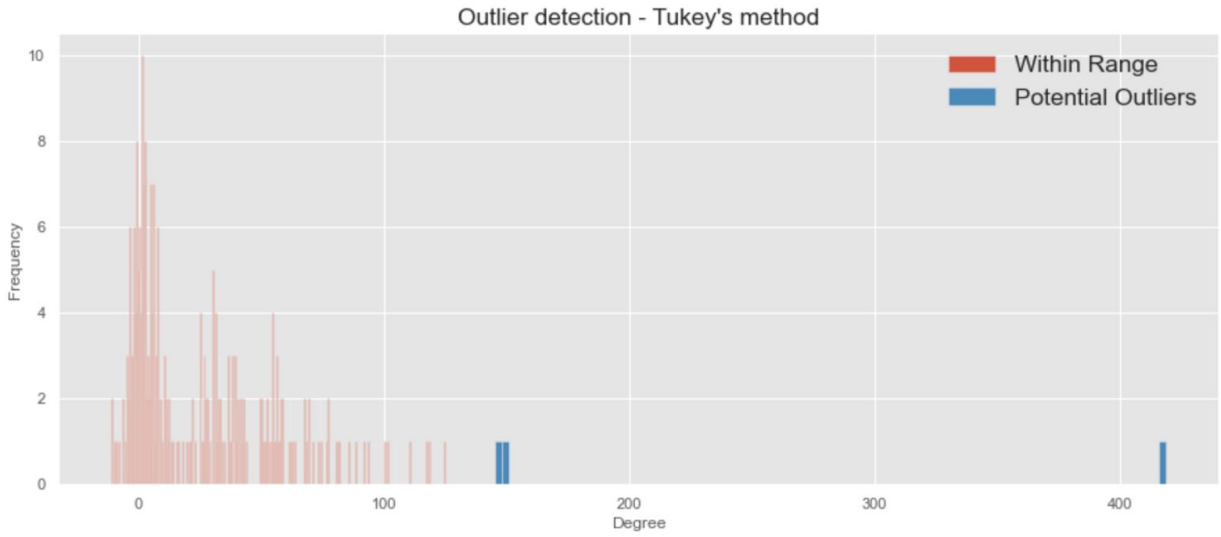
- No empty values were found in this data set.
- No strings or unusual text were identified.
- Non-unique values were found but expected, nothing highly unusual, no changes.
- Target: The string class was encoded for modeling purposes as:
    - 0 = Abnormal,
    - 1 = Normal

The violin plot below shows the mean, standard deviations, IQR and we can see that "degree" has potential outliers.



Looking at the distribution in the plot below, we see normal distributions for all, however, the degree distributions are skewed to the right, possibly the cause of the skewed violin plot.

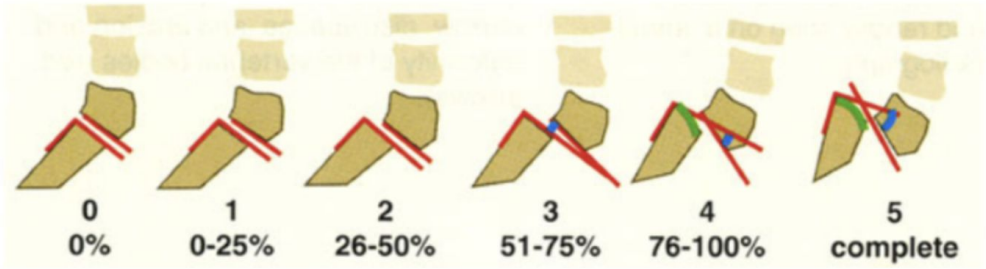




- Outliers:

With the skewed "degree" feature I choose Tukey's (z-score) method to identify potential outliers. With visual inspection of the 3 identified, 1 is greater than 400, this was not only outside the IQR, STD, and z-score, but very far from it's nearest neighbors and was selected for removal from the data. Without more input from the data source or team, the remaining 2 values were close to their nearest neighbors and therefore included for this project.

The image below demonstrates the physical meaning of these degree measures. While infrequent, degree's > 150 are potentially valid measurements, however a measurement > 400, while possible, indicates an unlikely measurement or unusual event.

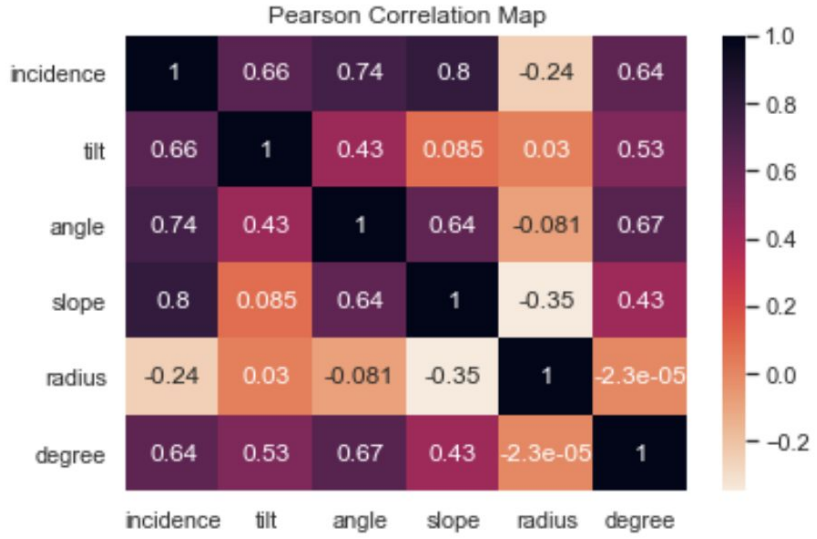


A single record was removed from the dataset and data exploration was performed starting with correlation comparisons.

EDA

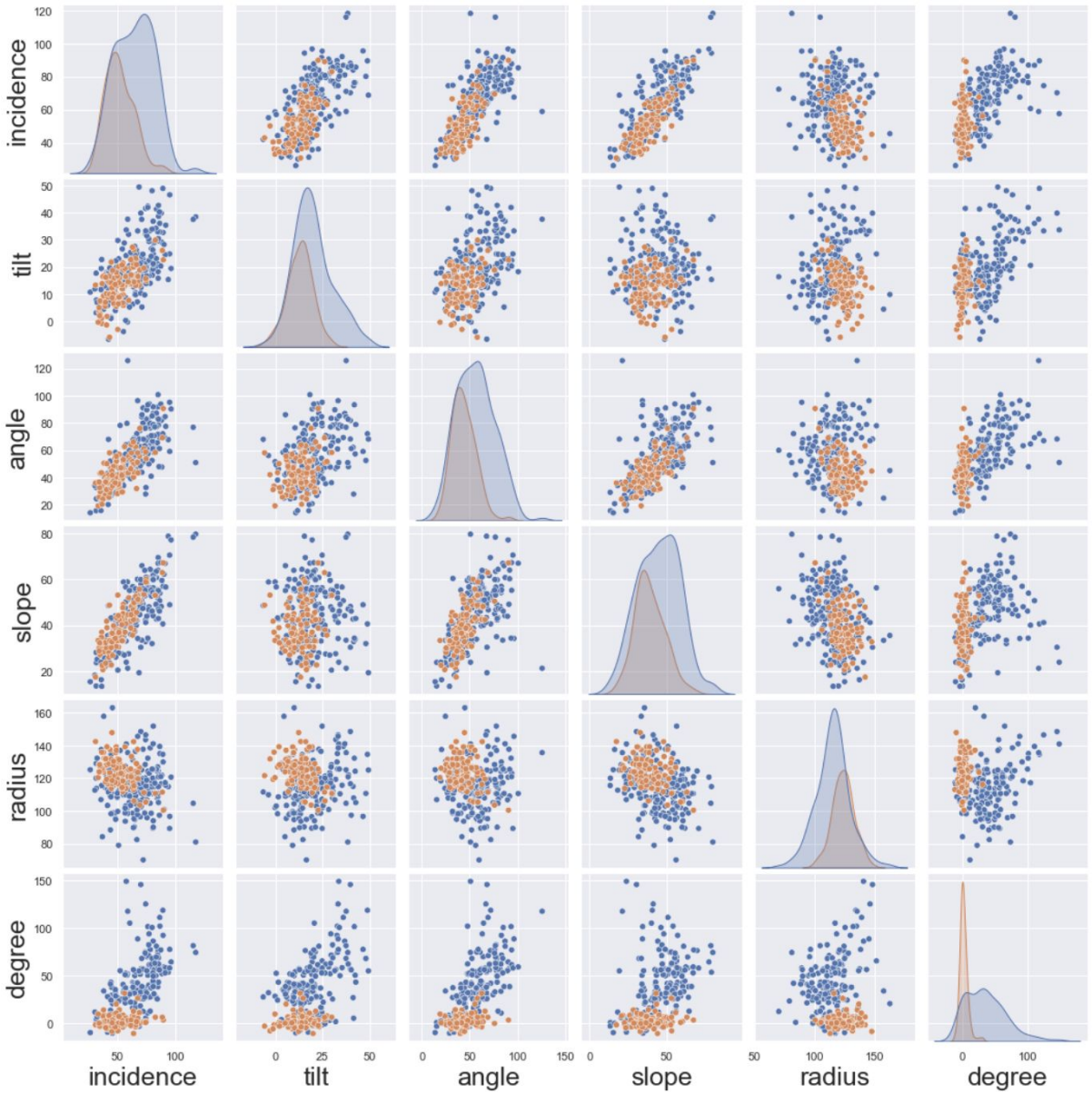
Correlations

A correlation plot, below, shows strong separation in the "degree" feature yet some overlap remains, indicating the need for additional features for classification. All features show some separation but not complete separation. Potential over-correlation between features is explored in the heatmap that follows.



Left, a Pearson's correlation map shows a high correlation between "slope" and "incidence".

All are under the 90% limit and all are included in the model.


incidence
tilt
angle
slope
radius
degree

Hypothesis Testing

The correlation plot above displays a normal distribution of all features. The "degree" feature demonstrate the best separation yet some overlap still exists. Remaining features have clear overlap and a Null hypothesis test was performed to validate the statistical significance of the features relationship with the target.

Alpha was defined as >0.05%

H0 asserts: there exists no statistically significant (>0.05%) difference between "normal' versus "abnormal" within feature data and is not correlated with the target variable.

A student t-test was performed on all features. Results are as follows:

```
The p-value for  degree :   1.0866025811155934e-36
The p-value for  incidence :   1.720325765945183e-12
The p-value for  radius :   1.9625404743220364e-10
The p-value for  slope :   3.327823020643103e-05
The p-value for  tilt :   1.258928899255553e-11
The p-value for  angle :   8.335986160767663e-11
```

All p-values were <0.05%, failing the Null hypothesis proving statistically significant relationships exist for all features. Therefore the null hypothesis is false, all features were proved useful for this project and included in models.

Deliverables

After cleaning there remained 309 records (209 Abnormal, 100 Normal).
The following variables were created and stored in csv files for use in ML models:

- **Data**: 6 Features, 1 Target
- **Y**: 1 Target, encoded binomial
    - 0 (abnormal)'
    - 1 (normal)
- **X:** 6 quantitative Features:
    - tilt, angle, slope, incidence, degree, radius

Machine Learning

Method

- Stratified Train-Test Split, 30% test
- Resampling: SMOTE, ADASYN
- Parameter Tuning: GridSearch with KFold split

Models

- Logistic Regression
- Gradient Boost
- Support Vector Machine
- Random Forest
- KNearest Neighbors
- Naive Bayes

Results

A summary of accuracy scores below compare models using f1 as the scoring method:

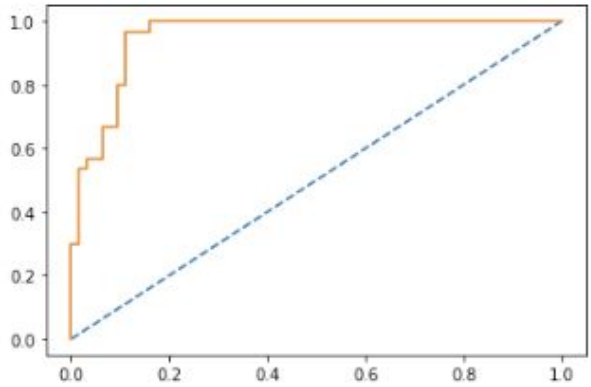
	SMOTE	ADASYN
Model	**Score**	***Model***
logreg	87.88%	85.71%
BGC	80.00%	73.02%
SVM	80.00%	77.61%
RandomForest	81.25%	81.16%
KNN	73.85%	73.85%
Naive Bayes	73.24%	72.22%

Logistic Regression provided the best performance for both resampling methods. With a recall rate of 89 for normal and 97 for abnormal values, using a SMOTE resampling method,

if a total of 100 abnormal results existed, we can expect 97 of these to be properly identified, for 100 normal results 89 would be found. ADASYN's performance was similar with a recall of 100 for abnormal and 84 for normal. Below is a summary of the Logistic Regression scores and confusion matrix.

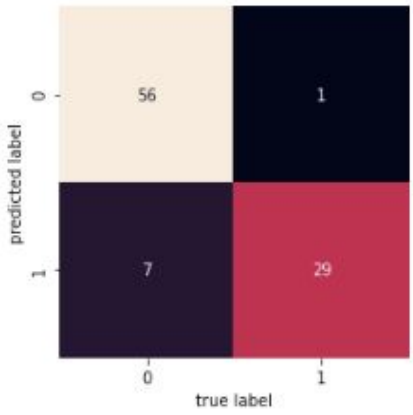
SMOTE

```
area under curve (auc):  0.9523809523809523
```



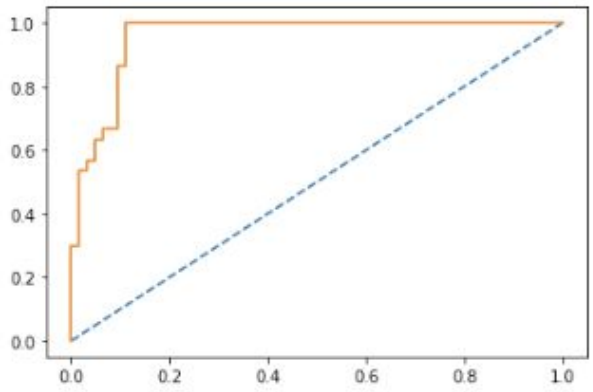
	precision	recall	f1-score	support
0	0.98	0.89	0.93	63
1	0.81	0.97	0.88	30
accuracy			0.91	93
macro avg	0.89	0.93	0.91	93
weighted avg	0.93	0.91	0.92	93

```
Best Parameters:  {'class__C': 10, 'class__penalty': 'l2', 'class__solver': 'newton-cg'}
```



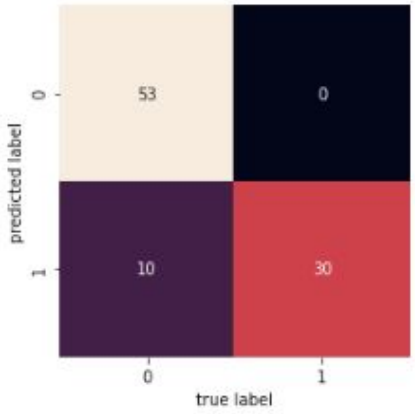
ADASYN

```
area under curve (auc):  0.9560846560846561
```



	precision	recall	f1-score	support
0	1.00	0.84	0.91	63
1	0.75	1.00	0.86	30
accuracy			0.89	93
macro avg	0.88	0.92	0.89	93
weighted avg	0.92	0.89	0.90	93

```
Best Parameters:  {'class__C': 10, 'class__penalty': 'l2', 'class__solver': 'newton-cg'}
```



Detailed results are contained in the addendum at the end of this report.

ADDENDUM

Summary of scores by resampling method:

CONFUSION MATRIX SUMMARY

SMOTE

		ABN	NOR
LR	ABN	56	1
	NOR	7	29
GB	ABN	54	4
	NOR	9	26
SVM	ABN	54	4
	NOR	9	26
RF	ABN	55	4
	NOR	8	26
KNN	ABN	52	6
	NOR	11	24
NB	ABN	48	4
	NOR	15	26

ADASYN

		ABN	NOR
LR	ABN	53	0
	NOR	10	30
GB	ABN	53	7
	NOR	10	23
SVM	ABN	52	4
	NOR	11	26
RF	ABN	52	2
	NOR	11	28
KNN	ABN	52	6
	NOR	11	24
NB	ABN	47	4
	NOR	16	26

CORRECTLY LABELED

SMOTE

	ROC AUC	Recall - Abnormal	Recall - Normal	Precision - Abnormal	Precision - Normal	Weighted Average - Precision
Logistic Regression	95%	89%	97%	98%	81%	93%
Gadient Boosting	96%	86%	87%	93%	74%	87%
SVM	NA	86%	87%	93%	74%	87%
Random Forest	95%	87%	87%	93%	76%	88%
KNeighbors	89%	83%	80%	90%	69%	83%
Gaussian NB	89%	76%	87%	92%	63%	83%

ADASYN

	ROC AUC	Recall - Abnormal	Recall - Normal	Precision - Abnormal	Precision - Normal	Weighted Average - Precision
Logistic Regression	96%	84%	100%	100%	75%	92%
Gadient Boosting	94%	84%	77%	88%	70%	82%
SVM	N/A	93%	70%	83%	87%	86%
Random Forest	95%	83%	93%	96%	72%	88%
KNeighbors	81%	83%	80%	90%	69%	83%
Gaussian NB	86%	75%	87%	92%	62%	82%

Conclusion

While the Logistic Regression model demonstrated great performance in labeling x-ray results correctly, all models performed similarly and clearly demonstrates usefulness of machine learning to assist physicians in validation and determination of appropriate diagnosis.

In order to take advantage of the full possible uses of these technologies I would recommend the use of this model as a validation system initially, continuing as a deep learning project creating a model that learns from all physician labeling of results that can improve with these inputs and develop into a model that can be relied on more and more saving time and money.