Predicting Disk Hernia and Spondylolisthesis using Geometry Measures Derived from Medical Imagery

Leighton Greenstein

2023-03-19

1 Introduction

Machine learning continues to become more prevalent in daily life. For example, despite having a search function, Amazon uses recommender systems to ease the information overload of having millions of products to choose from, so that online shopping is more enjoyable and sales are positively impacted. In addition, traffic predictions are made from GPS navigation position and velocity information, which are loaded into real-time maps to visualize predicted areas of congestion, which can help reduce commute times (Daffodil Software, 2017). Furthermore, machine learning can also bring value to the medical field.

For this project, the Vertebral Column Data Set (Mota et al., 2011) from the University of California's Machine Learning Repository was used to develop a preliminary framework for assessing the potential of machine learning solutions for use in future medical image augmentation systems. This Vertebral Column Data Set contains six anatomical predictors derived from lumbar spine and pelvis relationships that were measured from medical imagery, three associated lumbar spine classification types, and 310 patient records. More specifically, each of the 310 patient records contain measures of Pelvic Incidence (PI), Pelvic tilt (PT), Lumbar Lordosis (LL), Sacral Slope (SS), Pelvic Radius (PR), and Spondylolisthesis Grade (SG). In total, the 310 patients consist of 100 Normal (NO) classes, 60 Disk Hernia (DH) classes, and 150 Spondylolisthesis (SL) classes.

For this supervised machine learning exercise, a number of key steps were performed in pursuit achieving the goal of producing a machine learning solution that could be considered for use in a future machine learning medical image augmentation system. These steps included data acquisition; data preparation; data cleaning; data splitting into a training, test, and verification data set; data exploration and data visualization; a modeling approach; and model fitting which incorporated cross-validation along with parameter tuning for some of the models. After model cross-validation and parameter tuning was complete, the classification solutions were applied to the verification data set, and the results were assessed using metrics derived from the confusion matrix to determine suitability as imaging augmentation information. Overall, a mean score of 0.70 computed from balanced accuracy, F1-Score, and

precision was set as a threshold to be met or exceeded for the machine learning class prediction to achieve the goal of being considered for implementation within a future medical image augmentation system.

2 Methods and Analysis

In general, data science projects require the identification of patterns and trends within data and other important data characteristics, all of which tell a story. Using R, a programming language with origins in statistical computation and graphics (The R Foundation, n.d.), these noteworthy insights are communicated throughout, and ultimately, these insights guided the model selection and development.

2.1 Data Aquisition

Initially, the Vertebral Column Data Set (Mota et al., 2011) was obtained from the University of California's Machine Learning Repository through an automated download and in-memory storage of R objects, which were created from the downloaded data. As the foundation of the data download and storage, a previous script supplied by the course team for the HarvardX Professional Certificate in Data Science (Irizarry, n.d.) was used as vignette and modified as needed to complete the task. Within the downloaded data, there were two data sets: one binary classification data set, and the three class data set described in the introduction with the file name column_3C.dat that was used here. To conform to the recommendations of this project, the column_3C.dat file was loaded to the author's GitHub repository, where the process to obtain the data now is to download the data from the GitHub repository. Should the GitHub repository be unavailable, then the data download reverts back to the University of California's Machine Learning Repository. With the data downloaded and accessible, further data preparation tasks were undertaken.

2.2 Data Preparation

Data preparation is an important first step in any data science project (Wickham, 2014). Upon initial inspection, the Vertebral Column Data Set (Mota et al., 2011) is consistent with the characteristics of Tidy Data, where every row is an observation, every column is a variable, and any observational units are independent tables (Wickham, 2014). The rationale for working with Tidy Data comes from the increased ease with which data modeling, data manipulating, and data visualization can be achieved (Wickham, 2014). Because the data was tidy upon acquisition, little data preparation was required with the exception of providing meaningful column names for the predictors and classes.

The following table displays six randomly sampled rows from the Vertebral Column Data Set:

Table 1: Sample Vertebral Column Data Acquired from UCI Machine Learning Repository

pelvicIncidence	pelvicTilt	lumbarLordosis	sacralSlope	pelvicRadius	spondylolisthesisGrade	diseaseClassification
88.62	29.09	47.56	59.53	121.76	51.81	SL
80.99	36.84	86.96	44.14	141.09	85.87	SL
67.03	13.28	66.15	53.75	100.72	33.99	SL
40.41	-1.33	30.98	41.74	119.34	-6.17	NO
89.68	32.70	83.13	56.98	129.96	92.03	SL
43.19	9.98	28.94	33.22	123.47	1.74	NO

2.3 Data Cleaning

A reliable analysis and repeatable results are difficult to attain if the input data has errors and omissions. For example, missing values in a data set that are zero when zero is not representative of the feature can result in a mean that is biased, and machine learning algorithms may fail to train or produce reliable solutions on data that is missing values.

To ensure the Vertebral Column Data was free from the pitfalls that data error and omissions create, all columns of the data set were tested for NA values. The search for NA data within the Vertebral Column Data Set was achieved by counting the number of NA values present in each predictor and each classification. This NA count information is summarized in the table below:

Table 2: Counts of NA Values Within Vertebral Column Data Set

Column	NA Count
pelvicIncidence	0
pelvicTilt	0
lumbarLordosis	0
sacralSlope	0
pelvicRadius	0
spondylolisthesisGrade	0
diseaseClassification	0

Seeing that the Vertebral Column Data was free of NA values, the next data cleaning step focused on data type verification, and confirming that the anatomical predictor values are representative of geometric measures. The data type for each predictor are summarized in the following table, along with the predictor range expressed as the minimum and maximum values:

Table 3: Vertebral Column Data Set Metadata

Column Name	Class	Minimum	Maximum
pelvicIncidence	numeric	26.15	129.83
pelvicTilt	numeric	-6.55	49.43
lumbarLordosis	numeric	14.00	125.74
sacralSlope	numeric	13.37	121.43
pelvicRadius	numeric	70.08	163.07
spondylolisthesisGrade	numeric	-11.06	418.54
diseaseClassification	character	NA	NA

Because Pelvic Incidence, Pelvic Tilt, Lumbar Lordosis, Sacral Slope, Pelvic Radius, and Spondylolisthesis Grade are measured anatomical values, their numeric class is a valid representation. Although Pelvic Incidence, Pelvic Tilt, Lumbar Lordosis, Sacral Slope, and Pelvic Radius appear to be within reasonable ranges, Spondylolisthesis Grade stood out based on the connection of the name to the Meyerding Classification System (Koslosky & Gendelberg, 2020), which suggested that it may be a class instead of a predictor. Within the Meyerding Classification System, the degree of slip between vertebrae obtained from medical imagery is provided as a percentage (Koslosky & Gendelberg, 2020); more precisely,

> [t]he grade percent is determined by drawing a line through the posterior wall of the superior and inferior vertebral bodies and measuring the translation of the superior vertebral body as a percentage of the distance between the two lines. (Koslosky & Gendelberg, 2020)

However, the minimum value of -11.06 and maximum value of 418.54 affirm that the Meyerding Classification System was not the unit of measure for Spondylolisthesis Grade within the Vertebral Column Data set because of the negative values. For reference, the Meyerding Classification System grades and percentages are shown in the following table:

Table 4: Meyerding Classification Grades and Vertebrae Slip Percentages

Grade	Percentage Range
1	0% to 25%
2	25% to 50%
3	50% t0 75%
4	75% to 100%
5	$>$ 100%

(Koslosky & Gendelberg, 2020)

Even though the physical interpretation of the Spondylolisthesis Grade measure in the Vertebral Column Data Set was unknown, since the Vertebral Column values for Spondylolisthesis Grade were not Meyerding Classifications and the predictor definitions within the University of California's Machine Learning Repository identified Spondylolisthesis Grade as an attribute (not class), they were deemed to be valid for use as predictors.

As the last data check before moving forward with data exploration and data visualization, Le Huec et al. (2011) reveals the relationship between Pelvic Incidence (PI), Pelvic Tilt (PT), and Sacral Slope (SS) as shown in the following equation:

$$PI = PT + SS$$

As described by the equation above, the sum of the Pelvic Tilt and Sacral Slope from the Vertebral Column Data Set should match the Pelvic Incidence from the Vertebral Column Data Set within a reasonable margin of error, considering that Pelvic Tilt and Sacral Slope are measured values and measurements are random variables with uncertainty. The following table displays a random sample of the difference between the computed Pelvic Incidence and the Pelvic Incidence provided within the Vertebral Column Data Set:

Table 5: Pelvic Incidence, Pelvic Tilt, and Sacral Slope Data Verification

pelvicIncidence	pelvicTilt	sacralSlope	PI_Computed	Delta_PI
88.62	29.09	59.53	88.62	0.00
80.99	36.84	44.14	80.98	0.01
67.03	13.28	53.75	67.03	0.00
40.41	-1.33	41.74	40.41	0.00
89.68	32.70	56.98	89.68	0.00
43.19	9.98	33.22	43.20	-0.01

From the six random samples shown in the table above, the differences in the Vertebral Column Data Set Pelvic Incidence and the computed Pelvic Incidence were zero or close to zero. However, the table shows only six of the 310 differences. To further explore the Pelvic Incidence verification results, the following summary statistics of the difference between the Vertebral Column Data Set Pelvic Incidence and the computed Pelvic Incidence (Delta_PI) are presented in the table below:

Table 6: Summary Statistics for Delta Pelvic Incidence

Minimum	FirstQuartile	Median	Mean	ThirdQuartile	Maximum
-0.01	0	0	-0.0002903	0	0.01

With a range of -0.01 to 0.01 and a mean difference between the Vertebral Column Data Set Pelvic Incidence and the computed Pelvic Incidence of -0.0002903, the Vertebral Column Data Set predictors of Pelvic Incidence, Pelvic Tilt, and Sacral slope were deemed to be reliable for machine learning purposes.

Overall, the Vertebral Column Data downloaded from the University of California's Machine Learning Repository did not require additional modification to transform it to a Tidy Data set, nor did it require additional cleaning. With the integrity of Vertebral Column Data Set confirmed, the data was ready to be split into training, test, and validation subsets.

2.4 Data Splitting

In general, very large data sets and very small data sets present their own machine learning challenges. Although large data sets offer the benefit of creating training, test, and verification data where class prevalence disparity can be eliminated, managing memory and having adequate processing power can be difficult to achieve. Conversely, small data sets require little memory management and computing power to implement computationally expensive methods, but bias due to prevalence of the classes within the data set can exist within the split data and be transferred to the machine learning solutions, which can cloud the results. Since the Vertebral Column Data Set is small, with only 310 patient classifications and their associated predictors, key to producing reliable results required a strategic data splitting methodology.

Based on the size of the Vertebral Column Data Set, the Law of Large Numbers and the Central Limit Theorem can help guide the data splitting process. More specifically, the Law

of Large Numbers states that as the number of observations increases, the standard error decreases and the mean of the observations becomes closer to the true mean (Irizzary, 2022, Chapter 14). In addition, the Central Limit Theorem states that large sample sizes result in a normal distribution (Irizarry, 2022, Chapter 14), which is an fundamental requirement of many machine learning methods and a requirement for reliable statistical properties of the predictors. Often, 30 observations will produce data that is normal, and sometimes as little as 10 observations will suffice (Irizarry, 2022, Chapter 14).

To test the application of the Central Limit Theorem and the Law of Large Numbers to the Vertebral Column Data set, a Monte Carlo Simulation could be preformed on each of the training, test, and validation set predictors to reveal the statistical stability of the sample means and standard errors of those means (Irizarry, 2022). Alternatively, histograms and QQ-Plots are visualization tools that can be used to asses the normality of a sample, which is the method that was used to test the training data normality in the following data exploration and data visualization section.

Unfortunately, the data size did not allow for class prevalence correction during data splitting while still maintaining enough data in each split to produce normally distributed predictors. Therefore, 155 rows, one half of each class of the raw data, was split out of the Vertebral Column Data into a training data set and a verification data set. To complete the data splitting, the training set was again split approximately in half to contain 77 to 78 observations of training and test data with approximately the same class structure as the verification data. In addition, based on the knowledge of class prevalence presented in the introduction and the discussion thus far, class prevalence disparity is expected. Therefore, the sample function was selected instead of the caret package's createDataPartition function to split the data into training, test, and verification data sets because the createDataPartition function attempts to create data splits that are somewhat statistically similar (Dalpaiz, 2020, Chapter 21). Yet, the goal is to limit the statistical similarity of the data between the splits while maintaining the same class structure, so that biases have as little transference as possible within the split data. In other words, the sample function is preferable based on the prevalence of classes in the data set because the sampling of each class will be random, which should increase the likelihood that the final algorithm can generalize better to new data and help mitigate the expected bias.

As noted, as many patient records as possible were kept instead of splitting the data into even smaller sets to correct for class prevalence. This approach was necessary to increase the likelihood that the training, test, and verification data would be normally distributed, and therefore, specific machine learning methods that require normally distributed data would not be excluded from use. Fortunately, even the performance of methods with bias in the data where normality has been preserved by using all of the data, such as this case, can still be reasonably assessed using specific metrics that include, balanced accuracy and F1-Score, because those metrics have the ability to account for the class prevalence bias (Olugbenga, 2023).
Consequently, determining the best performing method is possible using the entire data set, while retaining as many machine learning methods to choose from. Still, caution must be exercised given the possibility that class prevalence bias could jeopardize the ability of the

solution to generalize to new data.

To summarize the data splitting technique, all records were kept to provide the best chance for predictors to follow a normal distribution, which helped qualify the split data for use within a variety of machine learning algorithms that require normal data. In addition, the sample function was used over the createDataPartition function to assist in mitigating the class prevalence bias within the Vertebral Column Data Set, yet bias is still expected to persist in the training, test, and verification data to an unknown degree. The following table provides the class counts that were used to construct the training and verification data sets:

Table 7: Training Data and Verification Data Class Counts

diseaseClassification	count	training_counts	verification_counts
DH	60	30	30
NO	100	50	50
SL	150	75	75

With the training and verification data set split determined, the split of the training data into a 50 percent training and 50 percent test set using the foregoing data splitting method was performed. This further division of the training data into a training and test data set will allow for model development, cross-validation, and testing of the cross-validated methods prior to final testing on the verification data set. The following tables present the number of patient records and their class count for the training, test, and verification data sets:

Table 8: Training Data Configuration

diseaseClassification	count
DH	15
NO	25
SL	38

Table 9: Test Data Configuration

diseaseClassification	count
DH	15
NO	25
SL	37

Table 10: Verification Data Configuration

diseaseClassification	count
DH	30
NO	50
SL	75

With the data splitting completed, data exploration and visualization was performed next with the initial goal of testing whether the data splitting method succeeded in producing predictors that follow a normal distribution, and therefore are valid to be used in machine learning methods that require normally distributed predictors.

2.5 Data Exploration and Visualization

Data exploration and data visualization are used throughout this project to assist in constructing, improving, and refining models that predict Vertebral Column Data patient classes. More specifically, the model development follows an iterative analysis and visualization process on the training data set. Ultimately, this process provided the insights that guided the model development and model fitting.

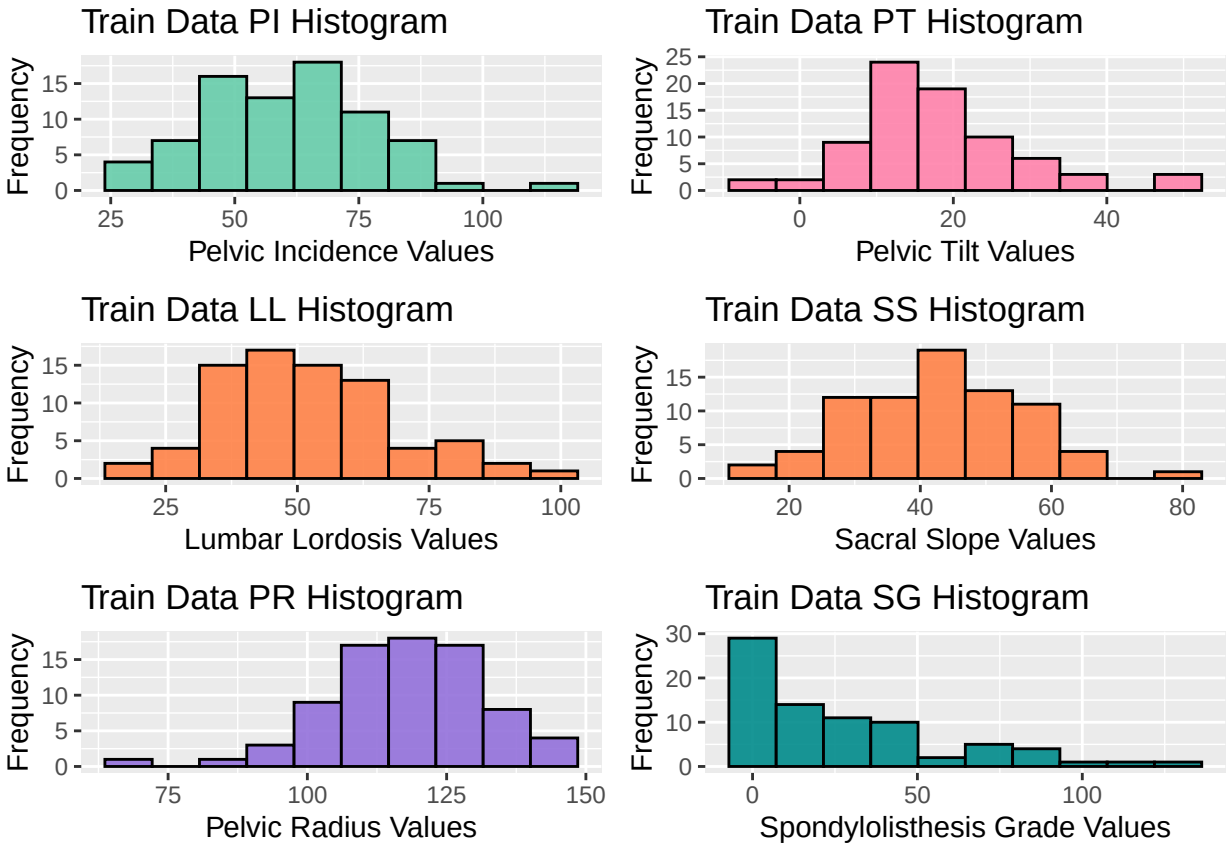
2.5.1 Training Data Predictor Summary Statistics and Histograms

To begin the data exploration and visualization process, summary statistics and histograms were prepared for each predictor in the training data set. A few of the many reasons why summary statistics and histograms are useful in preliminary data analysis is that the distribution, measures of variability, and measures of central tendency (Irizarry, 2022, Chapter 12) can be determined. In turn, these statistics and data visualizations can lead to insights that drive machine learning method selection, greater understanding of the data, and how it can potentially be used. The following table presents summary statistics of the minimum, first quartile, median, mean, third quartile, and maximum values for each predictor within the training data:

Table 11: Summary Statistics of Vertebral Column Data Training Set Predictors

Predictor	Minimum	FirstQuartile	Median	Mean	ThirdQuartile	Maximum
Pelvic Incidence	30.15	48.9050	60.230	60.81910	72.0525	115.92
Pelvic Tilt	-6.55	12.1975	16.710	18.14256	23.1100	48.90
Lumbar Lordosis	15.50	40.0650	51.210	51.66756	61.1525	96.28
Sacral Slope	13.52	33.3925	42.825	42.67679	52.8800	78.41
Pelvic Radius	70.08	110.6575	117.810	117.89603	126.3725	146.47
Spondylolisthesis Grade	-4.08	2.6625	16.080	27.23692	40.1125	124.98

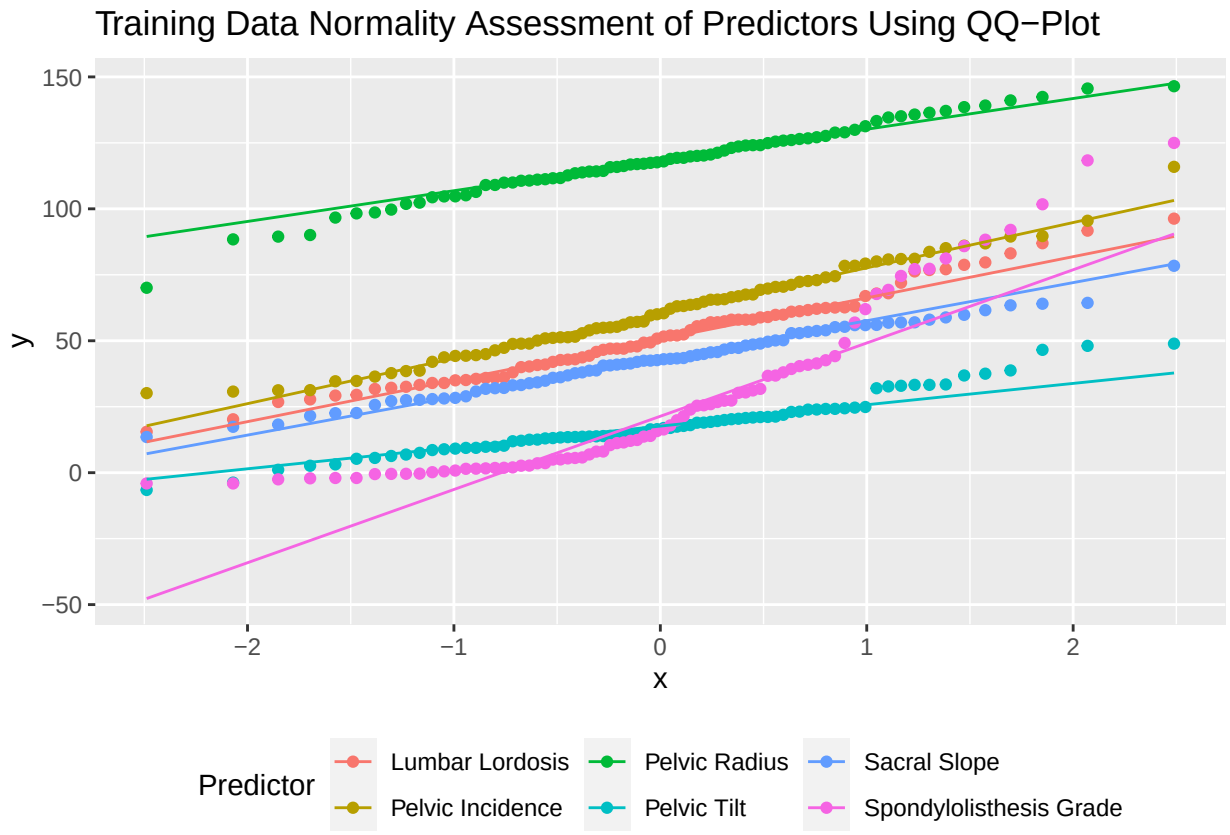
To generate the summary statistics presented in the table above, the base R summary function was used. With the exception of Spondylolisthesis Grade, the mean and median are reasonably close for all predictors, which suggests that all predictors but Spondylolisthesis Grade follow a normal distribution. Although the first and third quartiles are provided, which gives an idea of the where 75 percent of the data for each predictor in the training data lies, greater insight into the distribution of each predictor was achieved by producing histograms of each predictor as displayed in the following plot grid:



The ggplot package was used to generate the histogram objects using geom_histogram with the number of bins set to 10. The number of bins was set to 10 to attain a general idea (not coarse, not fine) of the predictor distribution. Consistent with the mean and median comparison of the predictors made using the summary statistics, the predictors are appear to follow a normal distribution with some shifts in central tendency, but without any substantial skew, except for Spondylolisthesis Grade. However, a QQ-plot would provide a better visual frame to evaluate the degree of predictor normality.

2.5.2 Degree of Predictor Normality

As shown in the predictor histograms derived from the training data, with the exception of the Spondylolisthesis Grade, the predictors, overall, present with characteristics that would fit a normal curve (Gaussian distribution). However, more insight into the suitability of the predictors to be used in models such as Quadratic Discriminant Analysis (QDA), a method that requires predictors to be bivariate normal, could be gained by assessing the degree of normality with at least two independent samples of the data. Since QQ-plots are useful for comparing a set of data to the mathematical normal distribution (Irizarry, 2022, Chapter 12), QQ-plots of each predictor should provide the normality assessment desired. The following QQ-plot shows each predictor compared to the mathematical Normal distribution:



Confirming what the histograms and summary statistics presented, the predictors more or less follow a normal distribution with the exception of Spondylolisthesis Grade. In assessing the degree of normality for the predictors other than Spondylolisthesis Grade, visually, Pelvic Incidence had the largest number of points with the shortest distance to its best fit line, and Pelvic Tilt visually has the greatest number of points with the largest distance to its best fit line. Therefore, Pelvic Incidence shows the greatest degree of normality, Pelvic Tilt shows the least degree of normality, and Pelvic Radius, Lumbar Lordosis, and Sacral Slope fall somewhere in between. Consequently, Pelvic Incidence should have the greatest acceptance from machine learning methods that require normally distributed predictors, and Pelvic Tilt should be the least acceptable.

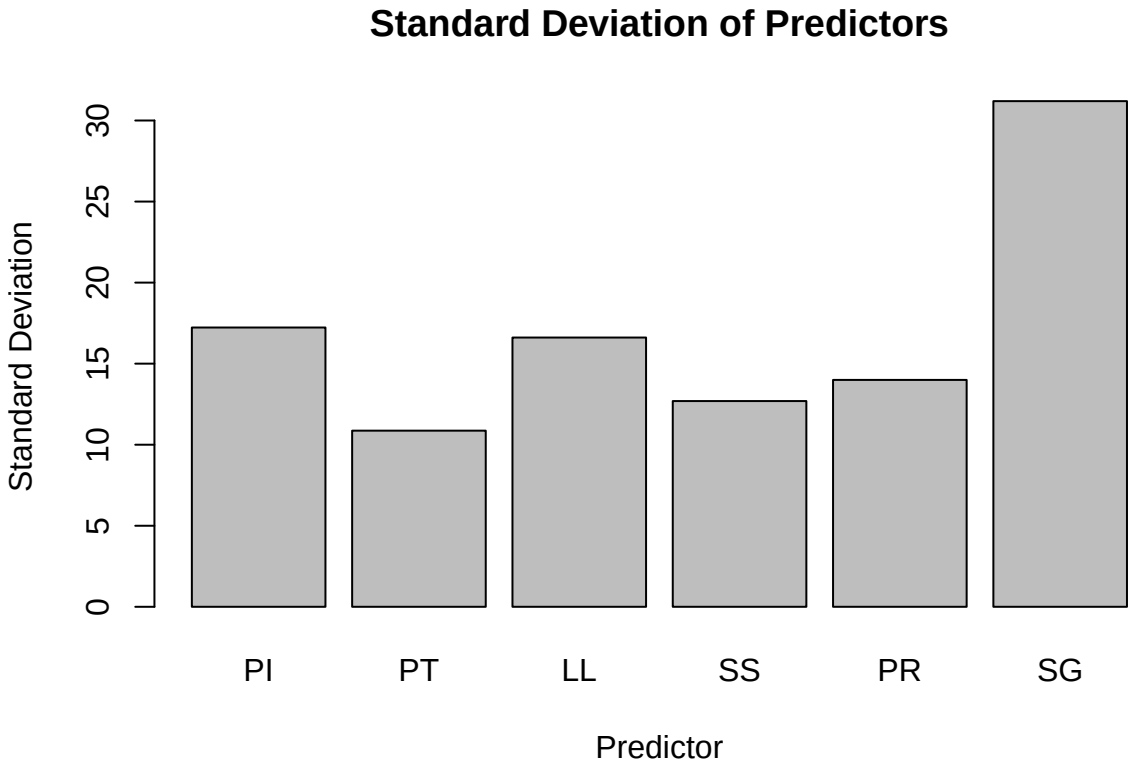
Regardless, of the suitability of a predictor for a particular machine learning method, the utility of the predictors within a machine learning method depends on their variability (Irizarry, n.d). Despite being able to discern some variability from the histograms, the base R summary function that was used to provide summary statistic information did not provide an important measure of variability: standard deviation. Therefore, the variability of the predictors required further exploration because predictors with high variability often have greater importance and predictive power in machine learning, whereas predictors with low variability may provide little information to potentially enhance model performance (Irizarry, n.d).

2.5.3 Predictor Variability

As stated in the Degree of Normality section, more variability in a predictor implies that the predictor may provide more machine learning utility. Therefore, the following table and plot summarizes the variability of each predictor, so that initial interpretations of predictive power can be assessed:

Table 12: Variability of Predictors

predictor	standardDeviation
pelvicIncidence	17.22956
pelvicTilt	10.86634
lumbarLordosis	16.61025
sacralSlope	12.69174
pelvicRadius	13.99701
spondylolisthesisGrade	31.19512



The predictors of Pelvic Incidence, Pelvic Tilt, Lumbar Lordosis, Sacral Slope, and Pelvic Radius have comparable variability. However, as the histogram above shows, Spondylolisthesis Grade on average has nearly twice the variability as all other predictors, which suggests Spondylolisthesis Grade should prove valuable when used in machine learning methods. Unfortunately, the QQ-plot revealed that Spondylolisthesis Grade is not normally distributed, and therefore, only machine learning methods that do not consider predictor distribution or do not require normally distributed predictors are valid to use with Spondylolisthesis Grade.

2.6 Modeling Approach and Model Building

The foregoing data splitting and data exploration and visualization sections identified important characteristics and insights attributed to the Vertebral Column Data set that can be leveraged to guide the modeling approach and model construction. In summary, those insights are as follows:

- Although Spondylolisthesis Grade has the most variability out of all predictors, which from a preliminary perspective suggests that it may be one of the most useful predictors, it can be applied only in machine learning methods that do not consider the predictor distribution or do not require normally distributed predictors, since the QQ-plot revealed Spondylolisthesis Grade is not normally distributed.
- Despite attempting to implement a data splitting strategy that mitigates class prevalence bias in the training, test, and verification data, a balance was struck between mitigating the degree of bias by sampling the Vertebral Column Data Set instead of using createDataPartition, and using all of the patient records to help ensure that the predictors follow a normal distribution, so that machine learning methods were not eliminated from available options due to some method's requirement for normally distributed data.
- Despite the disadvantages of the small size of the data set for producing a prediction platform that would generalize well to new data, with such a small amount of data, baseline results can be skipped. Instead, each model can immediately be cross-validated on the training data and tested on the test data to assess the models ability to generalize to new data. In addition, computationally expensive models were not ruled out when selecting models to train, cross-validate, test, and verify because the time and computing power required should be minimal given the small data set size.

Based on the foregoing insights, characteristics of an ideal model for the Vertebral Column Data Set includes models that do not consider predictor distribution, and models that can accommodate for the class prevalence bias that is presumed to be inherent within the Vertebral Column Data Set and transferred to the training, test, and validation data sets.

Fortunately, the caret package contains a large number of supervised classification machine learning methods to choose from. In addition, the caret package makes working with the packages easy due to standardized function calls (Kuhn, 2019). Overall, the models selected to predicted patient classifications were determined from exposure to models from Irizarry's machine learning course (n.d.), the documentation provided in the caret repository (Kuhn, 2019), and the insights obtained from the data splitting and data exploration and data visualization sections. Based on this information, Quadratic Discriminant Analysis (QDA), Random Forest, and K-Nearest Neighbor (KNN) methods were chosen to generate cross-validated predictions of patient class while using suitable predictors for each method.

2.6.1 Quadratic Discriminant Analysis

Before starting the prediction process with QDA, which is a version of Naive Bayes where the predictors are assumed to be multivariate normal (Irizarry, 2022, Chapter 31), the data splitting was performed multiple times by re-running the computation script without setting a seed value, so the split data was randomly sampled. Although the results of script re-runs are not shown here, by randomly sampling the Vertebral Column data multiple times to generate variations of the training data and analyzing the summary statistics, histograms, and QQ-plots for the predictors, the predictors with the exception of Spondylolisthesis Grade were determined to be multivariate normal. In addition, seeing that class prevalence bias is expected within the training data due to unequal distribution of patient classifications, QDA was an attractive method because Naive Bayes controls for prevalence (Irizarry, 2022, Chapter 31). The table and histogram below provide a reminder of the class prevalence within the training data set:

Table 13: Training Data Class Prevalence

diseaseClassification	count
DH	15
NO	25
SL	38

Considering the small size of the training data set, the potential to over fit the model was also assessed. The following formula was used to determine the estimated number of parameters required for the QDA solution, which was used to understand the over fitting risk:

$$Parameters = K\left[\frac{2p + p(p-1)}{2}\right]$$

where,

$$K = the\ number\ of\ classes$$
$$p = the\ number\ of\ predictors$$

(Irizarry, n.d.)

To assess the prospect of over fitting, the closer the number of parameters is to the size of the data set, the higher the risk of over fitting, which is also know as the curse of dimensionality; as the number of parameters increases and approaches the number of data set records, the utility of the method declines (Irizarry, 2022, Chapter 31).

For the Vertebral Column Data Set, the following table shows the estimated number of computed parameters required for QDA to operate with three classes of DH, SL, and NO, and using a range of predictors sequentially added to the model, which include, Pelvic Incidence, Pelvic Tilt, Lumbar Lordosis, Sacral Slope, and Pelvic Radius:

Table 14: Estimated QDA Parameter Requirements

predictorsUsed	classesK	parametersRequired
1	3	3
2	3	9
3	3	18
4	3	30
5	3	45

From the table above, if all predictors represented by bivariate normal data are used in the QDA model, an estimated 45 parameters are required. Since there are 78 observations, there is still plenty of freedom left within the training data set, so over fitting was not expected be a problem. Therefore, QDA was applied to the training data using 5-fold cross-validation and the following five bivariate normal predictors: Pelvic Incidence, Pelvic Tilt, Lumbar Lordosis, Sacral Slope, and Pelvic Radius.

In the context of patient classification, the accuracy of 0.675 achieved by QDA left room for improvement. Perhaps, because Pelvic Incidence is a function of Pelvic Tilt and Sacral Slope, as noted in the data cleaning section, where $PI = PT + SS$, duplicate information may not be providing a performance boost, or those predictors simply could have very little impact to the accuracy of the QDA solution. To explore whether Pelvic Incidence, Pelvic Tilt and Sacral Slope had importance in patient class predictions (along with the other predictors in the Vertebral Column Data Set), the variable importance feature of the Random Forest method presented the opportunity to quantify predictor classification value.

2.6.2 Random Forest

Besides providing insight regarding the importance of predictors and being a method that is an extension of decision trees that elegantly escapes the curse of dimensionality that QDA is subject to (Irizarry, 2022, Chapter 31), the decision trees associated with Random Forests are also useful in medical diagnostics. According to Podgorelec et al. (2002), decision trees can be reliable, effective, and accurate in medical decision making. With the premise of escaping the curse of dimensionality and the relevance of decision trees in medical decision making, Random Forest was implemented using all of the predictors available and five-fold cross-validation to provide an accuracy comparative to QDA, and to gain insight into the predictors that provide the most and least utility.

The Random Forest method improved the accuracy of the model from 0.675 to 0.805, and as expected, provided insight into the importance of the predictors. A summary of the variable importance generated from the Random Forest method is shown in the bar plot and table below:

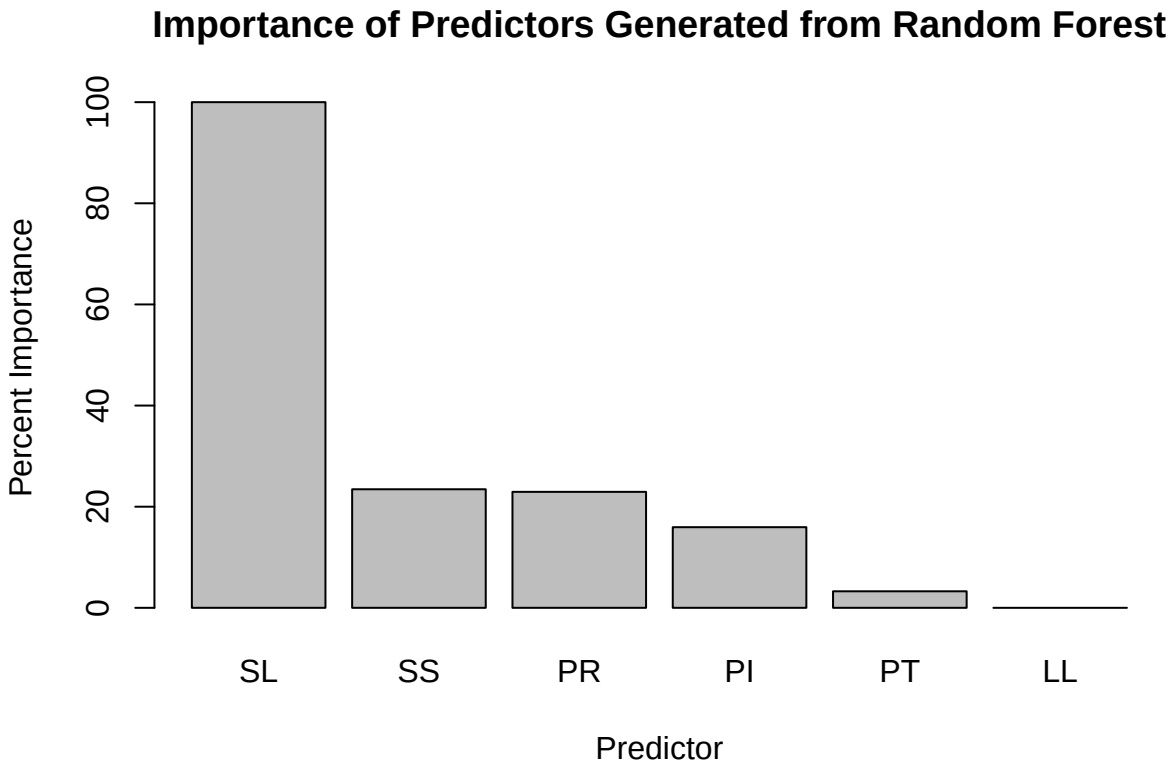


Table 15: Random Forest Variable Importance

Predictor	Overall
spondylolisthesisGrade	100.000000
sacralSlope	23.445413
pelvicRadius	22.932631
pelvicIncidence	15.949053
pelvicTilt	3.270546
lumbarLordosis	0.000000

Not surprising, Spondylolisthesis Grade presents with overwhelming importance compared to the other predictors. However, the substantial importance may simply be due to the class prevalence bias within the training data set, or it may be legitimate since Spondylolisthesis Grade had, on average, twice the variability of any other predictor, which implies its ability to provide useful classification information. To confirm, the importance of Spondylolisthesis Grade could be verified as over-valued if data sets with different class prevalence were used to compare variable importance results.

Nevertheless, Random Forest and the standard deviation comparison in the Predictor Variability section did show that Spondylolisthesis Grade is an important predictor. Therefore, since Spondylolisthesis Grade was identified as an important predictor and QDA was not able to take advantage of Spondylolisthesis Grade because the predictor does not meet QDA's requirement of being bivariate normal, K-Nearest Neighbors was implemented next as a comparative to Random Forest because KNN is a non-parametric classifier, which means KNN does not consider predictor distribution (Speck, 2017).

2.6.3 K-Nearest Neighbors

Despite KNN being immune from having to rely on data following a specific distribution, KNN comes with its own challenges. First of all, a major challenge with KNN (as with QDA), is the curse of dimensionality (Speck, 2017; Irizarry, 2022, Chapter 31). Fortunately, the variable importance derived from the five fold cross-validated Random Forest model helped identify predictors that should add value to the KNN solution. Based on predictor importance being predetermined, overcoming potential over fitting using KNN was protected against by using the most important predictor first, and then adding the next most important predictor, and so on. With this approach, the anticipated outcome was to find a balance between avoiding over fitting and optimizing the KNN method, where each KNN model added the next most important predictor, was cross-validated, and the smoothing window (K) was tuned.

Before the models could be assessed, a metric to compare the various KNN models was needed. Assessing the fit was selected because fit can be evaluated as over, under, or best. Models that are over fit and under fit can be identified by comparing the test accuracy of various models and individual model training and test accuracy in conjunction with model complexity (Dalpaiz, 2020; Dalpaiz, 2022). In general, a model is over fit when the test accuracy is less than the train accuracy; a model is under fit when the test accuracy is greater than the train accuracy, yet, not the best performing model; and the best fit model

has a test accuracy that is greater than the train accuracy while being the least complex out of all under trained models (Dalpaiz, 2020; Dalpaiz, 2022).

The following plot shows the cross-validation and tuning process for a variety of smoothing values (K) used during the KNN model building process of adding the next most important predictor per iteration. In addition, the following table summarizes the predictors used; the best tune for K; the accuracy attained for the training data and test data; and a fit result of over, under, or best:

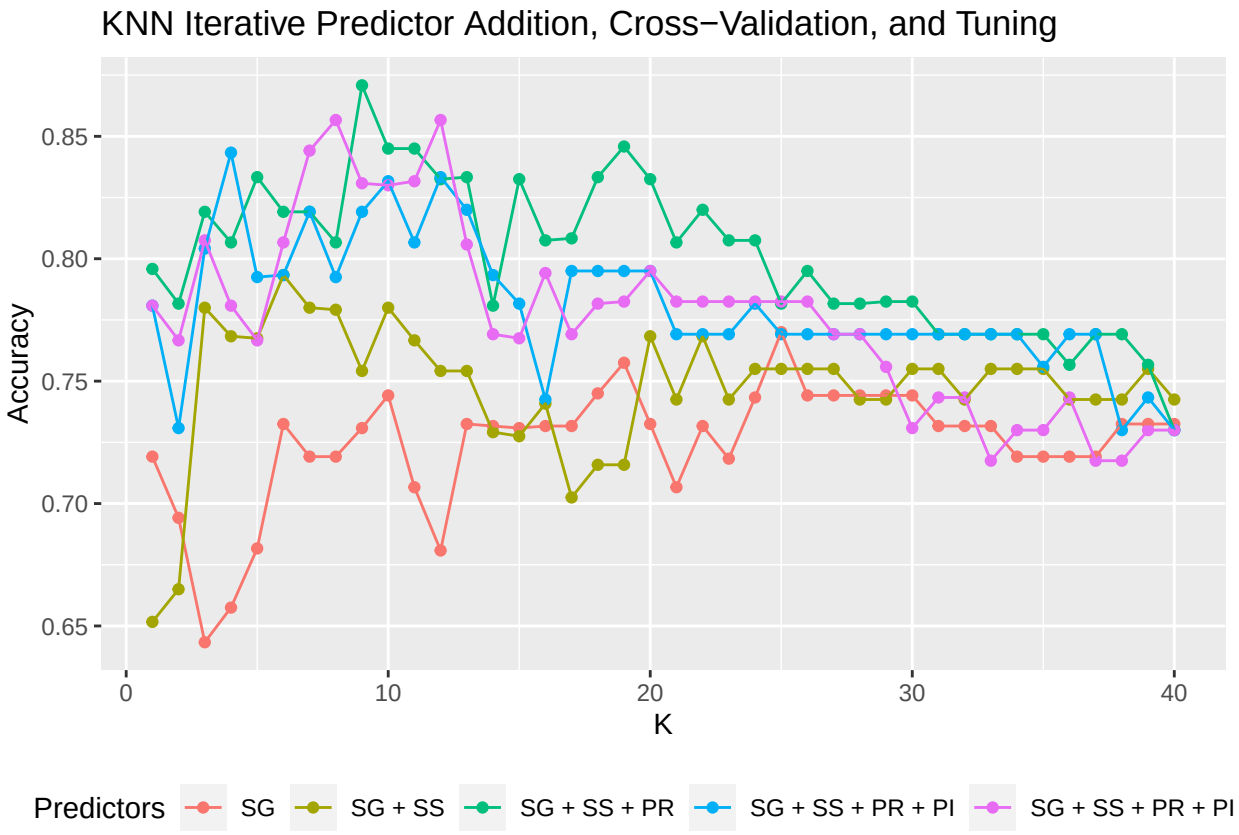


Table 16: KNN Iterative Predictor Addition, Cross-Validation, Tuning and Fit

SG	SS	PR	PI	PT	LL	K	Train Accuracy	Test Accuracy	Fit
Applied	NA	NA	NA	NA	NA	25	0.7700000	0.7922078	Under
Applied	Applied	NA	NA	NA	NA	6	0.7933333	0.8051948	Under
Applied	Applied	Applied	NA	NA	NA	9	0.8708333	0.7922078	Over
Applied	Applied	Applied	Applied	NA	NA	4	0.8433333	0.7142857	Over
Applied	Applied	Applied	Applied	Applied	NA	12	0.8566667	0.8571429	Best

Based on the foregoing definitions of fit, the best performing KNN model used predictors of Spondylolisthesis Grade (SG), Sacral Slope (SS), Pelvic Radius (PR), Pelvic Incidence (PI), and Pelvic Tilt (PT), but did not use Lumbar Lordosis, since the Random Forest variable importance reported Lumbar Lordosis to have zero contribution. Going forward, this best performing KNN model will be referred to as the KNN-5 model.

3 Results

Often in supervised machine learning classification, accuracy can be used as a general approach to access model performance, but it can be deceptive, particularly when some classes have greater prevalence than others (Irizarry, n.d.). Instead, a confusion matrix provides substantial information in grid form that can be used to assess solution performance. To better understand confusion matrices, the following image and definitions are provided.

Confusion Matrix and Parameter Definitions

Table 17: Confusion Matrix

	Actually Positive	Actually Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

(Irizarry, 2022, Chapter 27)

True Positive (TP)

Occurs when a positive sample is classified correctly as a positive (Kundu, 2023).

True Negative (TN)

Occurs when a negative sample is classified correctly as a negative (Kundu, 2023).

False Positive (FP)

Occurs when a negative sample is classified incorrectly as a positive (Kundu, 2023).

False Negative (FN)

Occurs when a positive sample is classified incorrectly as a negative (Kundu, 2023).

In addition, confusion matrix parameters can be used to compute other performance measures (Dalpaiz, 2022, Chapter 17). Of these available metrics, the following measures have been selected based on their connection to overall model performance, relevance for data sets that have class prevalence bias, and meaning to the medical field in terms of the cost of errors (Irizarry, n.d.).

Accuracy

The proportion between correct classifications and total classifications (Irizarry, n.d.):

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

(Olugbenga, 2023)

As noted previously, accuracy can be deceptive, particularly when some classes have greater prevalence than others (Irizarry, n.d.). For instance, in a binary classification where one class represents 90 percent of the observations, with random guessing an imbalance in accuracy between the two classes occurs simply because there is more of one class and less of another

(Irizarry, n.d.), which makes accuracy misleading for model to model comparison and assessment. Fortunately, other metrics derived from confusion matrix values are not clouded by class prevalence, such as studying sensitivity and specificity individually, using balanced accuracy, or using F1-Score (Irizarry, n.d.).

Sensitivity

The algorithms ability to make a positive prediction when the true value is positive; in other words $\hat{Y} = True\,, when\, Y = True$, which is also known as the True Positive Rate (TPR), or Recall (Irizarry, n.d.):

$$Sensitivity = \frac{TP}{TP + FN}$$

(Irizarry, n.d.)

Specificity

The algorithms ability to make a negative prediction when the true value is negative; in other words $\hat{Y} = False\,, when\, Y = False$, and is also known as the True Negative Rate (TNR) (Irizarry, n.d.):

$$Specificity = \frac{TN}{TN + FP}$$

(Irizarry, n.d.)

Sensitivity and specificity are particularly important to evaluate individually in the context of the cost of error in medicine. For example, failing to diagnose cancer when cancer is present (poor sensitivity) could result in end of life, whereas diagnosing cancer when cancer is not present (poor specificity) is not ideal due to the side effects of chemotherapy drugs used in cancer treatment, but the consequence is less costly than poor sensitivity, which could result in death.

Precision

The ratio of positive predictions derived from the positive class, also known as the Positive Predictive Value (PPV), which is the accuracy of the True Positive (TP) (Olugbenga, 2023).

$$Precision = \frac{TP}{TP + FP}$$

(Irizarry, n.d.)

An important note about precision is its relationship to prevalence. Because precision depends on prevalence, higher precision can be achieved even when guessing (Irizarry, n.d.).

Balanced Accuracy

The mean of sensitivity and specificity (Irizarry, n.d.):

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2}$$

(Irizarry, n.d.)

Balanced Accuracy is particularly useful when there is class prevalence imbalance within the data set (Olugbenga, 2023).

F1-Score

The harmonic mean of precision and recall (Irizarry, n.d.):

$$F1 = \frac{1}{\frac{1}{2}(\frac{1}{Recall} + \frac{1}{Precision})}$$

(Irizarry, n.d.)

In addition to F1-Score being useful when a class prevalence imbalance exists within a data set (Olugbenga, 2023), it can also be used to independently weight precision and recall (Irizarry, n.d.). In doing so, precision and recall can be weighted to better the apportion the cost of error given a specific use case (cancer diagnosis example provided previously), which ultimately allows for superior algorithm performance comparisons, and building a framework to determine when an algorithm is ready to be deployed.

With the applicable metrics for supervised machine learning classification defined along with their use, the analysis of the results for QDA, Random Forest, and KNN was ready to proceed.

3.1 Model Performance Comparisons

Based on the expected class prevalence bias in the data and the implications of incorrect classifications in medicine, accuracy alone is not enough to determine the best performing model. Yet, accuracy does have value because of its ability to provide a general performance assessment, and the ability to reveal whether a model over fits, under fits, or is the best fit when applied to new data. The following table shows the results of applying the same fit assessment methodology used in the Modeling Approach and Model Building section that was applied to the various KNN models, which is based on the fit criteria established by Dalpaiz (2020; 2021):

Table 18: Verification Data Fit Assesment Based on Model Accuracy

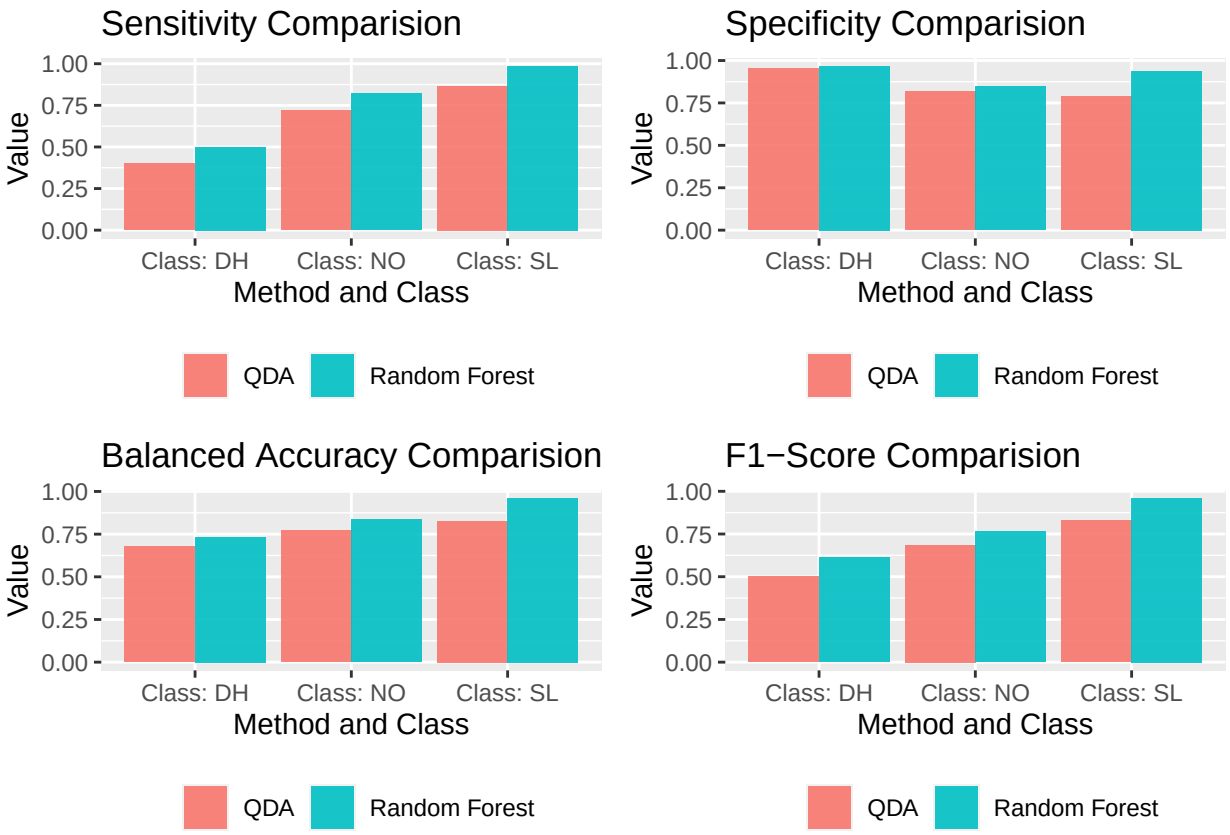
Method	Test Accuracy	Verification Accuracy	Fit
QDA	0.6753247	0.7290323	Under
KNN-5	0.8571429	0.8258065	Over
Random Forest	0.8051948	0.8387097	Best

Surprising or not, the KNN-5 model appears to be susceptible to the curse of dimesionality as the KNN-5 model over fit the verification data. Perhaps not surprising, Random Forest performed well because decision trees are unaffected by the curse of dimensionality, and as described by Podgorelec et al. (2002) can be a reliable method in medicine for diagnosis. Furthermore, since Random Forest was implemented over other decision tree methods such as rpart, the solution also escaped the pitfalls of over fitting that decision tree packages are susceptible to, but that Random Forest is not (Irizarry, n.d.). Overall, because KNN-5 over fit the verification data, and therefore, renders the model unreliable for the application to new data, the confusion matrix derived performance metrics identified as useful given the context and project goal will not be discussed for the KNN-5 model. Instead, the results of QDA and Random Forest became the focus of the results analysis because both solutions did not over fit.

Although Random Forest outperformed QDA when considering accuracy only, sensitivity, specificity, balanced accuracy, and F1-Score, can provide a more precise perspective for comparing model performance. These measures are automatically generated by the caret package's confusionMatrix function, and can easily be extract from the returned object. The expectation is that sensitivity and specificity will provide information as to whether QDA or Random Forest results are more suitable for integration into a medical image augmentation system for spine classifications, since sensitivity and specificity are measures of the cost of error (Irizarry, n.d.). Moreover, balanced accuracy and F1-Score will provide insight as to whether QDA or Random Forest would likely generalize to other future data sets with different class prevalence, since balanced accuracy and F1-Score work well for model assessment when class prevalence bias is present. To complete this analysis, the following table summarizes the classes and their associated metrics for QDA and Random Forest, and the following plot provides a visual comparison of these performance metrics for QDA and Random Forest:

Table 19: Random Forest and QDA Performance Metrics

Method	Class	Prevalence	Sensitivity	Specificity	Balanced Accuracy	F1
QDA	Class: DH	0.1935484	0.4000000	0.9520000	0.6760000	0.5000000
QDA	Class: NO	0.3225806	0.7200000	0.8190476	0.7695238	0.6857143
QDA	Class: SL	0.4838710	0.8666667	0.7875000	0.8270833	0.8280255
Random Forest	Class: DH	0.1935484	0.5000000	0.9680000	0.7340000	0.6122449
Random Forest	Class: NO	0.3225806	0.8200000	0.8476190	0.8338095	0.7663551
Random Forest	Class: SL	0.4838710	0.9866667	0.9375000	0.9620833	0.9610390



The sensitivity, specificity, balanced accuracy, and F1-Score for Random Forest exceed those of QDA for all classes. Perhaps QDA did not perform as well as Random Forest because the degree of normality of the predictors used in QDA cross-validation were inadequate. Alternatively, QDA may be a more realistic solution compared to Random Forest because QDA is unaffected by the class prevalence bias that is expected to exist within the data set, where Random Forest does not have the same bias controlling characteristics (O'Brien & Ishwaran, 2019).

Regardless of the class prevalence bias, as noted previously, balanced accuracy and F1-Score are useful as machine learning solution metrics for imbalanced data sets because they can evaluate an algorithm in a way that prevalence does not cloud the assessment (Irizarry, n.d.). Therefore, from the bar plots and table above, the following performance comparisons between QDA and Random Forest were drawn:

- For both balanced accuracy and F1-Score, Random Forest out performed QDA in classifying Spondylolisthesis with metrics of 0.962 for balanced accuracy and 0.961 for F1-Score
- For both balanced accuracy and F1-Score, Random Forest out performed QDA in classifying Normal spine characteristics with metrics of 0.833 for balanced accuracy and 0.766 for F1-Score;

- For both balanced accuracy and F1-Score, Random Forest out performed QDA in classifying Disk Hernia with metrics of 0.734 for balanced accuracy and 0.612 for F1-Score

Overall, specificity, sensitivity, balanced accuracy, and F1-Score indicate that Random Forest better classifies Spondylolisthesis, Normal spine, and Disk Hernia compared to QDA.

Despite the ability of balanced accuracy and F1-Score to avoid being clouded by class prevalence, the relationship between real world prevalence of Disk Hernia and Spondylolisthesis and this Vertebral Column Data Set has yet to be considered. To attempt to gain a sense of any disparity between real world prevalence and the class prevalence in the Vertebral Column Data Set, secondary research was sought. According to Dydyk et al. (2022), the prevalence of Disk Hernia is somewhere in the range of 5 to 20 cases per 1,000, and Tenny & Gillis (2022) state that up to 18 percent of lumbar Magnetic Resonance Imaging (MRI) patients present with Spondylolisthesis of various grades. Although some prevalence information was found, being able to compare these measures to those in the Vertebral Column Data Set was difficult, because the secondary research presented binary classification between Disk Hernia and Normal spine characteristics and binary classification between Spondylolisthesis and Normal spine characteristics, whereas the Vertebral Column Data Set has three classes. For this reason, assumptions were required to compare the prevalence from secondary research with the prevalence in the Vertebral Column Data Set. For simplicity, the normal classifications within the Vertebral Column Data Set were assumed to be representative of the real world, which left a direct comparison to the secondary research for Disk Hernia and Spondylolisthesis prevalence.

The following table shows the class percentages from the verification data (which is approximately the same for the training and test data sets):

Table 20: Random Forest and QDA Verification Data Prevalence

Method	Class	Prevalence
QDA	Class: DH	0.1935484
QDA	Class: NO	0.3225806
QDA	Class: SL	0.4838710
Random Forest	Class: DH	0.1935484
Random Forest	Class: NO	0.3225806
Random Forest	Class: SL	0.4838710

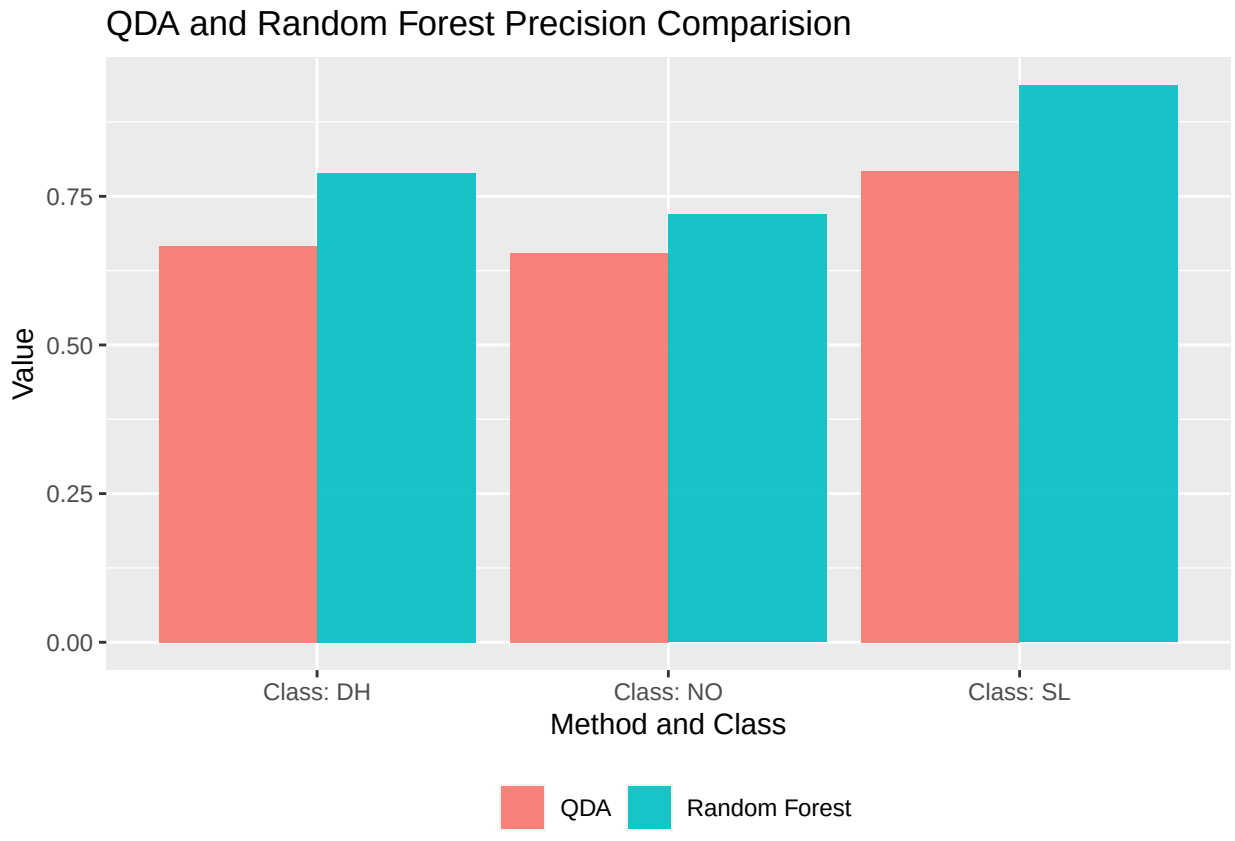
The prevalence from the Vertebral Column Data Set shown in the table above differs from the values presented by Dydyk et al. (2022) and Tenny & Gillis (2022). More specifically, two percent prevalence for Disk Hernia was noted in the secondary research versus 19 percent in the Vertebral Column data, and up to 18 percent prevalence for Spondylolisthesis in the secondary research (which is based on MRI imagery collected – true real world prevalence may be much smaller as healthy people generally are not subjects of MRI collection) versus 48 percent in the Vertebral Column data. Regardless of the prior assumption of the Normal class in the Vertebral Column Data Set being representative of the real world prevalence, the merits of the assumption, or others without better real world prevalence data, is weak,

and therefore, exposes a limitation of this small data set and the ability to connect it to real world prevalence.

In addition to real world prevalence, precision also matters given the goal of determining the feasibility of creating a future medical image augmentation system. Therefore, a summary table for class precision for QDA and Random Forest are shown in the table below, and a comparison of precision for QDA and Random Forest are shown in the plot below because precision matters in medical diagnosis for its ability to quantify the accuracy of True Positives (specificity) (Irizarry, n.d.); in other words, the capability of making a positive diagnosis when a medical condition exists. However, precision should be scrutinized due to its limitation of potentially producing optimistic values because of its tie to prevalence:

Table 21: QDA and Random Forest Precision Comparision

Method	Class	Precision
QDA	Class: DH	0.6666667
QDA	Class: NO	0.6545455
QDA	Class: SL	0.7926829
Random Forest	Class: DH	0.7894737
Random Forest	Class: NO	0.7192982
Random Forest	Class: SL	0.9367089



Overall, since imbalanced data has the ability to cloud Random Forests decision making (O'Brien & Ishwaran, 2019), despite Random Forest having better performance metrics than QDA, which is believed to be devoid of class prevalence effects, given the small data set with limited ability to control the class prevalence, and the challenges in mirroring real world prevalence, as a precautionary approach, QDA was selected as the most trustworthy model to assess suitability for a future medical image augmentation system.

Based on selecting QDA as the preferred machine learning solution, QDA may provide meaningful augmentation information to medical image assessments based on the performance metrics observed for balanced accuracy, F1-Score, and precision. These three metrics were selected because balanced accuracy and F1-score incorporate measures of sensitivity and specificity, which are useful in quantifying the cost of errors and their characteristics of having relief from the clouding of class prevalence, and precision was selected as a specificity measure since correctly diagnosing is important in medicine.

Keeping the goal in mind of assessing the potential use of QDA in a future medical image augmentation system, setting a threshold to determine whether or not a method would be valuable for augmentation helps to bring objectivity to the assessment. Therefore, a measure of 0.70 for the mean of balanced accuracy, F1-Score, and precision, which can be thought of as ensembling metrics relevant for the data characteristics of class prevalence bias and the medical use case of considering the cost of errors, was selected as a cut-off to judge the predictive utility for the potential use of each class in a future medical image augmentation system. In addition, and in the context of developing a framework for medical image augmentation systems, the metrics used in the assessment ensemble could be weighted by health professionals based on their specific needs and their approach to diagnostics, which can provide the level of control that they may desire.

To facilitate the threshold analysis, the means of balanced accuracy, F1-Score, and precision for the QDA classifications are shown in the following table:

Table 22: Mean QDA Performance Metrics (Ensemble of Balanced Accuracy, F1-Score, and Precision)

Method	DH Mean	NO Mean	SL Mean
QDA	0.6142222	0.7032612	0.8159306

From the table above, the mean score for Spondylolisthesis and Normal spine classes met or exceeded the threshold of 0.7, and are therefore, recommended for consideration in a future medical image augmentation system. However, the mean score for Disk Hernia did not meet the 0.70 threshold. Therefore, QDA Disk Hernia classifications were not recommended to be included as an information source within a future future medical image augmentation system.

4 Conclusion and Future Work

Out of three machine learning solutions of QDA, Random Forest, and KNN used to predict spine classifications of Disk Hernia, Normal, and Spondylolisthesis, the accuracy for the cross-validated solutions generated using Random Forest and QDA had higher accuracy on the verification data than the preliminary test accuracy results, which was evidence that Random Forest and QDA did not over fit. Since the KNN-5 solution produced accuracy on the verification set that was three percent lower than the test set, the KNN-5 solution was categorized as over fit.

In comparing the performance of QDA and Random Forest, Random Forest out performed QDA for every metric used to analyze the results. More specifically, these metrics consisted of sensitivity, specificity, balanced accuracy, F1-Score, and precision. Despite Random Forest exhibiting better performance than QDA, QDA was selected as the most appropriate model to consider for a future medical image augmentation system due to the class prevalence characteristics of the Vertebral Column Data Set, which invoked the belief that the decision trees generated by Random Forest have bias (O'Brien & Ishwaran, 2019) that is less controlled compared to QDA.

The class prevalence characteristics of the Vertebral Column Data Set and the small size of the data set (310 patients) were the major limiting factors in producing solutions that were free from bias. More specifically, the class prevalence disparity within the Vertebral column Data Set and the difficulty ascertaining whether the class prevalence disparity was representative of real world prevalence of Disk Hernia and Spondylolisthesis was one of the major contributors to selecting QDA over Random Forest, because QDA is a version of Naive Bayes that controls for prevalence (Irizarry, 2022, Chapter 31).

Another major factor that influenced selecting QDA over Random Forest was discovering that highly imbalanced data has been observed to reduce the reliability of Random Forest solutions to generalize well to new data, because tree formation within the model may be based on classification error (O'Brien & Ishwaran, 2019). This implied potential for classification error could make the decisions for tree algorithms non-nonsensical when applied to future data that has a different class prevalence. Although class imbalance was evident, stating with confidence whether the imbalanced data is extreme enough to render nonsensical decision trees cannot be proven nor evaluated with this small data set. Therefore, the conservative decision was to reject Random Forest until the solution could be retrained and evaluated on a larger data set where prevalence could be controlled.

Although opinions may differ between health professionals and data scientists, a cut-off score for the mean of balanced accuracy, F1-Score and precision of 0.7 was used to determine the classes that would potentially qualify for implementation in a future medical image augmentation system. For the QDA solution, a Normal classification achieved a cut-off score of 0.703, Spondylolisthesis achieved a cut-off score of 0.816, and Disk Hernia achieved a cut-off score of 0.614. Based on these scores and the threshold, Spondylolisthesis and Normal classifications could be considered for use in a future medical image augmentation system. More importantly, in the spirit of developing a framework for medical image augmentation systems, the metrics used in the assessment ensemble could be weighted by health profes-

sionals based on their specific needs and their approach to diagnostics, which could provide the level of control that they may desire.

Given the foregoing limitations of the small data set size and the class prevalence issues identified, future work requires a much larger data set that is representative of the prevalence of Disk Hernia and Spondylolisthesis in the real world. With a data set of sufficient size and real world prevalence, Random Forest could potentially be retrained to obtain a reliable, effective, and accurate medical decision making model (Podgorelec et al., 2002). In addition, an opportunity to obtain a more robust data set may present itself by developing a Shiny web application, where health professionals can upload imagery and measurements provided patient consent has been obtained.

Should a Shiny web application be developed, the expanded data set could be continually retrained and evaluated in order to advance this specific application along with providing a rich data set that can be used in the future for other research. Potentially, the long term impacts may include, advancement in conceptualization and building upon the framework laid out here for the assessment of machine learning results for medical applications, and increased interest in developing augmentation systems. However, to be clear, the goal of creating and promoting machine learning augmentation systems in medicine is not to replace the judgment of qualified health professionals; as the classification results here show, the predictions are by no means perfect. Instead, the potential benefits include providing a data rich environment to assist health professionals, and the construction of various data platforms for collecting and gaining insights from medical data that in the future can benefit society in new and novel ways.

5 References

Daffodil Software. (2017, July 30). *9 Applications of Machine Learning from Day-to-Day Life.* Medium. https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Dalpaiz, D. (2020). *R for Statistical Learning.* https://daviddalpiaz.github.io/r4sl/

Dalpaiz, D. (2022). *Applied Statistics with R.* https://book.stat420.org

Dydyk, A.M., Massa, R.N., Mesfin, F.B. (2022, January 18). *Disc Herniation.* National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK441822/#:~:text=The%20incidence%20of%20a%20herniated,1%2D3%20percent%20of%20patients.

Irizarry, R.A., (n.d.) Professional Certificate in Data Science [MOOC]. HarvardX https://www.edx.org/professional-certificate/harvardx-data-science

Irizarry, R.A., (2022). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R* bookdown. http://rafalab.dfci.harvard.edu/dsbook/

Koslosky, E., & Gendelberg, D. (2020). Classification in Brief: The Meyerding Classification System of Spondylolisthesis. *Clinical Orthopaedics and Related Research, 478*(5), 1125-1130. 10.1097/CORR.0000000000001153

Kuhn, M. (2019, March 27). *The caret Package.* GitHub. https://topepo.github.io/caret/index.html

Kundu, R. (2023, March 2). *Confusion Matrix: How To Use It & Interpret Results [Examples].* V7 Labs. https://www.v7labs.com/blog/confusion-matrix-guide

Le Huec, J.C., Aunoble, S., Philippe, L. (2011). Pelvic parameters: origins and significance. *European spine journal, 20*(5), 564-571. 10.1007/s00586-011-1940-1.

Mota, H., Barreto, G., Neto, A. (2011). Vertebral Column Data Set [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/vertebral%2Bcolumn

O'Brien, R., & Ishwaran, H. (2019). A Random Forests Quantile Classifier for Class Imbalanced Data. *Pattern Recognition, 90*, 232–249. doi:10.1016/j.patcog.2019.01.036.

Olugbenga, M. (2023, January 25). *Balanced Accuracy: When Should You Use It?* neptune.ai. https://neptune.ai/blog/balanced-accuracy

Pileggi, S. (2022, January 23). *Report Ready PDF tables with rmarkdown, knitr, kableExtra, and LaTeX.* Piping hot data. https://www.pipinghotdata.com/posts/2022-01-24-report-ready-pdf-tables-with-rmarkdown-knitr-kableextra-and-latex/

Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of Medical Systems, 26*(5), 445-463. 10.1023/a:1016409317640

Speck, M. (2017, May 15). *What is K-Nearest Neighbors?* Medium. https://medium.com/@mjspeck/what-is-k-nearest-neighbors-c9b4cdf9f35c

STHDA. (n.d.). *ggplot2 barplots : Quick start guide - R software and data visualization.* Statistical Tools for High-Throughput Data Analysis. http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

Tenny, S., & Gillis, C.C. (2022, May 24). *Spondylolisthesis.* https://www.ncbi.nlm.nih.gov/books/NBK430767/#:~:text=Current%20estimates%20for%20prevalence%20are,for%2075%25%20of%20all%20cases.

The R Foundation. (n.d.) What is R? https://www.r-project.org/about.html

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10), 1-23. https://doi.org/10.18637/jss.v059.i10

Wilke, C.O., (2022, December 15). *Introduction to cowplot.* CRAN. https://cran.r-project.org/web/packages/cowplot/vignettes/introduction.

html#:~:text=The%20cowplot%20package%20is%20a,or%20mix%20plots%20with%20images.

Xie, Y., Dervieux, C., Riederer, E. (2022). *R Markdown Cookbook.* Chapman & Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook/

Zhu, H. (2021). *Create Awesome LaTeX Table with knitr::kable and kableExtra.* R-project. https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_pdf.pdf

Acknowledgments

This work has been supported by the HarvardX Data Science Professional Certificate course material, and the various textbooks and vignettes provided within the references section of the report. Creating this report would not have been possible without the learning that took place from completing the first eight courses in the series, the knowledge gained from reading the sources noted, and the R-package documentation that was used to implement this work.