What clinical measures could be used to predict the presence of low back dysfunction and how reliable can prediction of low back pain be using machine learning methods?

Student Name: Hope McLeod

Submission Date: 3rd September 2018

Module Code/Name: INSTG099 – MSc Dissertation

Supervisor's Name: Dr. Luke Dickens

Word Count:

This dissertation is submitted in partial fulfilment of the requirements for the Master's degree in Information Science, UCL.

The Harvard referencing style has been used for this dissertation.

Abstract

The aim of this project was to look at which clinical measures could be used to predict the presence of low back dysfunction. It looked into how this data could be collected and selected using evidence based measurements and tests that would be typically done within a physical therapy outpatients setting. As a separate section it also looked at the importance of having good data to use in building predictive models for low back pain. The belief is that machine learning (ML) can discover good attributes to be used in these models with the intention of integrating them into a clinical decision support system.

The above was addressed in two sections. First, the project investigated the measurements and tests physical therapists use to differentially diagnose a patient's back complaint. It used results from studies to select ones with the greatest levels of accuracy and reliability. This part of the project also presents an idea for a research study that would use these measurements for building prediction models that identify a back condition from samples of data. Secondly, the project analysed a public dataset (containing measurements from spinal x-rays) to look at how valid the features values were. It then used them to improve prediction accuracy on ML models that have been built by others interested in the topic.

In summary, the first part of the project looks at how good data can potentially be collected for a dataset to be used in an ML model to predict back conditions and the second part of the project demonstrates how important it is to use good data in order to build reliable models.

From doing literature searches and analysis on datasets, this project shows that there are potentially a sufficient number of features that can be used for modelling back pain. It shows that with the little spinal data publicly available, back pain can be modelled successfully using ML methods. It also shows that there is more than one way of improving a model's predictive ability by part experimentation but having prior knowledge of ML algorithms and their parameters goes a long way towards building the best models .

Because not much research has been done on this subject in conjunction with ML, a lot more data needs to be collected before coming to solid conclusions about modelling back pain.

Overall, the conclusion is that ML could potentially be a useful tool to assist clinicians in making decisions about data related to the spine.

Declaration

I have read and understood the College and Departmental statements and guidelines concerning plagiarism. I declare that this submission is entirely my own original work; wherever published, unpublished, printed, electronic or other information sources have been used as a contribution or component of this work, these are explicitly, clearly and individually acknowledged by appropriate use of quotation marks, citations, references and statements in the text. It is **XXX** words in length

Table of Contents

List of Tables and Illustrations

Illustration Index

Index of Tables

List of Abbreviations

CLBP	Chronic Low Back Pain
CT	Computer tomography
DT	Decision Tree
GP	General practitioner
ICC	Intra-class coefficient
K	Kappa coefficient
K MOD	Kernel with moderate decreasing
KNN	K nearest neighbour
L1	Ridge regularisation
L2	LASSO regularisation
LASSO	Least absolute shrinkage and selection operator
LogR	Logistic Regression
LR	Likelihood ratio
ML	Machine learning
MRI	Magnetic Resonance Imaging
n.d.	no date
NB	Naive Bayes
NICE	National Institute for health and Clinical Evidence
NSLBP	Non-specific low back pain
RF	Random Forest
RBF	Radial Basis Function
SMOTE	Synthetic minority oversampling technique
SVM	Support Vector Machine
UCI	University of California Irvine

Acknowledgements

I would like to thank Dr. Luke Dickens for his time and help during the project; he has always been very generous in that respect. Also, to him and the department of Information Studies for introducing a more technical aspect to the department, albeit challenging, but without doubt, worth it.

Introduction

‘The various terms used to describe low back pain reflect our difficulty in accurately identifying discrete causes of low back pain and our inability to accurately define which characteristics might help to identify specific causes.’

(NICE, 2016)

This statement from the National Institute for health and Clinical Excellence (NICE) makes a valid point with regards to diagnosing back pain. Having worked as an osteopath for 15 years, diagnosing back pain could be difficult due to back complaints presenting in similar ways. With regards to identifying the causes of pain when questioning, patients do not associate the unreasonable demands they place on their bodies as a reason for a complaint; they view the intensity of their activities of daily living as normal. There are also no definitive tests done in a physical therapy outpatient setting which can identify a condition confidently.

To help identify the tissue causing symptoms, therapists rely on patients being able to describe where their pain is, how it feels and what aggravates and relieves it. Text books do the best they can to try and map a region of pain with an anatomical structure with structures nearer the surface of the body often more easily located by patients compared to structures deeper down which can give more diffuse and vague symptoms. Pain is not easy to describe and what one pain feels like to one person will be different to another. With regards to clinicians, the tests and measurements chosen are inconsistent with ones chosen by their colleagues. Therefore the tests with the best reliability are not always used. Data and methods are needed to help identify components involved in back pain and this is what this project set out to investigate.

It is the belief that ML can help understand the causes of back pain by building models capable of predicting it and its associated conditions. The aim of this project is to look into clinical features of back pain which could be used to train a machine learning model to predict it.

Arthur Samuel came up with the term 'machine learning' in 1959 (Hellard et al, 2018) but its application wasn't really applied until the noughties which is when the concept of big data became known (Press, 2013). Despite this, there is an "absence of numerical attributes that quantitatively describe the pathologies of interest to the field of orthopedics" (Neto et al, 2011). To put this statement into context, if look at the literature on using ML to predict vertebral column pathologies there are several studies that are using the same UCI dataset that is referred to later in this project. Akben (Akben, 2016), Huang (Huang, 2014), Unal (Unal, 2014) and Reddy (Reddy, 2012) all refer to this dataset in their research. But is it a problem that physical therapists and orthopaedic clinicians would be keen to solve using ML methods? Neto states clinicians are happy to use these systems in helping them make decisions. With clinical records moving away from paper to electronic based systems (Macaulay, 2016), the task of collecting data should become a lot easier.

With there being limited access to datasets, this project first focuses on a method of how a dataset could be created by describing a potential study in the "Methodology" chapter. The other part of the project will analyse a public dataset and show the effect good and bad data can have on a ML model created to predict back pain. Outcomes from both parts will be outlined in the a results and analysis chapter. The pre-ultimate chapter discusses how all this supports the idea of using ML for prediction and whether there are good clinical features that could be input into forming these models. The

dissertation ends with a conclusion chapter which talks about how the project contributed to the research and ideas for future studies other than the one proposed.

Literature Review

The project is about investigating clinical features that could be used to help predict the presence of low back dysfunction and the reliability of using ML methods to predict it. To fulfil the project objective the literature search consisted of finding out about:

- the prevalence of back pain and its effects on society
- physiotherapy measurements and tests that can be carried out reliably and accurately
- ready made datasets containing spinal data
- research that has involved building ML models for predicting back pain and algorithms capable of binary classification

Google scholar, UCL library and the internet including reputable ML blogs and websites were used

The prevalence of low back pain and its effects on society

UNISON, the second largest trade union in the UK, refers to a large European study (Breivik, 2006) which confirms that out of all chronic pains looked at, low back pain forms 19% of "persistent and intrusive pain". Within the workforce it is one of the main reasons for sickness with up to 12 million days lost a year (UNISON, 2018). NICE, an organisation that provides guidance to the NHS, public sector workers and social care services about various diseases and conditions including back pain state that low back pain "causes more disability, worldwide, than any other condition" (NICE, 2016). With regards to costs incurred due to chronic low back pain, in 1998 approximately £500 million for the cost of care was spent in the private sector, more than £1000 million was spent in the NHS and £3500 million was lost in hours of production (NICE, 2009). The Lancet, a reputable

medical journal published a paper (Buchbinder et al, 2018) that reported that back pain "is the leading worldwide cause of years lost to disability" and because people are living for longer, it is becoming even more of a problem.

Finding reliable and accurate tests and measurements to help diagnose low back dysfunction

The methodology section on finding data for a dataset suggests that data be collected from physiotherapists because they see a lot of patients and record a substantial amount of information about their backs. The literature review therefore required searching for measurements and tests that could be used for a dataset for future study. Unfortunately NICE is not optimistic with regards to finding reliable and accurate measures and tests. They state:

"… there are no reliable clinical features or imaging findings that allow us to identify" spinal structures "with any confidence." (NICE, 2016)

However, this does not mean that the most reliable measurements and tests should not be endeavoured to be found and explored within the realms of ML.

Originally Simpson (Simpson, 2006) was thought to be a good source as it is a systematic review on the accuracy of spinal orthopaedic tests. But this paper referred to a researcher that carried out a few studies with unconvincing results. They mentioned two studies done by this researcher and each time 48 participants were used. In the first study each test had the same specificity and sensitivity. The next study by this researcher involved the same tests but this time the specificity and sensitivity for each test was very different to the previous study. This might just be that their methods of checking accuracy and reliability were different in the two studies.

Cleland's (Cleland, 2016), on the other hand, was presented well and scored tests and measurements with the same methods as Simpson. But it also included other methods such as likelihood ratio which indicates the power of a test to confidently diagnose or disregard a condition and it showed the scores for methods used to assess intra and inter-reliability of clinician measurements. The measurements and tests from this book were the main source for a new dataset. Cleland focuses on the following spinal conditions: lumbar spinal stenosis (narrowing of holes where nerve tissue passes), lumbosacral radiculopathy (single nerve dysfunction), ankylosing spondylitis (inflammation and fusing of the spine), disc herniation (slipped disc) and non-specific low back pain (NSLBP).

Finding ready made spinal datasets

Datasets from Kaggle (Kaggle, 2016) and UCI (Barreto, n.d.) were used in the section of the dissertation where models are built. The datasets are from a repository of publicly available datasets and were the only sources found for back pain. Their content will be described in more detail later on in this project but they contain data samples of measurements taken from spinal x-rays.

Terminology used in machine learning

Machine learning comes in various forms (supervised, unsupervised and reinforcement learning). The most common type is supervised learning and this is the type used in this project. Supervised ML is where the available input data to learn from is associated with a value or category that needs to be predicted. If the output to be predicted is a category then the ML task is classification. If the task is to predict a numeric value on a continuous scale, then it is a regression task. The input data to a ML program can sometimes be referred to as features, attributes and independent variables. Whereas the output can be known as a target or dependent variable.

ML use algorithms to learn from data with the goal of creating a general rule (model) that is able to map a new input value to an output. When an algorithm is learning this general rule the model is, at this point, going through a training process. Training a model needs input data, an algorithm and usually some form of extra parameters that can be used to fine tune the model.

During training, the algorithm chosen knows the input (and output in the case of supervised learning) and its task is to find out coefficients that when applied to the feature values will map that sample to an output. If the input data is known to follow particular distributions such as the normal distribution then there are two parameters to be learnt such as standard deviation and mean. Some algorithms will make assumptions about the input data following a distribution and when this is the case a parametric model is produced (linear SVMs, logistic regression, linear regression). In other cases, no assumptions are made about the distribution of the data and a non-parametric model is produced (KNN, non-linear SVM). The non-parametric model still has parameters but the number of them is unknown.

The idea is for the model produced to be able to generalise about data which means that its predictions accuracy during training is reflected when testing. To get good generalisation the input dataset is split into a train and test set during training. The algorithm uses the train portion to generalise about the data it has “in front of it” and then use the unseen test data to test its generalisability. If the training set gives a good prediction rate but a poor one with the test data then the model is said to be over fitting during training. What this means is that the training data thinks that the true points and noise in the dataset are all valid. So part of training is learning to distinguish

between true data and noise. There are methods which allow this to be controlled and regularisation is one such method.

Regularisation is a process that penalises an input weight if it is too high. Some regularisation methods can act by eliminating a feature completely (LASSO) by setting an input weight to zero and as a result these regularisation methods act as a form of feature selection.

Finally the classification models can be referred to being generative of discriminative. A generative model will try and learn how the data itself was generated. On doing this it is able to make assumptions about the data and associate it with a category. Because generative models learn about how the data itself is generated, it is a good way of producing synthetic data. Discriminative models just learn about the data in the dataset and therefore makes fewer assumptions. Because it bases its assumptions about the data it has in front of it, it depends on good quality data. Discriminative models can be sub categorised into non-probabilistic and probabilistic. SVMs are non-probabilistic and logistic regression is probabilistic. Logistic regression works out the probability of a data point belonging to both classes and bases its prediction on the highest probability.

Current research on predictive models for low back pain

Three papers were found whose objective was to build ML models that predict back pain: Gaonkar (Gaonkar, 2017), Neto (Neto, 2011) and Mingle (Mingle, 2015). There are other studies to predict back conditions but they tend to classify MRIs rather than measurements from the spine.

Mingle states Support Vector Machines (SVM), a type of ML algorithm, tends to be used to model spinal problems the most and with the few studies available he was right as this algorithm was common to all three studies. Both Mingle and Neto talk about a concept known as the reject option. This is a method used to reduce misclassifying samples by rejecting data that does not have a strong bias for a specific class. Once rejected, a human can make the final decision as to which class the sample belongs to. This is a very useful idea to use in medical systems where the cost of misclassifying could, in some cases of health, be detrimental.

Gaonkar and Mingle are keen on feature selection which is a method used in ML to identify the features that contribute the most useful information in making a prediction.

The following table summarises the algorithms used in each of the three research papers (an 'x' indicates that it was used):

Author	SVM	NB	KNN	DT	RF
Gaonkar	x		x	x	x
Mingle	x	x			
Neto	x				

Table 1: ML algorithms used by other researchers to model back pain

Mingle and Neto use SVM with a KMOD kernel. KMOD (Kernel with MODerate decreasing) penalises input feature vectors that are far apart "while maintaining the closeness information from vanishing" (Ayat, 2001). Gaonkar doesn't state which SVM kernel is used in their research.

In order to decide which algorithms were to be used for building ML models, the literature search involved learning about the different types of ML algorithms capable of performing binary classification. Between the three papers, they cover popular algorithms. The following classification algorithms will be given a brief and simple explanation: Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LogR), K Nearest Neighbour (KNN), Naive Bayes (NB)

Machine Learning algorithms

Decision tree: Conceptually a DT consists of a root node, internal nodes and terminal leaves. The nodes represent features and the features that have the ability to cleanly separate the data into classes the best are higher up in the tree. For example, if a feature has values that mainly (or solely) belong to one class and a different set of values belong to the other class, then that feature separates the data very well and is considered to be quite pure. However, pure nodes are not usually the case and there will be an intersection of values which spill into both classes.

The root and internal nodes have conditions attached to them, similar to if-then statements in traditional programming, and as a result two pathways exist for a new data point to choose from as it traverses the tree. A decision as to which class a new data point belongs to is decided upon once it reaches a terminal node.

A DT is good in that, unlike logistic regression correlated features do not need to be eliminated (Le, n.d.). A DT is parametric and the depth (complexity) of a DT can be set using a hyper-parameter which is a parameter the person building the model can set.

Random Forest: RF can be described as a collection of DTs. Unlike DTs which decides from all the features which one is the best class separator, a random forest algorithm will randomly select several subsets of samples and features to build several trees. The RF result is based upon an aggregation of the results from all the DTs

Random forest classifiers tend to over fit less than DTs but depending on the number of features it can be slow to train because the more features there are, the more DTs there are. However, an RF can be fine tuned using hyperparameters to set the number of trees and features per tree. The Python class for Random Forests will also allow you to see the predictive power (importance of each feature). Random forest is a non-parametric classifier (Donges, 2018).

Naive Bayes: NB uses Bayes' rule and can work out the probability of new data belonging to a class and the probability of getting a particular class given a particular value for a feature. With that information it can then look at new data and decide which class it is most likely to belong to. It is naive because it naively assumes that the inputs are independent from each other. NB creates generative models and can be either parametric or non-parametric.

K-Nearest Neighbour: K nearest neighbour chooses the class a new data point belongs to by looking at the K number of points closest to it. It uses the data in the dataset at the point it is making a prediction and is non-parametric.

With each new data point a KNN model will search through the whole data set looking for the K samples closest in distance to the new point. Because KNN relies on the data at the time of prediction, they can use up a fair bit of memory. KNN doesn't always work well if there are many features in the dataset. It can become 'confused' as to what closeness means. If think of each feature as a dimension on a graph, it can get more difficult to interpret where that point is as the number of features (and therefore dimensions) that represent that point increases. KNN features should be standardised meaning that the features should be expressed using the same scale. Otherwise it can become insensitive to features with smaller scales.

Logistic Regression: LogR is popular for binary classification. Unlike linear regression it (which looks to predict a continuous number) it predicts a categorical variable. LogR can classify samples whose features are categorical or continuous data types. LogR decides on the class a data point belongs to by working out the probability of that point belonging to a class. Usually the classes e.g. "back pain" and "no back pain" will be re-assigned to 1 and 0. During training it can be seen which class a particular sample belongs to but with unseen data, the goal is to work that out. In order to do that there needs to be a way to link the sample of feature vectors to a probability of 1 or 0. This link is achieved by using the logit function.

The logit function can be described as the natural log of the odds. The odds is defined as:

probability of something being true (p)/ probability of that something not being true (1-p)

and the logit is therefore:

ln(odds) = ln(p/1-p) = logit(p)

It is known that logit(p) is equal to a linear combination of the sample feature values e.g.

logit(p) = a

where $a = w_0 + w_1*spine\_feature\_value_1 + w_2*spine\_feature\_value_2 + \ldots + w_n*spine\_feature\_value_n$

and

'w' are the regression coefficients. They are calculated using maximum likelihood estimation).

Algebra and rule of logs can be used to isolate p and get an estimation for it:

$p = e^a/(1+e^a)$

To help see where e^a came from:

$p/1-p = e^a$

With logR a probability threshold is chosen and if the probability of data belonging to a class is greater than a specified threshold then the data will be assigned to that class. LogR creates discriminative probabilistic models.

SVM: SVM is a binary classifier but it can still do multi-class classification by using methods known as one-vs-one or one-vs-rest. Because this project is interested in binary classification, these

methods will not be discussed but they could be useful for the proposed study mentioned earlier where a model would be produced to classify one of several spinal conditions.

An SVM uses the concept of a decision boundary to decide which class a data point belongs to and essentially separate data points belonging from one class from the data points belonging to the other class. The decision boundary is determined by a discriminant function of the form $f(x) = wX + b$ (the equation for a straight line). X represents the features' values of a sample (feature vector), w represents the coefficients (or weights) vector and b is where the decision boundary sits in relation to the origin when visualising the model graphically. X values are known but in their raw state they do not map directly to their associated class. Therefore the job of the training process is to work out the unknown w and b values so that mapping can take place. In order to find the best weight and bias values a mathematical method known as Lagrange is used.

The weights and bias for the decision boundary are chosen so as to allow $wX + b = 0$ to be fulfilled. A data point will be mapped to one of the sides depending on whether its representative feature values when applied to the $wX + b$ equation calculates to be > 0 or < 0. Basically, we only care if this calculation returns a negative or positive value and are not so interested in the value itself. If all points from one class successfully fall to one side of the boundary and the points for the other class to the other side then the data is said to be linearly separable. The training process attempts to have as wide a separation as possible between the classes by demarcating a margin. Ideally no points lie in this margin but in reality perfect linearly separable data is not likely and so SVMs have a concept of soft margin. The margin boundaries are defined by selected points from each class called support vectors. If the location of the support vectors change then so too does the size of the margin.

If the data is not linearly separable then the margin can be made to be “softer” meaning that it won’t be so strict about classifying the data perfectly. To decide on how strict the model should be, a soft margin parameter can be set which is often referred to as “C”. If C is set to a large value then the model will care a lot about making sure that all data points are classified correctly. This, however, can mean that noise is also captured and over fitting occurs resulting in poor generalisation. Setting a smaller C means that the model is not so strict about classing all values correctly but this means that there is likely to be better generalisation for new data points. A small C means that the errors matter less. C is also known as a hyperparameter which can be chosen by the person building the model whereas parameters w and b are calculated during the modelling process and not controlled by the programmer.

To pick the best hyperparameters, a technique known as cross validation is used to experiment with a different selection of hyperparameters on different subsets of the training data. By splitting the training dataset into subsets, the hyperparameters get a good run on effectively many different inputs of data.

An SVM can be modified to allow non-linearly data to be separated linearly by setting a what is known as a kernel trick. The kernel trick uses mathematical methods called kernels to essentially attempts to separate data in a higher dimension. Kernelling is the process of finding the space in which data can be separated into their respective classes by a plane. To visualise why this is the case, if imagine some red and blue balls resting on a flat surface. The red balls are arranged in a smaller inner circle and the blue balls in a larger outer circle. With this arrangement, the balls

cannot be separated by a straight line. However, if toss the balls in the air (assuming the balls stop and remain static) there will eventually be some throw that will result in the balls being arranged in space where they can be separated by a plane. At this point, they will be said to be linearly separable.

One of the SVM kernels is a radial basis function and it requires C to be set and another hyperparameter called gamma to be set. Gamma allows the model to define what the closeness of two points mean. A small gamma means that a new data point will be considered as part of the same class as a support vector even if it is quite a distance away from that support vector. A large gamma is less willing to consider 'far' points.

To select the best values for these hyperparameters a technique known as cross validation is used. Cross validation allows different values of hyperparameters to be tested on different samples of data. The hyperparameter(s) are chosen from the model that produced the best prediction result.

Algorithms chosen for the modelling process

The project used both SVM and LogR because they are popular algorithms for classification. By using an SVM with both a linear and RBF kernel the data could be tested for linear and non-linear separation and see which one performs the best. Also chose SVM so that the generalisation ability of the model could be compared with that in the research papers as part of this project's goal is to improve prediction accuracy. LogR and SVMs can be viewed as opposite with regards to their preference for number of features. SVMs cope well with a large number of features as they can find it easier to separate data whereas LogR can do well with less features. The twelve feature Kaggle

dataset is not considered to be high in dimensionality and therefore it would be interesting to see how an SVM model compares to LogR for those reasons.

The Project's Contribution Towards Research of Back Pain

This project contributes towards research that studies back pain. It will do so in two main ways. The first contribution describes a future research study that collects potentially good data for a dataset. The intention would be to use this dataset in an ML model that predicts spinal conditions. The second contribution is showing how the performance of an ML model, that predicts back pain, can be affected when good and bad input data is used and also how the model's predictive power can be improved.

Methodology

Overall, the methodology process involved doing a literature review and an analysis on a public dataset and it included a section for each of the contributions mentioned in the last chapter. The first section describes the process of developing a dataset for a proposal of a future study. The second section describes the procedures involved in analysing and implementing an ML model to predict back pain using data from a public dataset. The "Results and Analysis" chapter present the final dataset proposed for a future study and show the outcome of building models with the public datasets mentioned earlier.

Contribution 1: Creating a Dataset for a Future Study

Justification of data source

There are many different types of data that could be collected for input into a back pain prediction model. But with the author being a trained physical therapist, the obvious place is to collect data from physical therapy outpatient departments where data is abundant. This would include mainly physiotherapists as they make up the highest proportion of physical therapists in the UK. In 2017, there were 57,000 registered physiotherapists (Chartered Society of Physiotherapy, 2018) and approximately 5,200 osteopaths (General Osteopathic Council).

Physiotherapists are well known within the field of musculo-skeletal and orthodox medicine and they are fairly central with regards to the care of a patient with back pain. By 'central' it is meant that they have contact with many other clinicians involved in the patient's care. They are one of the

first people to see back patients, record significant details with regards to the complaint and they do not depend on inaccessible equipment to make a diagnosis.

Preparing for Research: General Considerations

Part of working out the feasibility of a study is costing the project. A study needs to factor in the people to be involved in the study (patients, clinicians, project supervisors, researcher). advertising the project, the length of the study, where to host, research equipment, amount and resource of funding and other such as reimbursement for patient travel. Other processes to consider are getting patients' permission, application to ethics committee, what the control should be (patients with back pain or without) and other such as getting clearance from health and safety for patient visits.

The data collecting process

If the proposed study discovered good features for a dataset to be used in predicting back conditions, then physiotherapists and osteopaths could be involved. However, for the study itself the doctor the patient is registered with (GP), a physical therapist and hospital consultant would be the main contacts for the research. The patient would be referred to a physiotherapist by a patient's doctor as normal and given typical treatment and advice. If either the physiotherapist or GP decide that a patient requires further investigation to either an orthopaedic consultant or neurologist, it could be at this point that they invite the patient to take part in the study. The study is only interested in the conditions mentioned in the literature review.

Further investigation usually means being sent for an MRI, CT or electro-diagnostics which happen to be the procedures used for reference standards (diagnosis confirmation) in research studies. Once a patient has confirmation of the results from a reference standard test and a document to confirm the presence or absence of the condition referred for, they would then come to the hosting

research centre to have various measurements done. These measurements would be the ones to be used for the dataset and the preference would be for one therapist doing the measurements partly for cost reasons but also this would mean having to only be concerned with the intra-reliability of measurements. The patient would come to the research centre regardless of a positive or negative result and the proven result would be entered as opposed to the original diagnosis and reason for referral. The dataset could also include the original diagnosis but this would be diluting the dataset and taking it away from the objective. A separate analyses could be done comparing the original and confirmed diagnosis.

The results of the measurements and tests would be entered electronically into a spreadsheet including the diagnosis. This would ideally be their only visit to the research centre. Reasons to not take measurements might be because:

- it is inappropriate for a patient's diagnosis and might aggravate the condition
- taking measurements might generate apprehension to the point of interfering with the test
- it might not be appropriate to do a test for a symptom that doesn't exist

An alternative approach to the study would be for the initial physiotherapist to take the measurements. The problem with this is that any physiotherapist that registered as referring part of the study would need to be assessed for their competence in taking measurements. This could involve several centres. There would also be a need for a data collection tool and way of securely sending the measurements to the research centre. Anyone working in the NHS knows that having this extra workload could be quite a burden. However, it could also save on 'other' costs such as patient travel and realistically, if such a model was to be brought into production it would be

different therapists taking measurements therefore introducing more realistic data to the modelling process.

With regards to control participants, a decision would have to be made as to whether they should be back pain patients (those diagnosed with NSLBP specifically) or people without back pain. The author's preference is to have patients with NSLBP due to wanting to compare this type of back condition with other spinal ones. It might also be more justifiable to perform reference tests on them because they actually have symptoms as opposed to running tests on non-symptomatic people.

Reference standards used to confirm diagnoses

The reference standard tests tended to be MRI or CT for conditions that would show a structural defect or compromise such as a compression of a nerve. Electro-diagnostics were mentioned if there was suspected nerve dysfunction. Instruments such as inclinometers, electrogoniometers and plumb lines were used by the clinicians to measure active spinal extension and rotation respectively but they were not specified as being reference standards.

Choosing which data features to collect

The dataset is to be used to diagnose specific conditions and not just either the presence or absence of back pain. A good dataset would contain many samples which represents the diverse population of back pain sufferers. For example, collecting data from a clinic that serves a local population where there is a high proportion of an ethnic group or age would be biased and could have poor generalisation if tested on other groups. Initially the data collected is unlikely to contain ideal features to predict back conditions and therefore will need to go through a process of analysis and feature selection to select the best predictors. The sections related to building models using a public dataset will discuss such methods.

Kappa (K) and Intra-Class Correlation (ICC) coefficients scores were assessed for the intra and inter-reliability of clinician measurements and positive and negative likelihood ratios were assessed for the accuracy of diagnostic tests in proving whether a condition exists or not. Sensitivity and specificity were also considered as evaluation methods but on their own they were not considered to be as informative as the likelihood ratio which takes into account both the number of correct and incorrect predictions made.

Using cancer as an example of why this is the case. If take positive to mean a person has cancer and negative that they do not, sensitivity tells us the number of correctly identified positive cases made out of all the people that have cancer whereas specificity calculates the proportion of correctly identified negative cases out of all of those that do not have cancer. However, a problem with these evaluation methods is that a therapist could get 100% sensitivity or specificity which looks good but

what this calculation doesn't show is the number of people that they thought had cancer but actually didn't (false positives) and the ones they thought didn't have cancer but in fact did (false negatives). The seriousness of this might not be so apparent with back pain which is rarely a life threatening condition. But if relying on a system to predict a back condition, wrong predictions could still leave a patient in considerable discomfort if the advice and treatment given contra-indicates their complaint.

As previously mentioned the likelihood ratio (LR) is calculated by using sensitivity and specificity. There are two versions of LR: positive (+LR) and negative (-LR).

+LR = sensitivity / 1- specificity
-LR = 1- sensitivity / specificity

K is most often used to measure categorical data and ICC for continuous data such as measuring the range of motion in a spine.

It is down to the person interpreting these measurements to decide what level of reliability and accuracy are acceptable. For the proposed dataset, a test or measurement was accepted if it had a moderate or higher status. Appendix A shows a table for K/ICC reliability score categories and another for LR but tests were included for the dataset if K/ICC was > 0.61 and >= 5.00 for LR. The final dataset chosen is discussed in the results and analysis section.

Contribution 2: Building a model with a public dataset

The project is looking at which dataset variables could be used to predict back pain and also show that ML could be useful as part of a clinical support decision system in a physical therapy or orthopaedic setting. The last section showed a possible process of choosing features for a dataset to be used in a ML model. This section will explain the stages involved in building a model from a dataset and fine tuning it to increase its prediction accuracy.

The Python programming language was used to help analyse and build the models (Appendix C provides information about the scripts). This language was chosen because it is well known for analysing data and producing ML models. But it is also the language the author is most familiar with.

About the public datasets: Kaggle and UCI

As mentioned in the literature review two datasets were used in this part of the project from two separate websites. Both datasets contain the same 310 samples of measurements taken from spinal x-rays. 210 of the samples come from people with back pain whereas the remaining 100 are from people without it. The difference between the two datasets is that the Kaggle one contains 12 variables whereas the UCI one contains 6. In fact, the UCI dataset is a subset of the Kaggle one. The features in the datasets are **pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius**, **grade of spondylolithesis,** pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope (these are explained more in the glossary). The names in bold are contained in both datasets. The spondylolithesis feature is measured in terms of percentage whereas

all the eleven others have angle of degrees for their units. All of the data for these variables are numeric data types.

Datasets used for classification contain an extra feature called the target variable which is categorical in nature. So the Kaggle and UCI datasets have 13 and 7 features respectively. Because the target feature can be one of two values (normal and abnormal back) a binary classifier was used to build the model.

The UCI website makes it clear where the dataset has come from. It was collected by a Dr. Henrique da Mota (Neto, 2011) who is a consultant specialising in orthopaedics and trauma. This is not the case with the extra six features in the Kaggle dataset and at this point confidence in part of the Kaggle dataset diminishes. Therefore the first process involved in building the model was checking the validity of the data.

Checking data validity

The data was validated by drawing histograms, quantile-quantile plots, box plots and doing a literature search for expected normal ranges of the anatomical features. These aforementioned graphs and checks helped to identify any distributions and outliers. Raw data usually has outliers and a lot of populations follow a normal distribution. Many ML algorithms assume that the individual features follow a Gaussian distribution (Brownlee, 2016b). Another way that the data was validated was through building a model. If a model cannot make good generalisation, then this could indicate that the input data does not represent the target feature.

Pairwise graphs were also plotted to identify any linear relationships between variables (correlations). Algorithms like logistic regression can be affected by highly correlated variables (Rajarajan, 2014). If these features also change value by similar amounts to each other then they are collinear. Highly correlated variables can be counteracted by using a regularisation method and the Python method for Logistic Regression allows a regularisation parameter to be set.

Preparing the data for analysis

Imbalanced versus balanced

Earlier it was mentioned that the datasets have 210/100 split of abnormal/normal back samples. Having an imbalance of classes can cause problems when a ML algorithm is training a model because it biases the predictions towards the more dominant class. There are several ways that this can be addressed. The method used in this project was to oversample the non-dominant class and undersample the dominant class. To oversample, a copy of existing samples belonging to the non-dominant class was added to a copy of the 310 sample dataset. To undersample, several samples were taken away from a copy of the original dataset. This resulted in three different datasets: an augmented, reduced and original dataset with 420, 200 and 310 samples respectively. The modelling process experimented with all three datasets.

Choosing the optimum dataset size for training

In the proposed study section, it was mentioned that a good dataset would contain “many” samples but it is never clear what “many” means. For a model to be able to generalise about unseen data it needs enough data to learn from. To get a better idea of a good dataset size this project used a learning curve. The learning curve plotted the prediction accuracy for various size datasets.

Standardising the data

Some algorithms work better when the data has been standardised (rescale the data so that it has a mean of 0 and standard deviation of 1). Algorithms "might behave badly if the individual features do not more or less look like standard normally distributed data" (scikit-learn, 2017). The SVM algorithm, for example, "assumes that all features are centered around zero".

Building the model

First, the dataset was split into data for training (70% data) and data for testing (30% data to be used once the model was built) for whichever of the original, reduced or augmented datasets were being trained at the time. The percentage split was roughly based on the results from the learning curves. The Python sklearn library was used to program the models by initialising a SVM or LogR regression object. An array was set with various C penalty values and cross validation (see literature review) was performed on each C and the prediction accuracy for each outcome compared. The C value associated with the best performing model was chosen to create a final model and then tested with the data set aside.

Evaluating the model

Evaluating a model means calculating its prediction accuracy and there are several ways of doing this. Accuracy and F1 were used. Accuracy calculates the number of correctly identified predictions out of the total number of samples fed into the model. F1 uses precision and recall in its calculation whereby for this project F1 score = (precision x recall) / (precision + recall)

Results and Analysis

Contribution 1: The resulting dataset for a proposed study

The first part of the project was about proposing a way of getting useful features for a dataset to be used to build a ML model that predicts back pain. It was suggested that the data would come from measurements typically taken by physical therapists and the model was to predict one of the following conditions: lumbar spinal stenosis, lumbosacral radiculopathy, ankylosing spondylitis, disc herniation, lumbar nerve root compression and non-specific low back pain.

Findings: measurements of accuracy and reliability

The Kappa (K) and Intra-class correlation (ICC) scores the intra and inter reliability of measurements and the positive Likelihood Ratio (+LR) scores the accuracy of a test's ability to prove the presence of a condition. Tables 2 and 3 show the scale used to judge measurements and tests (Shrout (cited in Cleland, 2016)):

K/ICC	**Level of reliability/agreement**
0.0-0.1	none
0.11-0.40	slight
0.41-0.60	fair
0.61-0.80	moderate
0.81-1.0	substantial

Table 1: Categories of measurement reliability - Kappa and Intra-class correlation coefficient

And for accuracy:

+LR	-LR	Amount of proof that the test will correctly identify/rule out the condition
>10.0	<0.1	Good
5.0 – 10.0	0.1 – 0.2	Moderate
2.0 – 5.0	0.2 – 0.5	Small
1.0 – 2.0	0.5 – 1.0	Poor

Table 2: Measure of accuracy - Likelihood Ratio

K and ICC scores were chosen if they indicated “moderate” and “substantial” and +LR if it fell into the “moderate”. No test fell into the “good” category. With regards to conditions being tested for, there were no tests that were good enough identify ankylosing spondylitis or disc herniation.

While looking at tests, there was an occasional odd case where the reliability of a measurement was better when tested on one side of the body but poor when tested on the other. For example, there was more inter-reliability at identifying pain over the left joint in the back (0.73 - left joint versus 0.52 for the right). In cases like this, the test was not included as it is important to have reliability for both sides.

The Final Dataset

Appendix A shows details about the dataset but in summary, the final dataset would include 27 features (including the target variable). Approximately 60% of the features would be categorical and 40% numeric data types. The target variable would be categorical therefore a classification algorithm would be used to build a predictive model. However, it would need to be a technique that is capable of multi-class classification. Because some algorithms are designed to work as binary

classifiers, techniques such as one-versus-one and one-versus-rest could be used as a way of transforming multi-class problem into a binary one.

The dataset consists of results from tests that are considered reliable and accurate for testing functional and anatomical pain, nerve integrity, neuro-vascular compromise, ranges of movement, motor control, stiffness and inter-vertebral instability. However, what should have also been included are tests that were accurate and reliable enough to rule out conditions (-LR < 0.1). Using a test to rule out a condition could be useful if, for example, applying a test to confirm a condition was too painful for a patient.

With regards to the reliability of measurements, there were plenty of tests and measurements that had good inter and/or intra reliability. However, there were no tests that were considered to have 'good' proof for accuracy of a condition. Most accuracies were within the moderate range.

The author realises that this is not the definitive dataset but considers it to be a good start because the dataset would consist of measurements that can be done along a typical care pathway for patients with back pain therefore potentially lots of data and there would be some confidence in the accuracy and reliability of the data.

Contribution 2: The resulting back pain prediction models

Findings from checking the validity of the datasets

With regards to building a model using a public dataset, the data was validated using histograms, scatter graphs, box plots and quantile-quantile plots. Because the datasets use the terms “abnormal” (back pain) and “normal” (no back pain) for the target feature, they will also be used from this point onwards. The UCI dataset will be referred to whenever talking about the six shared features, kaggle-12 will mean the full twelve feature dataset and kaggle-6 to mean the six unique features in the Kaggle dataset.

Histograms and Kernel Density Plots: Distributions and Outliers

Histograms and kernel density plots were produced for each of the twelve features. Illustration 1 to 6 shown below are from the UCI dataset and the final six histograms are the additional features in the Kaggle dataset. The histograms show the proportion of samples of that feature associated with abnormal (red) and normal (green) backs.

Features common to both the UCI an Kaggle datasets:

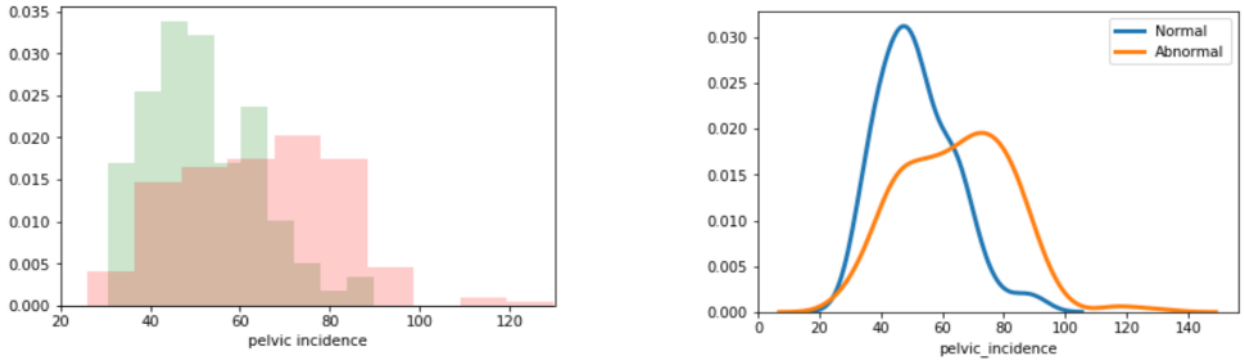


Illustration 1: Histogram and kernel density plots for pelvic incidence

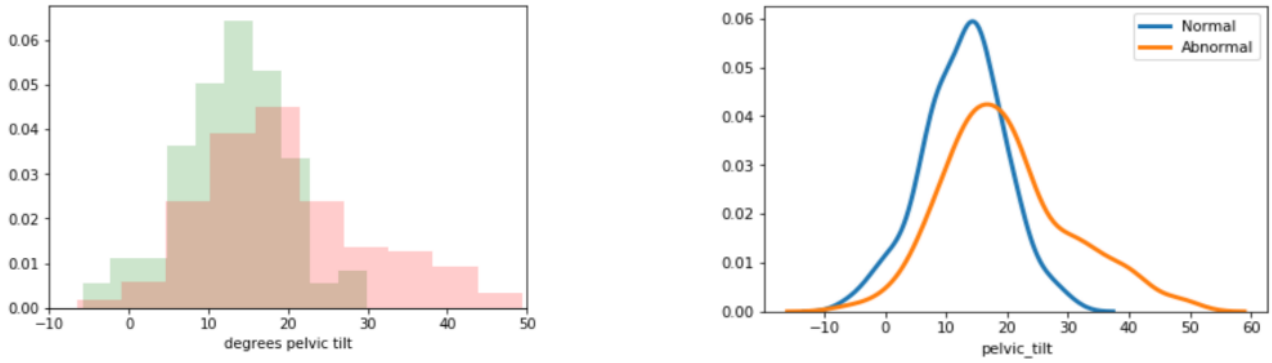


Illustration 2: Histogram and kernel density plots for pelvic tilt

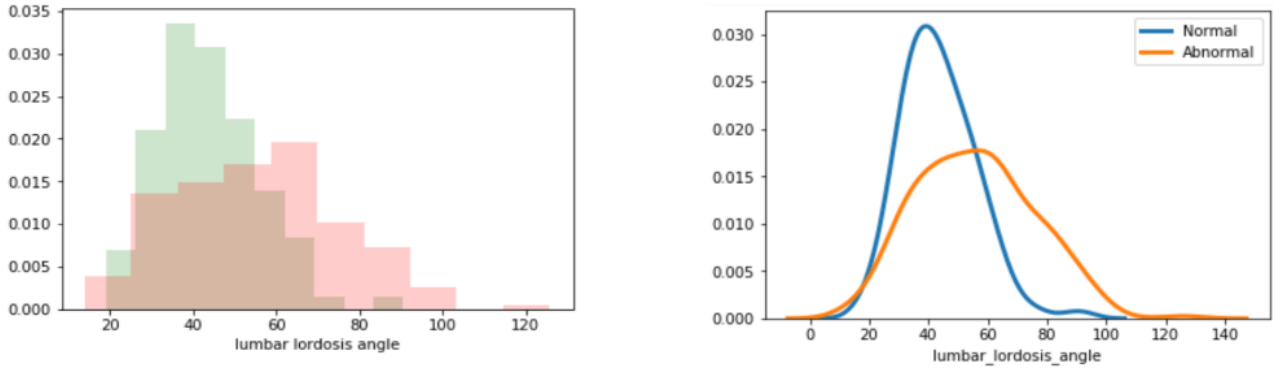


Illustration 3: Histogram and kernel density plots for lumbar lordosis angle

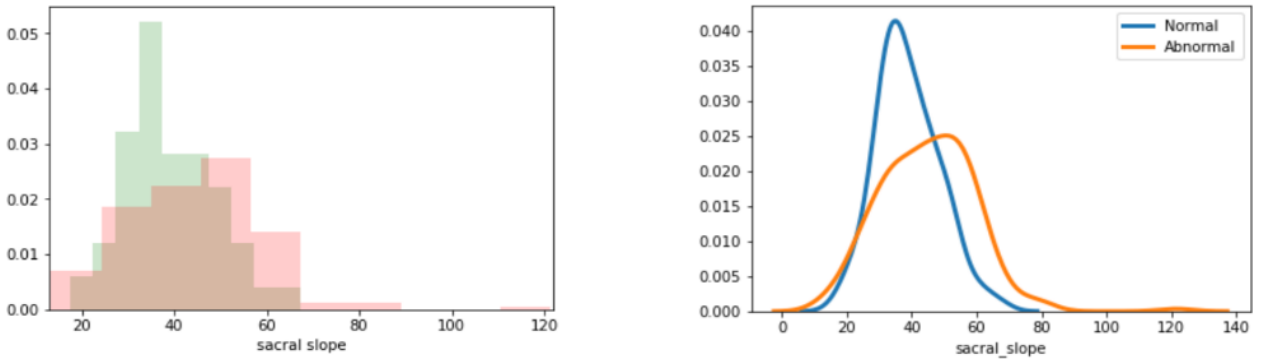


Illustration 4: Histogram and kernel density plots for sacral slope

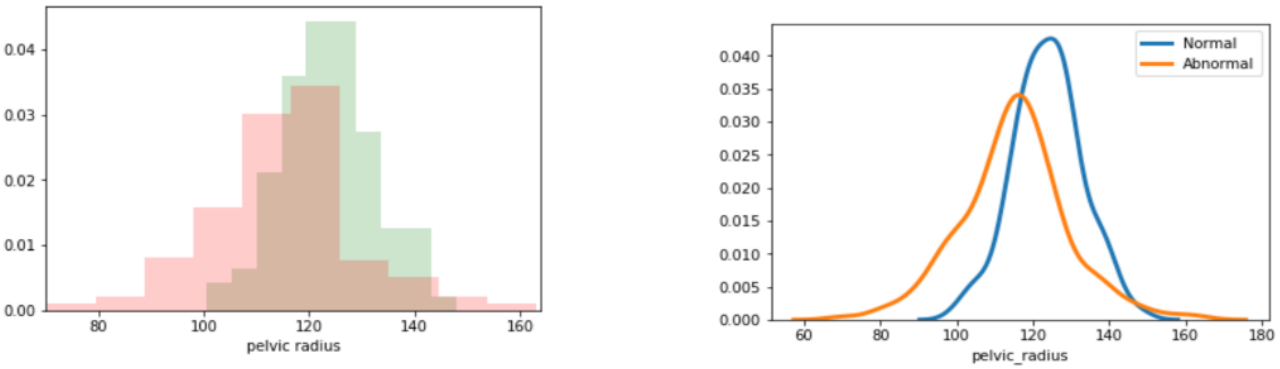


Illustration 5: Histogram and kernel density plots for pelvic radius

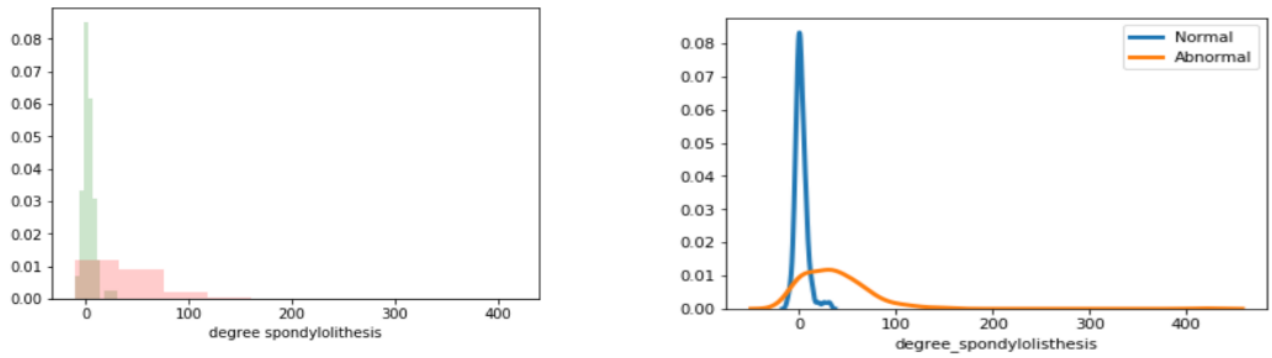
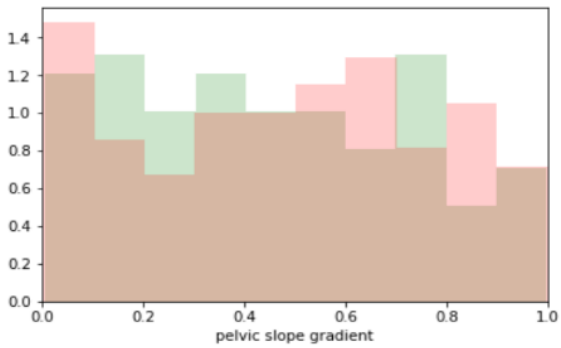


Illustration 6: Histogram and kernel density plots for degree spondylolithesis

Features unique to the Kaggle dataset:



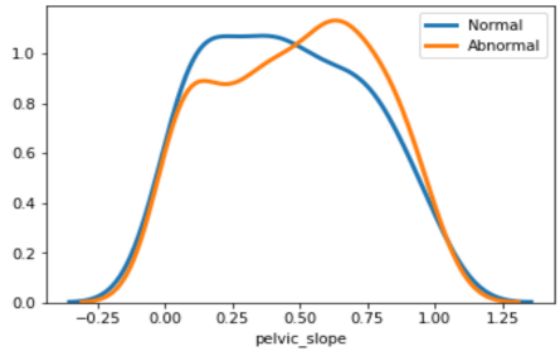
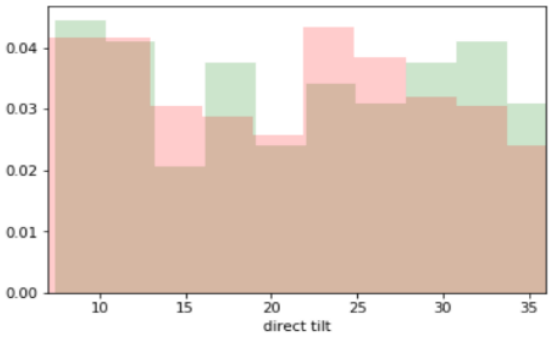


Illustration 7: Histogram and kernel density plots for pelvic slope



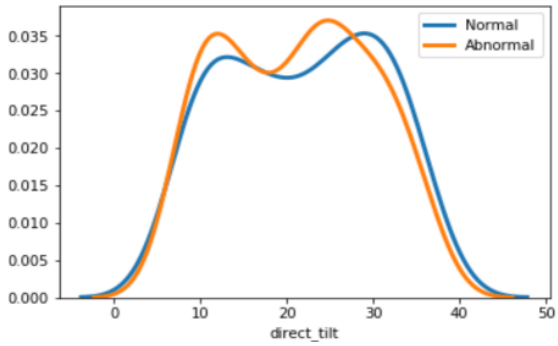
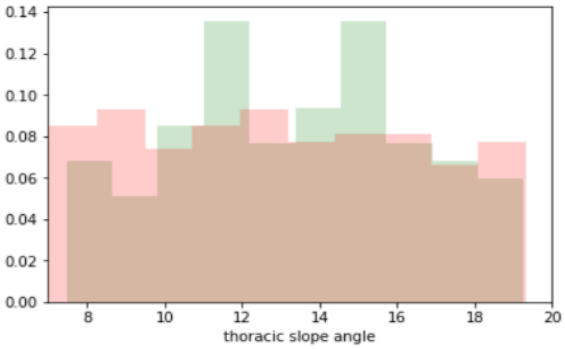


Illustration 8: Histogram and kernel density plots for direct tilt



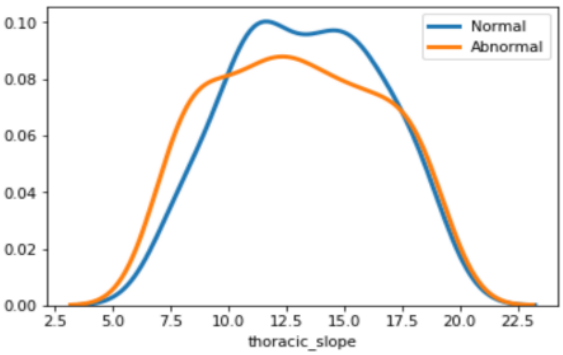


Illustration 9: Histogram and kernel density plots for thoracic slope

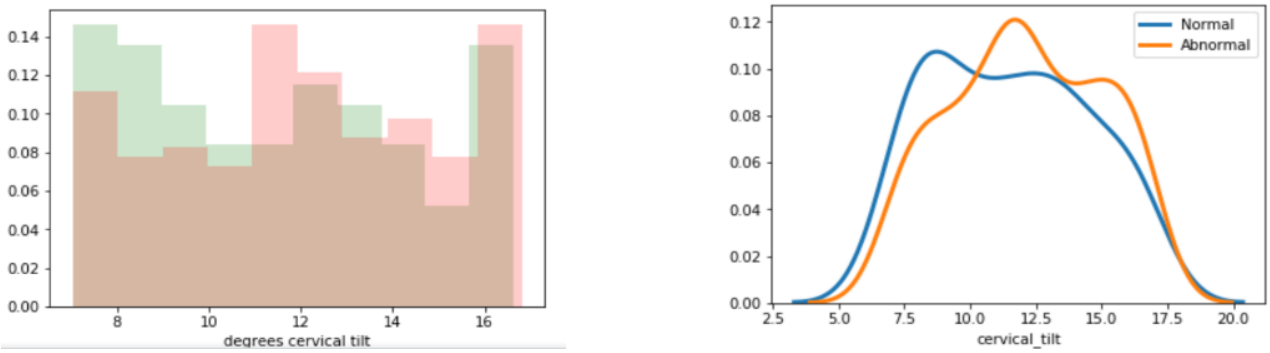


Illustration 10: Histogram and kernel density plots for cervical tilt

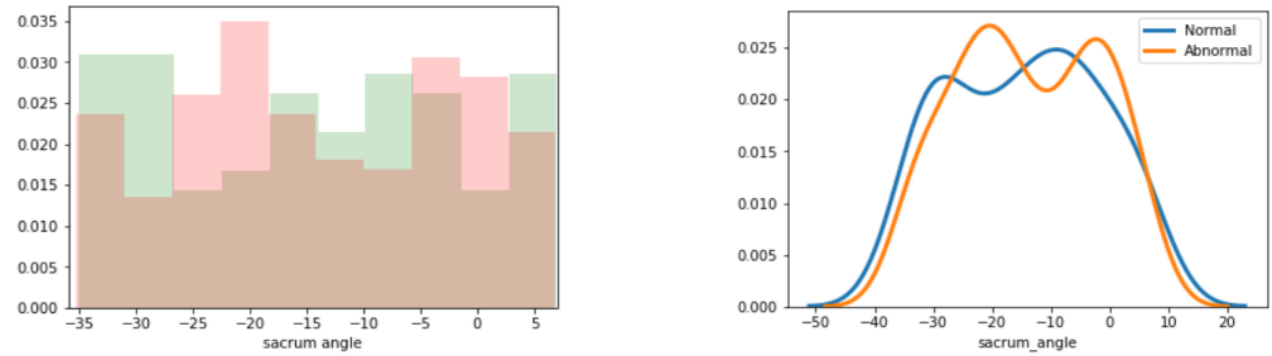


Illustration 11: Histogram and kernel density plots for sacrum angle

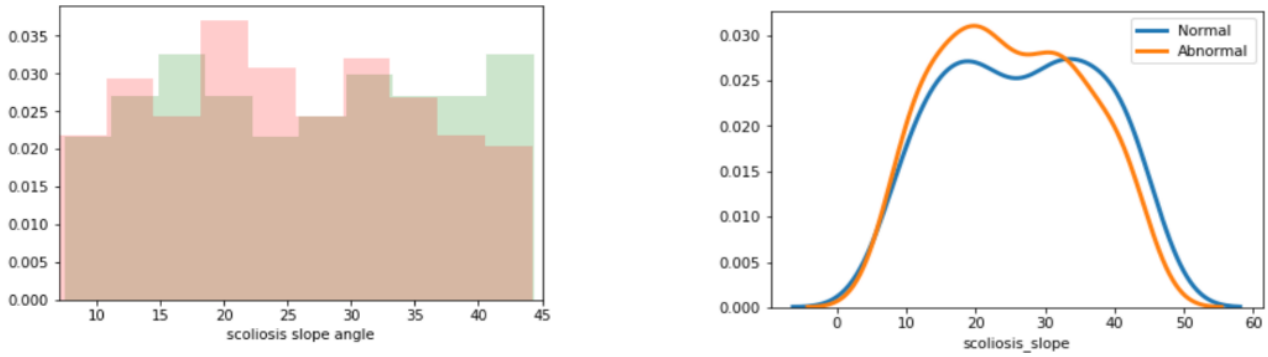


Illustration 12: Histogram and kernel density plots for scoliosis slope

The features of the UCI dataset show a rough normal distribution for both the abnormal (red) and normal data (green) in the histograms. The kernel density plots give a bit more insight into the distribution of the extra Kaggle features. They do not seem to follow any distribution with the bin sizes used. The kernel density plots show two peaks suggesting a bimodal distribution. If the data is valid, this is suggesting that there could be two different populations of data for each class. Gaonkar (Gaonkar et al, 2017) used the Kaggle dataset to create a back pain model and they report degree of spondylolithesis as being the strongest predictor. This is no surprise when looking at the spondylolithesis graph because there is quite a strong distinction between data belonging to each of the classes.

There are other ways of establishing whether data follows a normal distribution. A quantile-quantile plot (qqplot) is a method that can be used to check if a distribution is normally distributed or not (Brownlee, 2018). The points in a qqplot should ideally form along a 45 degree angle line if the data is normally distributed as shown in illustration-13 below for pelvic radius:

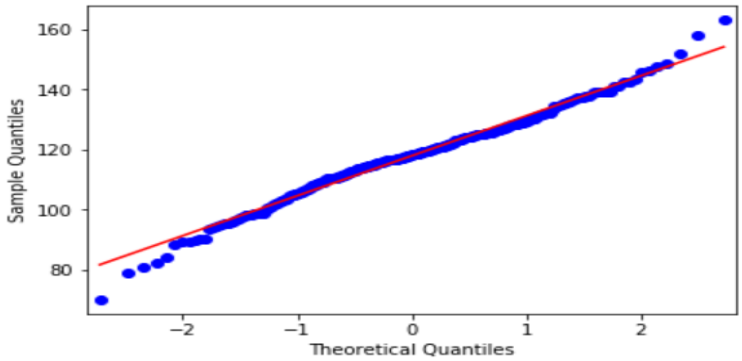


Illustration 13: Quantile-quantile plot: Positive for a normal distribution (pelvic radius)

Compare this to direct tilt:

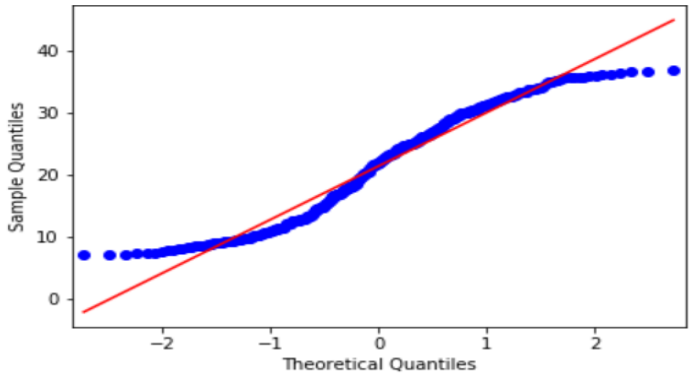


Illustration 14: Quantile-quantile plot: Negative for a normal distribution (direct tilt)

The Kaggle features had similar graphs to the one for direct tilt (illustration-14) again indicating that the six extra features do not follow a normal distribution and might be better fit with a non-parametric algorithm.

What is not so easy to see on the graphs are outliers therefore box plots were drawn to help identify them.

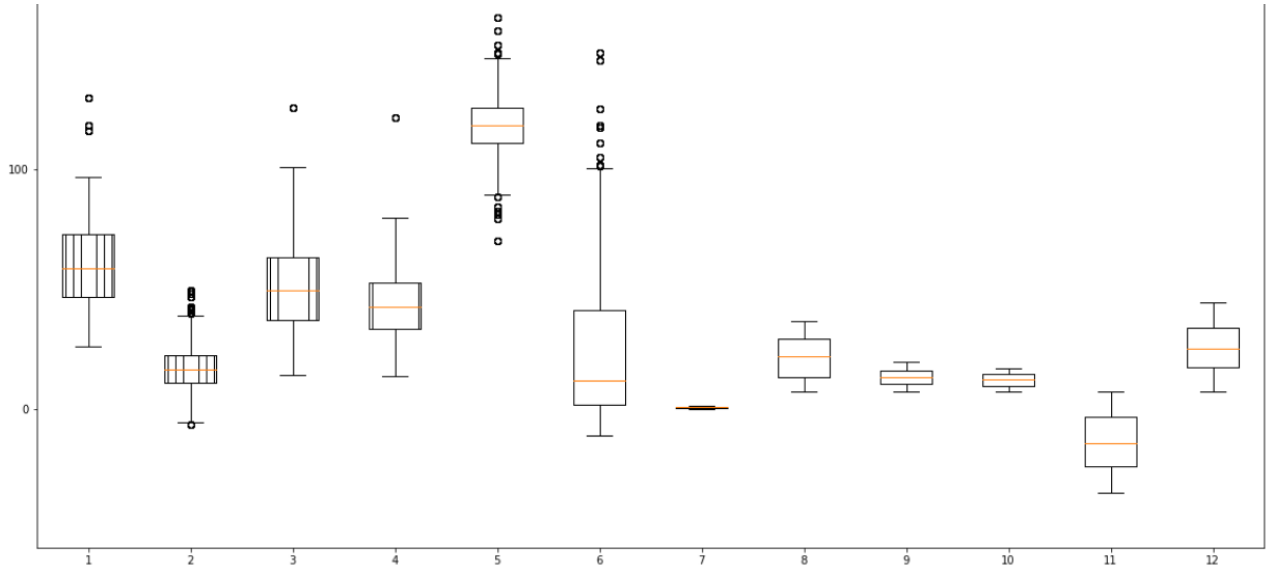


Illustration 15: Box plots for Kaggle-12

(1-pelvic incidence, 2-pelvic-tilt, 3-lumbar lordosis angle, 4-sacral slope, 5-pelvic radius, 6-degree spondylolithesis, 7-pelvic slope, 8-direct tilt, -9-thoracic slope, 10-cervical tilt, 11-sacrum angle, 12-scoliosis slope)

Not all outliers are shown due to the scale of the diagram but what stands out is that the features in the Kaggle dataset have no outliers. It is unusual to have raw data without noise.

The data was also analysed by looking at the range (minimum to maximum values). The table-4 below shows the range of normal and abnormal values for each feature.

Feature name	**Features**	**Dataset samples belonging to normal class**	**Dataset samples belonging to abnormal class**	**Normal range according to literature**
		(in degrees of angle except for degree spondylolithesis)		
Lumbar lordosis angle	UCI	*19.1 - 76.0 (19.1 - 90.6)	*14.8 – 100.7 (14.0 - 125.7)	21.1 – 45.3 (Lin, 1997) 51.3 – 62.0 (Been, 2014)
Pelvic incidence	UCI	30.7 - 89.8	*26.2 - 96.6 (26.2 - 129.8)	22 – 75 (Tyrakowski, 2015)
Pelvic tilt	UCI	-5.8 - 29.9	-6.6 - 49.4	6.2 - 19.8 (Donzelli, 2012) -12 – 30 (Tyrakowski, 2015)
Pelvic radius	UCI	100.5-147.9	70.1-163.1	128 - 146mm (Sergides, 2011)
Sacral slope	UCI	17.4 - 67.2	13.4 - 121.4	31.1 - 46.9 (Donzelli, 2012) 22.5 – 50.9 (Deinlein, 2013)
Degree of spondylolithesis	UCI	-11.1% – 31.2%	*-10.7% – 93.6% (-10.7% - 418.5%)	0 -100% (Neto, 2011) normal would be no slippage
Cervical tilt	Kaggle-6	7.0 - 16.6	7.0 - 16.8	11.4 - 24.6 (Lee, 2015)
Direct tilt	Kaggle-6	7.4 - 36.6	7.0 - 36.7	No range found
Thoracic slope	Kaggle-6	7.5 - 19.3	7.0 - 19.3	20.4 – 33.0 (Janusz, 2015)
Pelvic slope	Kaggle-6	0.0 - 1.0	0.0 - 1.0	No range found
Scoliosis slope	Kaggle-6	7.4 - 44.3	7.0 - 44.2	No range found
Sacrum (sacral) angle	Kaggle-6	-35.1 - 7.0	-35.1 - 7.0	21 – 44.5 (Gilliam, 1994)

Table 3: Actual and expected ranges of the features

* suspected outliers removed (original range)

But looking at data in this way does not make outliers obvious. Below is a box plot for degree spondylolithesis:

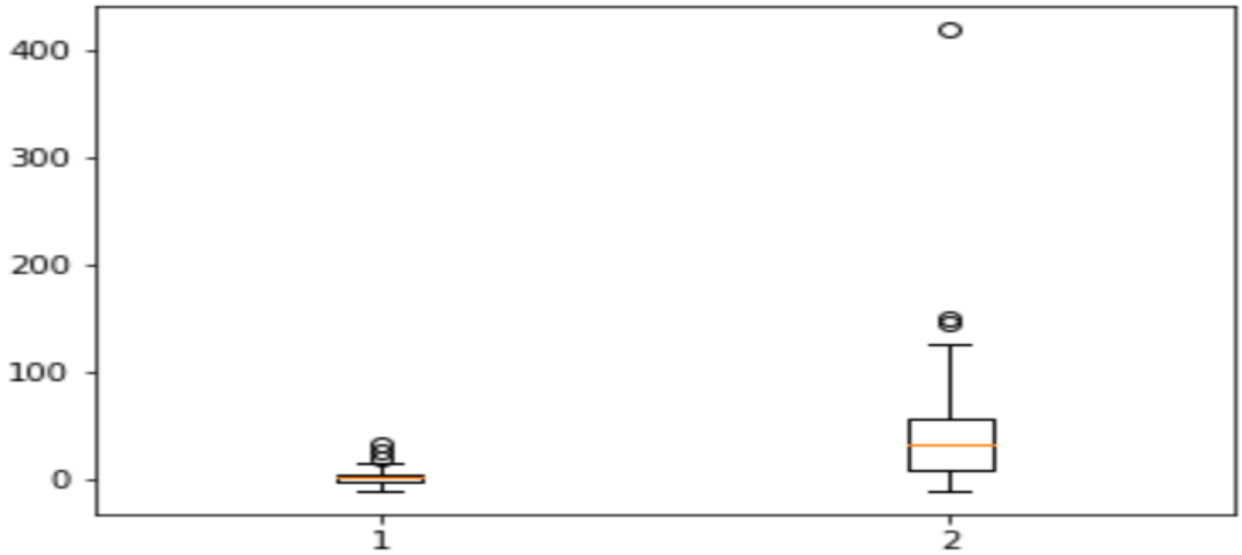


Illustration 16: Box plot for degree spondylolithesis – example of outliers

With regards to looking for the normal ranges in a literature search, "slope" was difficult to find because it is a new concept (Negrini, 2012). Other terms have alternative names. For example "neck and "cervical" are used interchangeably in anatomy and different ranges were found on both "neck tilt" and "cervical tilt". The range for cervical tilt was chosen because it matched more closely the range found in the dataset. "Direct tilt'"could not be found. Lumbar lordosis angle measurements can be measured from different points on the spine (Been, 2014) so the angles in literature could vary. The lumbar spine consists of five vertebrae (L1 – L5) and each vertebra has a top and bottom surface known as end plates. Some people take the lumbar lordosis angle as starting from the top end plate of L1 to the bottom end plate of L5, whereas others will, as Been points out, "measure lordosis starting as high as T10, others finish at L3" (T10 being the 10th thoracic vertebra of which there is a total of 12).

Feature correlation and collinearity

Part of deciding on whether features of a dataset could be helpful to train a ML model involved looking at the correlation between features. If there are features that are highly correlated in a high dimensional dataset, it can be useful to remove some of the variables especially if they do not offer additional help in making a prediction. The Kaggle dataset is not high dimensional but there were some features that were considered to be highly correlated. The pairwise scatter graphs show that pelvic incidence and sacral slope are highly correlated with a Pearson's correlation = 0.81 (-0.5 or greater than 0.5 are considered to have good correlation).

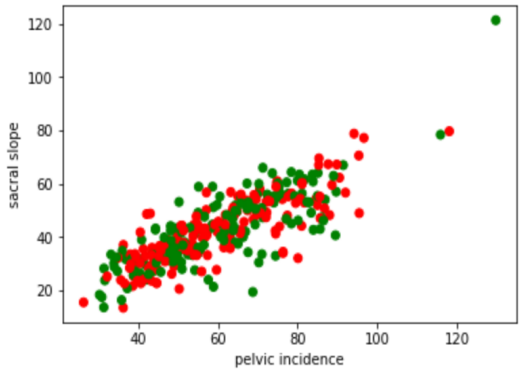


Illustration 17: Correlation between pelvic incidence and sacral slope

Anatomically, pelvic incidence is known as a morphological parameter which means that once someone has reached skeletal maturity the angle (in this case) does not change. However, pelvic incidence = pelvic tilt angle + sacral slope angle and these individual components can change. So it is no surprise that pelvic incidence and pelvic tilt also have a good correlation (Pearson's = 0.63)

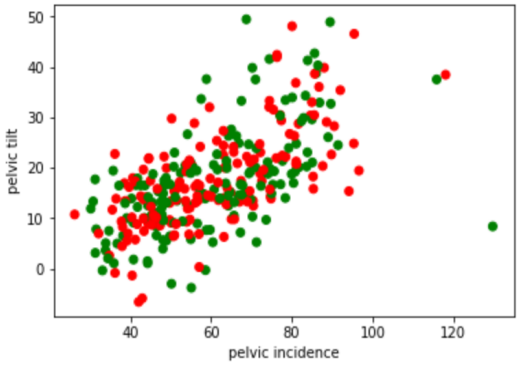


Illustration 18: Scatter plot for pelvic incidence and pelvic tilt

Neto (Neto, 2011) seems to put a fair bit of importance on pelvic incidence in terms of its biological behaviour by stating that the pelvic "incidence angle determines a normal condition" suggesting that they would not remove this parameter from the dataset. However, because it is a variable that can be calculated from its components, removing it is a consideration. Although any of the three variables could be removed and calculated by the other two, it would could be interesting to investigate which of the pelvic tilt or sacral slope angles influences the model more.

The pairwise correlations involving the extra Kaggle features demonstrated no correlations (high or otherwise). Examples of the pairwise graphs are shown in illustrations 19 and 20.

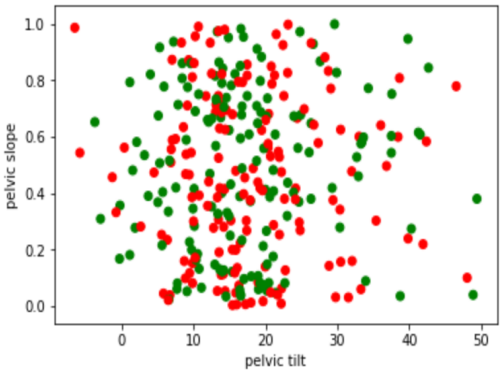


Illustration 19: Scatter plot for pelvic tilt and pelvic slope

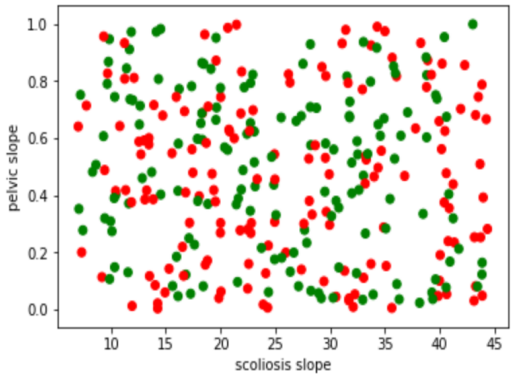


Illustration 20: Scatter plot for scoliosis slope and pelvic slope

Outcome of evaluating the dataset size

In order for a model to offer generalisation it needs enough data. But what does 'enough' mean? Models were created on different size datasets and their prediction accuracy were compared and visualised using learning curve graphs. Illustration-21 shows the learning curves for the reduced, augmented and original datasets:

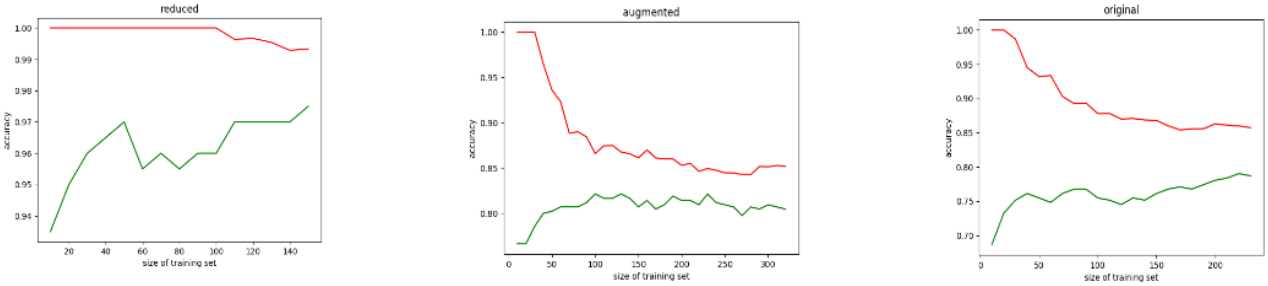


Illustration 21: Learning curves for reduced, augmented and original datasets, respectively

The augmented learning curve shows that the testing and training outputs are at their best approximately between 225 to 250 samples. After that, it looks as if the test error might start to get larger again suggesting generalisation is getting worse. So with a 310 sample dataset, 70% of the data should be used for training (225/310=0.72) and 30% data as input data for testing. Both the original and reduced dataset learning curves suggest that their models would generalise better if there was more data because the curves look as if they still want to converge.

The resulting model

The following tables show the prediction accuracy of models built using the Kaggle (Kaggle-12) and UCI datasets and alone with the six extra Kaggle (Kaggle-6) features. For each one of these datasets the effect of using the original 310, 200, and 420 samples from the original, undersampled and oversampled datasets is also included.

Accuracy scores for the SVM model

SVM (Linear) – Evaluation method = accuracy	UCI	Kaggle-6	Kaggle-12
Original	C: 0.021 training: 0.86 test: 0.82	C: 0.001 training: 0.66 test: 0.71	C: 0.002 training: 0.86 test: 0.84
Reduced	C: 0.004 training: 0.98 test: 0.96	C: 0.002 training: 0.54 test: 0.52	C: 0.001 training: 0.93 test: 0.93
Augmented	C: 0.010 training: 0.86 test: 0.86	C: 0.005 training: 0.52 test: 0.53	C: 1.000 training: 0.86 test: 0.83
Best performing dataset:accuracy score using unseen data	**Reduced: 0.96**	**Original: 0.71**	**Reduced: 0.93**

Table 4: Accuracy score for SVM (linear)

SVM (RBF) – evaluation method = accuracy	UCI	Kaggle-6	Kaggle-12
Original	C: 0.215 gamma: 0.001 train: 0.87 test: 0.82	C: 0.001 gamma: 0.000001 train: 0.66 test: 0.71	C: 0.460 gamma: 0.001 train: 0.85 test: 088
Reduced	C: 0.215 gamma: 0.01 train: 0.99 test: 0.95	C: 0.010 gamma: 0.000001 train: 0.51 test: 0.47	C: 1.000 gamma: 0.0001 train: 0.98 test: 0.97
Augmented	C: 1.000 gamma: 0.001 train: 0.87 test: 0.86	C: 0.460 gamma: 0.001 train: 0.57 test: 0.48	C: 0.460 gamma: 0.001 train: 0.87 test: 0.84
Best performing dataset:accuracy score using unseen data	**Reduced: 0.95**	**Original: 0.71**	**Reduced: 0.97**

Table 5: Accuracy score for SVM (RBF)

Overall the linear SVM doesn't perform any better than the RBF SVM. According to Hsu (Hsu et al, 2016) even though the linear and RBF kernels are using different methods to separate the data

“the linear kernel is a special case of RBF” and that the same results can be obtained for different C values for a linear kernel and when combined with gamma in a RBF kernel.

In general the reduced dataset allows for better models to be built and this could be because it has a balance of classes with unique samples. Whereas because the oversampled dataset contains copies of the original data, this could mean that the same data that was used for training could have over spilled into the test datasets. It is important that the test data uses unseen samples. This way the model is truly tested for its generalisation. The cross validation training data, cross validation test data, and the data that was put aside for ultimate testing with unseen samples could all potentially contain some same samples. Therefore, over sampling should have taken place after the cross validation process by over sampling the data that was used for cross validation. The test data, which was put aside when the dataset was split into train and test should be unique and untouched until ready for ultimate testing.

Because the histogram distributions for kaggle-6 showed little signs of a normal distribution, it is not surprising that a parametric model performs significantly worse than for the Kaggle dataset. However, even with a non-parametric model (RBF SVM) there is no improvement which again increases doubt over these extra features. This doubt is strengthened by the fact a model built using all twelve features doesn’t perform any better than when the UCI dataset is used alone. This implies that the kaggle-6 features offer very little to the modelling of back pain.

Illustrations 22-24 show a visualisation of model performance during testing for different values of C using different features - UCI, Kaggle-12 and Kaggle-6. Illustrations 25-27 show the mean scores for the same features but with different C and gamma values.

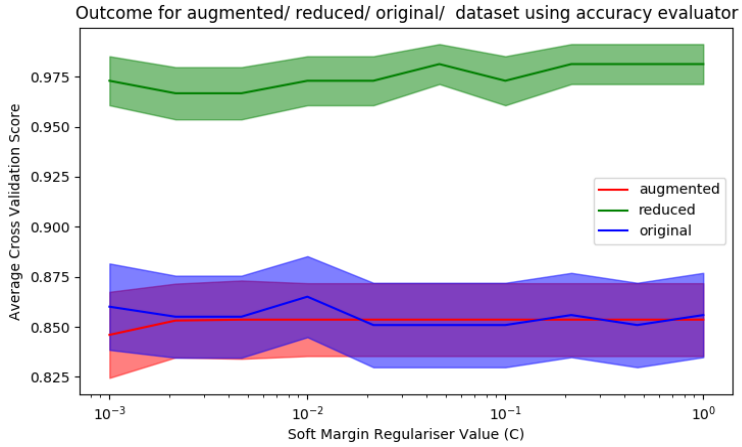


Illustration 22: Mean accuracy score for the UCI dataset

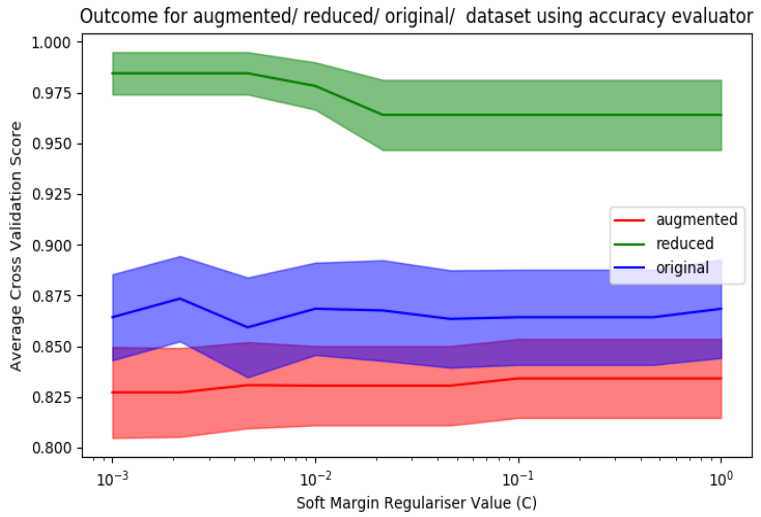


Illustration 23: Mean accuracy score for Kaggle-12

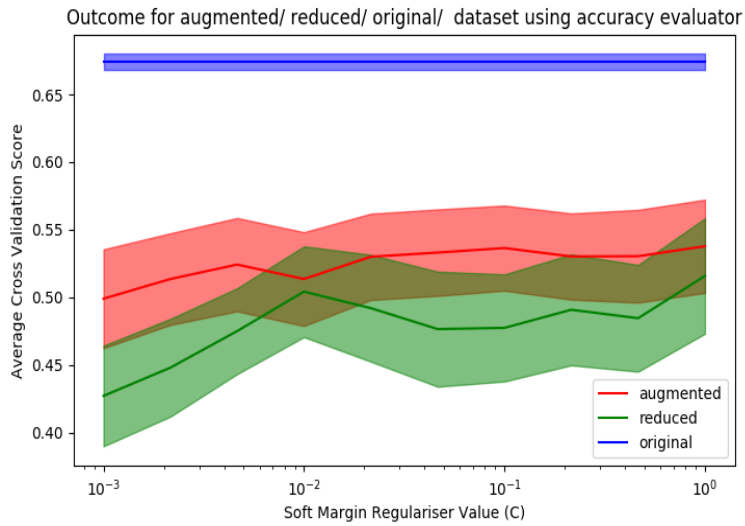


Illustration 24: Mean accuracy score for Kaggle-6

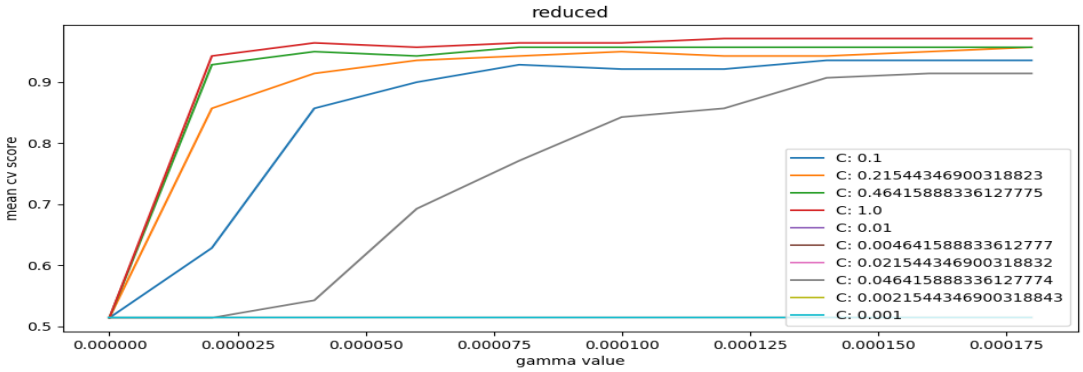


Illustration 25: SVM (RBF) - accuracy scores for various gammas – reduced dataset

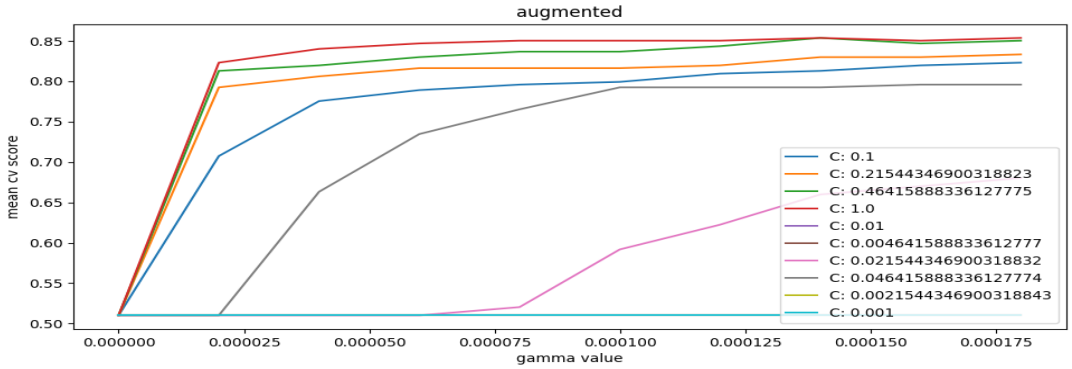


Illustration 26: SVM (RBF) - accuracy scores for various gammas – augmented dataset

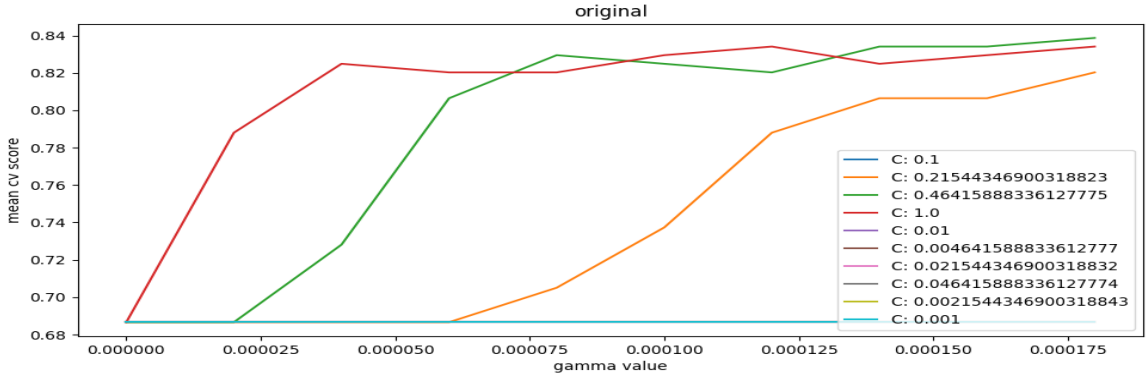


Illustration 27: SVM (RBF) - accuracy scores for various gammas – original dataset

Comparing the results to Gaonkar (Gaonkar et al, 2017) who used the twelve feature dataset in their research, they obtained 85% accuracy for their SVM model. However, they do not specify which kernel they used.

Accuracy scores for the logistic regression model

Logistic Regression (***L1 regularisation***)	UCI	Kaggle-6	Kaggle-12
Original	C=1.000 train: 0.83 test: 0.85	C=0.002 train: 0.66 test: 0.72	C=1.000 train: 0.83 test: 0.85
Reduced	C=0.002 train: 0.98 test: 0.97	C=0.050 train: 0.53 test: 0.52	C=0.001 train: 0.98 test: 0.98
Augmented	C=0.005 train: 0.84 test: 0.85	C=0.215 train: 0.60 test: 0.49	C=0.050 train: 0.84 test: 0.80
Best performing dataset:accuracy score using unseen data	**Reduced: 0.98**	**Original: 0.72**	**Reduced: 0.98**

Table 6: Accuracy scores for logistic regression using the different sets of features (L1 regularisation)

Logistic Regression (***L2 regularisation***)	UCI	Kaggle-6	Kaggle-12
Original	C=0.005 train: 0.83 test: 0.84	C=0.001 train: 0.69 test: 0.64	C=1.000 train: 0.82 test: 0.82
Reduced	C=0.001 train: 0.98 test: 0.95	C=0.010 train: 0.52 test: 0.48	C=0.005 train: 0.98 test: 0.97
Augmented	C: 0.001 train: 0.85 test: 0.84	C=1.000 train: 0.57 test: 0.53	C=0.022 train: 0.85 test: 0.81
Best performing dataset:accuracy score using unseen data	**Reduced: 0.98**	**Original: 0.64**	**Reduced: 0.98**

Table 7: Accuracy scores for logistic regression using the different sets of features (L2 regularisation)

The logistic regression models didn't perform any better than the SVM ones. One of the differences between the Python SVM and logistic regression methods is that the logistic regression method allows regularisation methods to be specified that help to reduce over fitting. A separate model was created for each of the regularisation methods L1 (Ridge regularisation) and L2 (LASSO regularisation). They are useful because they penalise a feature if it is not useful for the model. LASSO regularisation not only reduces the importance of a feature but it will set it to zero effectively eliminating the feature completely from the modelling process.

F1 scores for the SVM model

The accuracy score is not recommended for imbalanced datasets. So models were evaluated with the F1 score.

SVM (Linear) (evaluation method = F1)	UCI	Kaggle-6	Kaggle-12
Original	train = 0.89 test = 0.87	train= 0.81 test= 0.82	train= 0.89 test= 0.86
Reduced	train= 0.98 test=0.96	train= 0.36 test= 0.48	train= 0.98 test= 0.95
Augmented	train = 0.86 test = 0.81	train= 0.51 test= 0.45	train= 0.84 test= 0.83
Best performing dataset:accuracy score using unseen data	**reduced: 0.96**	**original: 0.82**	**reduced: 0.95**

Table 8: F1 scores using SVM (linear) on the different sets of features

Overall the results weren't much different to when the models were evaluated with the accuracy score. However, there is an improvement when a model is produced from kaggle-12 but again because the scores for the UCI features and all 12 features are very similar, this suggests that the

Kaggle extra features do not have much impact on predicting back pain. It has not gone unnoticed that the kaggle-6 dataset performed better with the original dataset for all algorithms used.

Model accuracy obtained from other researchers

The table below compares the results obtained from the Gaonkar (Gaonkar, 2017) and Mingle (Mingle, 2015) papers:

SVM (linear)	Gaonkar paper	Mingle paper	dissertation
UCI (original)	-	85%	82%
Kaggle-12 (original)	85.5%	-	84% (88% for SVM RBF)
UCI (undersampled)			96%
Kaggle-12 (undersampled)			93% (97% for SVM RBF)

Table 9: A comparison of results between other studies that have used the same datasets

Summary of results and analysis

In summary, the extra features found in the Kaggle dataset cannot be trusted because they show no particular distribution or correlation with other features, there are no outliers which is unusual, but the fact that reference values for a normal range cannot be found for most of the six extra variables doesn't give much trust in the data.

Discussion

This dissertation set out to investigate machine learning and its application to predicting back pain and conditions. It approached this in mainly two ways. The first task was to look for a means of getting good data and the second approach involved demonstrating the effect of having good and bad data as input for a machine learning task. Overall, the dissertation set out to investigate clinical features that could be used in an ML model and how reliable these models can be in predicting back pain.

Whilst researching for data from measurements and tests that could be used in a dataset for an ML task, the main considerations were looking for a source of data that would be reasonably accessible, abundant, representative of people with back pain and obtained reliably and accurately. Physical therapy outpatient departments was considered to potentially fit all of the criteria.

Tests and measurements for conditions and dysfunction associated with the lumbar region were chosen based on the highest reliability and accuracy scores. However, none of the measurements and tests from the Cleland (Cleland, 2016) resource were scored within the highest categories. Reliability (a measure of clinicians' ability to get the same measurement consistently and correctly) has five categories of confidence ranging from "None" to "substantial' saw some measurements being scored with the highest grade. Accuracy (a measure of a test's ability to prove the presence of a condition) failed to have any tests in the highest of its four categories. The best tests scored "moderate" and so moderate and good measurements and moderate tests were chosen as acceptable features for a dataset.

The final dataset suggested for the proposed study would really be in a raw form and doesn't guarantee that a good prediction model could be built from it. It would have to be analysed and the best features would need to be selected and engineered. Usually a therapist would ask questions other than what is in the suggested dataset. For example, is there pain on leaning backwards (extension) and not just record the amount of extension they have. Also knowing a patient's history of activities is important. Again as an example, spondylolithesis is higher in those that take part in certain sports because of the hyper-extension involved (Watkins, 2010). Therefore this is important information that should be captured by a dataset.

An alternative approach to picking the "best" data for a dataset is to use brute-force meaning that all the data recorded by physiotherapists could be included and then apply feature selection and engineering as needed. But that would be like using uncalibrated machines to measure objects. Therefore a protocol for measurements and tests is recommended.

The other part of the project set out to demonstrate that ML can be used to reliably predict back pain. Reliable here refers to whether a model can be built from data and consistently discriminate between samples from those with back pain and those that are pain free. Ultimately, can it be used to predict back pain? This dissertation shows that it can. What perhaps wasn't emphasised enough to the reader is how important cross validation is in choosing the best hyperparameters therefore increasing the chances of generalisation. Cross validation also gives confidence that the model can perform well with different sets of data. Also how it is important to compare the training accuracy score with the test score and that both a high *and* similar score is obtained for training and test. A

discrepancy between train and test scores could indicate that the training phase learns about the noise as well as the true data and as a result the model fails to generalise enough for new unseen data. A phenomenon known as over fitting.

The project also set out to improve accuracy scores of the prediction model by comparing it to scores obtained in the Mingle and Gaonkar papers. This dissertation demonstrated how the prediction accuracy could be improved by under sampling the dataset. Under sampling means training will not be biased towards the dominant class and therefore increase the chances of better generalisation. The better model can be built from using a balanced dataset in the training process.

Given time this dissertation could have used models trained from different classification algorithms and incorporated a reject option that is a popular concept in medicine. The reject option is a method that can be used to limit mis-classifications by having the option to reject a sample if there is not a strong enough indication that a sample belongs to one class or another. The system would flag a rejected sample with the expectation of human intervention.

Other methods of oversampling could have been used such as SMOTE (Synthetic Minority Over-Sampling Technique). The synthetically produced data would produce unique and similar data to the dataset and therefore address the possible problem encountered with the replacement method used in this dissertation.

There were some unexpected results such as getting similar results for both the linear and non-linear SVM models but this is more a reflection of the author's lack of understanding with regards to the details of how the SVM algorithm works underneath.

At the very beginning of the dissertation, alternative datasets were sought. Fortunately a contact was made and there is an opportunity to use a dataset that includes spinal measurements from a fluroscopy (x-ray movie) study. The dataset will be used with the intention of understanding NSLBP better.

Conclusion

This dissertation set out to ask if ML can be used to predict back pain and investigate clinical features that could be used to do so. A suggestion was made as to how good clinical features could be found but until the proposed study is carried out, their ability to produce a good model cannot be tested. But it was highlighted that the dataset is likely to be limited and not necessarily any better than using more of a brute-force method instead of careful selection. The dissertation also showed that models with higher prediction accuracy can be built and highlights how important the pre-processing, selection, and engineering of data is before building the final model.

The work included in this document has contributed by suggesting ways good data could be collected for understanding back pain better (a condition that is s substantial problem in society). It demonstrated some techniques of how prediction models can be improved and how important it is for the programmer building these models to understand machine learning at a deeper level and not just be able to call pre-defined functions from Python libraries.

As for future studies, there is potential for the author to model a dataset that includes spinal measurements from fluroscopy (x-ray movie) with the intention of understanding NSLBP better. NSLBP is main bugbear of back pain because the underlying cause is unknown and it is usually chronic in nature therefore very difficult to treat and manage. The fluroscopy project will look at inter-vertebral motion control which is something that cannot be done in outpatient settings. As NSLBP is considered more of an issue, maybe this is where the focus of back pain research should be.

Bibliography

Akben, S. (2016). *Importance of the shape and orientation of the spine and pelvis for the vertebral column pathologies diagnosis using machine learning methods*. Available at: https://www.alliedacademies.org/articles/importance-of-the-shape-and-orientation-of-the-spine-and-pelvis-for-the-vertebral-column-pathologies-diagnosis-with-using-machine-.pdf [Accessed 8 Aug 2018]

Altini, M. (2015). Dealing with Imbalanced Data: Undersampling, Oversampling and Proper Cross Validation. Available from: https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation. [Accessed 18 Aug 2018]

Ayat, N., Cheriet, M., Remaki, L., Suen, C. (2001). KMOD – a new support vector machine kernel with moderate decreasing pattern recognition. Application to digit image recognition. Proceedings of Sixth International Conference on Document Analysis and Recognition. p1215-1219

Bambrick, N. (2016) *Support Vector Machines: A Simple Explanation*. Available from: https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html [Accessed 3 Aug 2018]

Barreto, G., Neto, A., da Mota Filho, H. (n.d.) Vertebral Column Data Set. Available from: https://archive.ics.uci.edu/ml/datasets/Vertebral%20Column [Accessed on: 15 Aug 2018]

Been, E., Kalichman, L. (2014) Lumbar Lordosis. The Spine Journal. Vol 14 (1), pp97-97

Bernardone, C. (n.d) *Machine Learning: Understanding Feature Selection.* Available from: https://www.nastel.com/blog/machine-learning-understanding-feature-selection [Accessed 5 Aug 2018]

Breivik, H., Collett, B., Ventafridda, V., Cohen, R., Gallacher, D. (2006). Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. European Journal of Pain Vol.10 pp287-333

Brent, L. (2018). Ankylosing Spondylitis and Undifferentiated Spondyloarthropathy Differential Diagnoses. Available at https://emedicine.medscape.com/article/332945-differential. [Accessed 4 Aug 2018]

Brooks-Bartlett, j. (n.d.). *Probability Concepts Explained: Maximum Likelihood Estimation* Available at: https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1 [Accessed 2 Sep 2018]

Brownhill, K. (2007). Back pain and homeostatic requirements of the spinal system. International Journal of Osteopathic Medicine. Volume 10, Issue 1, p 18-23

Brownlee, J. (2016a) *Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning.* Available at: https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning. [Accessed 4 Aug 2018]

Brownlee, J. (2016b) *Visualize Machine Learning Data in Python With Pandas*. Available at: https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas [Accessed 11 Aug 2018]

Brownlee, J., (2017). *What is the difference between a parameter and a hyperparameter?* Available at: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter [Accessed 20 Aug 2018]

Brownlee, J. (2018). *A Gentle Introduction to Normality Tests*. Available from: https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/ [Accessed 13 Aug 2018]

Buchbinder, R., Tulder, M., Őberg, B., Costa, L., Woolf, A., Schoene, M., Croft, P. (2018). Low back pain: a call for action. The Lancet. Vol 391 (10137) p2384-2388

Chartered Society of Physiotherapists (2018) Promoting Physiotherapy. Available at: https://www.csp.org.uk/professional-clinical/promoting-physiotherapy) [Accessed 27 Aug 2018]

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research Vol. 16 pp321-357

Cleland, J., Koppenhaver, S., Su, J. (2016). Netter's orthopaedic clinical examination: An evidence based approach, 3rd Edition. Philadelphia: Elsevier

Coursera. (2018). *What is machine learning.* Available at: https://www.coursera.org/lecture/machine-learning/what-is-machine-learning-Ujm7v [Accessed 2 Sep 2018]

Deinlein, D., Bhandarkar, A., Vernon, P., McGwin, G., Wall, K., Reece, B., McKay, J., Theiss, S. (2013). Correlation of Pelvic and Spinal Parameters in Adult Deformity Patients With Neutral Sagittal Balance. Spine Deformity Vol.1 pp 458-463

Donges, N. (2018). The Random Forest Algorithm. Available at: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd [Accessed 20 Aug 2018]

Donzelli, S., Zaina, F., Negrini, S. (2012). Sagittal and pelvic parameters analysis in patients with adolescent idiopathic scoliosis. Scoliosis. Vol. 7 (suppl 1) pO15-O15

Fayaz, A., Croft., Langford, R., Donaldson, L., Jones, G. (2015). Prevalence of chronic low back pain in the UL: a systematic review and meta-analysis of population studies. Available from: https://bmjopen.bmj.com/content/bmjopen/6/6/e010364.full.pdf [Accessed 23 Aug 2018]

Gaonkar, A. et al. (2017). Classification of lower back pain disorder using multiple machine learning techniques and indentifying degree of importance of each parameter. International Journal of Advanced Science and Technology. Volume. 105, p11-24

General Osteopathic Council. (2018). Training and Registering. Available at: https://www.osteopathy.org.uk/training-and-registering [Accessed 27 Aug 2018]

Gilliam, J., Brunt, D., Macmillan, M., Kinard, R., Montgomery, W. (1994). Relationship of the pelvic angle to the sacral angle: measurement of clinical reliability and validity. The Journal of Orthopaedic and Sports Physical Therapy. Vol. 20 (4), pp. 193-199

Hellard, B., Hopping, C., Walker, D., Fearn, N. (2018) What is machine learning? IT Pro

Huang, M., Hung, Y., Liu, D. (2014). *Diagnostic Prediction of vertebral column using rough set theory and neural network technique*. Available from: http://docsdrive.com/pdfs/ansinet/itj/2014/874-884.pdf [Accessed 8 Aug 2018]

Hsu, W., Chang, C., Lin, J. (2016). A Practical Guide to Support Vector Classification. Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed 14 Aug 2018]

Imrie, R. (2016). Fibromyalgia. InnovAiT. Vol.10(1), pp.45-50

Janusz, P., et al. (2015). Influence of cervical spine position on the radiographic parameters of the thoracic inlet alignment. European Spine Journal. Vol. 24, Issue 12, pp2880-2884

Kaggle. (2016). Lower Back Pain Symptoms Dataset. Available from: https://www.kaggle.com/sammy123/lower-back-pain-symptoms-dataset [Accessed: May 2018]

Kassirer, J. (1989). Our stubborn quest for diagnostic certainty: a cause of excessive testing. New England Journal of Medicine. Volume 320 No. 22 p1489-1491

Le, J. (n.d.). At Tour of The Top 10 Algorithms for Machine Learning Newbies. Available at: https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11 [Accessed 18 Aug 2018]

Lee, S., Son, E., Seo, E., Suk,K., Kim, K. (2015). Factors determining cervical spine sagittal balance in asymptomatic adults: correlation with spinopelvic balance and thoracic inlet alignment. Vol.15(4), pp705-712

Lewis, R., Williams, N., Matar, H., Din, N., Fitzsimmons, D., Phillips, C., Jones, M., Sutton, A., Burton, K., Nafees, S., Hendry, M., Rickard, I., Chakraverty, R., Wilkinson, C. (2011). Background. In: *The clinical effectiveness and cost-effectivenss of management strategies for sciatica: systematic review and economic model*. Winchester: Health technology assessment. Vol.15(39) p5

Lima, M., Neto, M., Zuiani, G., Veiga, I., Tebet, A., Pasqualini, W., Landim, E., Cavali, M. (n.d.) Parameters for the evaluation of cervical sagittal balance in idiopathic scoliosis. Coluna/Columna. Vol.16(1), pp38-41

Lin RM, Jou IM, Yu CY. (1992) Lumbar lordosis: normal adults. J Formos Med Assoc. Mar;91(3):329-33

Loong, T. (2003). Understanding sensitivity and specificity with the right side of the brain. British Medical Journal. Volume 327 Issue 7417 p716-719

McKenna, F. (2010). Spondylolithesis. Arthritis Research UK. Available at: https://www.arthritisresearchuk.org/health-professionals-and-students/reports/hands-on/hands-on-spring-2010.aspx [Accessed on 21 Aug 2018]

(The) McKenzie Institute International. (n.d.). *The McKenzie Method: How it all began* Available at: http://www.mckenzieinstitute.org/about-us/the-legend-of-the-mckenzie-method/ [Accessed 2 Sept 2018]

Macaulay, T. (2016). *Progress towards a paperless NHS*. Available from: https://search-proquest-com.libproxy.ucl.ac.uk/docview/1852995458?rfr_id=info%3Axri%2Fsid%3Aprimo [Accessed 8 Aug 2018]

Matheny, M., Ohno-Machado, L. 2nd ed., (2014). Generation of Knowledge for Clinical Decision Support: Statistical and Machine Learning Techniques. In: *Clinical Decision Support: The Road to Broad Adoption.* Amsterdam; Boston: Elsevier, p. 310-311

Mayo, M. (2016). *Support Vector Machines: A Concise Technical Overview*. Available from: https://www.kdnuggets.com/2016/09/support-vector-machines-concise-technical-overview.html [Accessed 3 Aug 2018]

Mellor, F. (2014). An evaluation of passive recumbent quantitative fluroscopy to measure mid-lumbar intervertebral motion in patients with chronic non-specific low back pain and healthy volunteers. PhD. School of Health & Social Care, Bournemouth University.

Merrill, R., Kim, J. S., Leven, M., Kim, J. H., Meaike, J., Bronheim, R., Suchman, K., Nowacki, D., Gidumal, S., Cho, S. (2018). Differences in Fundamental Saggital Pelvic Parameters Based on Age, Sex, and Race. Clinical Spine Surgery. Vol.31(2), ppE109-E114

Mingle, D. (2015) A Discriminative Feature Space for Detecting and Recognizing Pathologies of the Vertebral Column. Available from: file:///home/hope/Downloads/a-discriminative-feature-space-for-detecting-and-recognizing-pathologies-of-the-vertebral-column-2090-4924-1000114.pdf [Accessed on 13 Aug 2018]

Negrini, S., Atanasio, S., Donzelli, S., Zaina, F. (2012). “Slopes”: a new approach to scoliosis radiographic measurement and evaluation, related to the horizontal plane in a bodily view. [online video] Available at: https://www.youtube.com/watch?v=7Aypc4Owup4 [Accessed 18 Apr 2018]

Neto, A. (2011). Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option in Vitrià, J., Sanches, J., Hernández, M. (eds), Pattern Recognition and Image Analysis 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Springer, Berlin, Heidelberg, p588-595|

NHS England. (2016). *The Forward View Into Action: Paper-Free at the Point of Care. Guidance for Developing Local Digital Roadmaps*. Available from: https://www.england.nhs.uk/digitaltechnology-old/wp-content/uploads/sites/31/2016/11/develp-ldrs-guid.pdf [Accessed 8 Aug 2018]

NICE. (2009). Low Back Pain: Costing Report. Available at: http://webarchive.nationalarchives.gov.uk/20090706100332/http:/www.nice.org.uk/nicemedia/pdf/CG88CostReport.pdf [Accessed 18 Aug 2018]

NICE. (2016). Low Back pain and sciatica in overs 16s: assessment and management. Available at https://www.nice.org.uk/guidance/ng59/evidence/full-guideline-assessment-and-noninvasive-treatments-pdf-2726158003 [Accessed 30 July 2018]

Park, C., Allaby, M. (2013). A Dictionary of Environment and Conservation. 2nd Ed. Oxford University Press

Press, G. (2013). A Very Short History Of Big Data. Available from: https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#248c611c65a1 [Accessed on 13 Aug 2018]

Project Jupyter. (2018). Available at: http://jupyter.org [Accessed 8 Aug 2018]

Rajarajan, J. (2014). Does the performance of logistic regression get adversely affected by highly-correlated features? [Blog] *Quora*. Available at: https://www.quora.com/Does-the-performance-of-logistic-regression-get-adversely-affected-by-highly-correlated-features [Accessed 10 August 2018]

Reddy, S. (2012). *Classification of vertebral column using Naïve Bayes technique*. Available from: https://pdfs.semanticscholar.org/7288/f591d5421d704fce17605129e16b144e9b08.pdf [Accessed 8 Aug 2018]

Sadeghi-Naini, M., Taghipour, S., Savadkouhi, A., Kuzyk, P., León, S., Ghazavi, M., Abolghasemian, M. (2018). Hemi-pelvic slope is correlated with the acetabular depth in adults – a radiological study. Skeletal Radiology. Vol. 47(8), pp119-1125

scikit-learn. (2017). Preprocessing data. Availabel at: http://scikit-learn.org/stable/modules/preprocessing.html [Accessed 28 August 2018]

Sergides, I., McCombe, P., White, G., Mokhtar, S., Sears, W. (2011). Lumbo-pelvi lordosis and the pelvic radius technique in the assessment of spinal sagittal balance: strengths and caveats. European Spine Journal. Vol. 20 (S5) pp. 591-601

Simpson, R., Gemmell, H. (2006). Accuracy of spinal orthopaedic tests: a systematic review. Chiropractic & Osteopathy [online] Available at: https://chiromt.biomedcentral.com/articles/10.1186/1746-1340-14-26 [Accessed 30 July 2018]

Skymind. (n.d.) *Datasets and Machine* Learning. Available from: https://skymind.ai/wiki/datasets-ml [Accessed 3 Aug 2018]

Snodgrass, S., Rivett, D., Robertson, V. (2006). Manual Forces Applied During Posterior-to-Anterior Spinal Mobilisation: A Review of the Evidence. Journal of Manipulative and Physiological Therapeutics. 29 (4), 316-329

Stralen, K,. Stel, V., Reitsma, J., Dekker, F., Zoccali, C., Jager, K. (2009). Diagnostic methods 1: sensitivity, specificity, and other measures of accuracy. Kidney International. 75, pp1257-1263

Tyrakowski, M., Hailong, Y. Siemionow, K. (2015). Pelvic incidence and pelvic tilt measurements using femoral heads or acetabular domes to identify centers of the hips: comparison of the two methods. European Spine Journal. Vol.24(6) pp1259-1264

Unal, Y., Polat, K., Kocer, H. (2014). Pairwise FCM based feature weighting for improved classification of vertebral column disorders. Computers in Biology and Medicine. Vol. 46, pp61-70

Unal, Y., Polat, K., Kocer, H. (2016). Classification of vertebral column disorders and lumbar discs disease using attribute weighting algorithm with mean shift clustering. Measurement Vol.77.pp278-291

Unison (2018) Back Pain. Available from: https://www.unison.org.uk/get-help/knowledge/health-and-safety/back-pain/ [Accessed 22 Aug 2018]

Ushewokunze, S., Abbas, N., Dardis, R. (2008). Spontaneously disappearing lumbar disc protrusion. British Journal of General Practice Vol 58(554). p646-647

Waldman, S. (2009). Spinal Stenosis. In: *Pain Review*. Elsevier, p.302-303

Watkins, R. (2010). Lumbar Spondylolysis and Spondylolithesis in Athletes. Seminars in Spinal Surgery

Woolfson, T. (2008). Synopsis of Causation: Spondylolithesis (incorporating Spondylolysis and Spondyloptosis) Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/384554/spondylolisthesis.pdf [Accessed 25 Aug 2018]

Appendix A – Proposal for a dataset

The following table was put together using Cleland (Cleland, 2016) by selecting tests and measurements depending on their associated reliability and accuracy.

Test/Measurement	**Data type**	**Type of measurement**	**Possible values**	**Accuracy/reliability accepted if +LR >5 and intra/inter-reliability >0.6**	**Notes**
Pain sitting	Categorical - nominal	Presence of pain	Yes/No	+LR = 6.60	Indicates patient has lumbar spinal stenosis Reference standard: CT/MRI
Patellar reflex	Categorical - nominal	Muscle reaction or joint movement	Absent/Present	+LR = 6.29 – 7.14	Indicates patient has lumbar radiculopathy (L3/4 nerve) Reference standard: Electro-diagnostics
Medial hamstring reflex	Categorical - nominal	Muscle reaction	Absent/Present	+LR = 5.07	Indicates patient has lumbar radiculopathy (L5 nerve) Reference standard: Electro-diagnostics
Slump knee bend test	Categorical - nominal	Reproduction of neural pain	Yes/No	+LR = 6.00	Indicates nerve root compression Reference standard: MRI
Two stage treadmill test	Categorical -nominal	Longer time for pain to come on when walking on incline	Yes/No	+LR = 6.46	Indicates neurogenic claudication Reference standard: CT/MRI
Forward flexion	Numeric – ratio	Range of movement		Using distance fingertip distance from floor: Intra-reliability = 0.91-0.95 Inter-reliability = 0.77-0.99 Using an inclinometer:	

Test/Measurement	Data type	Type of measurement	Possible values	Accuracy/reliability accepted if +LR >5 and intra/inter-reliability >0.6	Notes
				Inter-reliability = 0.60-0.74	
Lateral flexion	Numeric – ratio	Range of movement		Intra-reliability: Left side bending=0.92-0.94 Right side-bending=0.89-0.99 Inter-reliability: Left side bending=0.81-0.95 right side bending=0.89-0.93	
Extension	Numeric – ratio	Range of movement		Inter-reliability=0.61	Using inclinometer instrument
Trunk rotation	Numeric – ratio	Angle of movement from neutral		Intra-reliability: Left side bending=0.96 Right side-bending=0.92 Inter-reliability: Left side bending=0.85 Right side bending=0.82	
Active rotation standing	Numeric – ratio	Angle of movement from neutral		Intra-reliability: Left side bending=0.80 Right side-bending=0.86 Inter-reliability: Left side bending=0.85 right side bending=0.82	Using plumb line instrument
Modified Schobar test	Numeric – ratio or categorical - ordinal	Range of movement or presence of restriction	If numeric = distance moved in cm if categorical = positive or negative for restriction	Intra-reliability=0.87 Inter-reliability=0.77-0.79	Using specific landmarks on the spine, the patient is asked to bend forward. If the change in distance between the landmarks is less than 5cm then this is a positive test for lumbar restriction.
Pain lateral bending	Categorical - nominal	Presence of pain	Yes/no	Inter-reliability=0.60	
Lumbar motor control	Categorical –	Presence of poor motor	Nominal – Yes/No	Inter-reliability=0.90-0.98	Not clear whether measure amount

Test/Measurement	Data type	Type of measurement	Possible values	Accuracy/reliability accepted if +LR >5 and intra/inter-reliability >0.6	Notes
	nominal or ordinal	control	Ordinal – degree of control		moved from neutral or if just report that poor control exists or not. This could be ordinal
Lumbar segmental instability	Categorical - nominal	Symptoms reproduced during test	Yes/No	Inter-reliability=0.61-0.87	
Posterior-anterior force – least mobile segment	Categorical - nominal	Presence of restriction	L1/2, L2/3, L3/4, L45, L5/S1	Inter-reliability=0.71	
Tenderness Piriformis muscle	Categorical - nominal	Presence of pain	Yes/No	Inter-reliability=0.66	
Tenderness Tensor Fascia Latae muscle	Categorical - nominal	Presence of pain	Yes/No	Inter-reliability=0.75	
Fibromyalgia tender points	Numeric - ratio	Number of specific tender points	0-18	Inter-reliability=0.87	Diagnosing fibromyalgia doesn't necessarily use these tender points any more (Imrie, 2016)
Tenderness Multifidus muscle at L4/5 level	Categorical - nominal	Presence of pain	Yes/No	Inter-reliability=0.75	
Tenderness Multifidus muscle at L5/S1 level	Categorical - nominal	Presence of pain	Yes/No	Inter-reliability=0.81	
Centralisation	Categorical - nominal	Has occurred	Yes/No	Inter-reliability=0.70-0.90	
Straight Leg Raise -reproduce dermatomal pain	Numeric - ratio	Angle lower limb symptoms are reproduced	0-90º	Inter-reliability=0.68	Checking for neuropathic pain
Straight Leg Raise – resistance angle	Numeric - ratio	Angle resistance begins to be felt in tissues	0-90º	Inter-reliability=0.83-0.86	Checking for end of range resistance
Slump Test -angle feel pain	Numeric - ratio	Angle of knee discomfort felt as knee extended	90º to 0º	Intra-reliability=0.95	Use electrogoniometer. The patient is seated with their knee in a starting position of 90º
Slump knee bend test with neck extension	Categorical - nominal	Pain diminishes with neck extension	Yes/No	Intra-reliability=0.71	

Test/Measurement	Data type	Type of measurement	Possible values	Accuracy/reliability accepted if +LR >5 and intra/inter-reliability >0.6	Notes
McKenzie classification	Categorical - nominal	Identify syndrome type	Postural Derangement Dysfunction	Inter-reliability=0.61-0.70	
Diagnosis	Categorical - nominal	Identified by reference standard	Lumbar segmental instability Lumbar spinal stenosis Lumbar instability **Ankylosing spondylitis Lumbar radiculopathy **Disc bulge/herniation Normal or non-specific low back pain (depending on control chosen)	N/A	

** No tests met the threshold for diagnosing these conditions

K/ICC	Level of reliability/agreement
0.0-0.1	none
0.11-0.40	slight
0.41-0.60	fair
0.61-0.80	moderate
0.81-1.0	substantial

+LR	-LR	Amount of proof that the test will correctly identify/rule out the condition
>10.0	<0.1	Good
5.0 – 10.0	0.1 – 0.2	Moderate
2.0 – 5.0	0.2 – 0.5	Small
1.0 – 2.0	0.5 – 1.0	Poor

Appendix B – Information about physical tests and measurements used in the proposed dataset

Test/Measurement	Description
Pain sitting	Indicates patient has lumbar spinal stenosis because this movement involves a flexed spine and flexed spines relieve stenotic backs. Opens up the foramen where the nerves exit. Reference standard: CT/MRI
Patellar reflex	Indicates patient has lumbar radiculopathy (L3/4 nerve) Reference standard: Electro-diagnostics
Medial hamstring reflex	Indicates patient has lumbar radiculopathy (L5 nerve) Reference standard: Electro-diagnostics
Slump knee bend test	Indicates nerve root compression Reference standard: MRI
Two stage treadmill test	Indicates neurogenic claudication Reference standard: CT/MRI Using two stage treadmill test – longer walking total time met threshold whereas time to onset of symptoms and prolonged recovery after level walking (variations of the test) didn't. Basically, if a patient can walk for longer uphill (flexed spine) then indicates claudication when walking on flat surface.
Forward flexion	This measurement is not used in isolation for any specific condition. Cleland reliability was for distance of fingertips from floor and using an inclinometer.
Lateral flexion	This measurement is not used in isolation for any specific condition. Distance fingertips slid down
Extension	This measurement is not used in isolation for any specific condition. Using an inclinometer
Trunk rotation	This measurement is not used in isolation for any specific

	condition. Patients sat with horizontal bar on sternum Plumb weight hung down to floor angle measured with a protractor
Active rotation standing	This measurement is not used in isolation for any specific condition. Horizontal bar resting on shoulders plumb weight hung from end of bar to the floor
Modified Schobar test	
Pain lateral bending	
Lumbar motor control	Several tests exist and for all there was good reliability. Tests were repositioning, sitting forward lean, sitting knee extension, bent knee fall out, leg lowering (Cleland, 2016) p180 for a description.
Lumbar segmental instability	Various tests met the threshold. They are: hip extension test, painful arc in flexion, painful arc on return from flexion, aberrant movement pattern, prone instability test, trendelenburg, active straight leg raise
Posterior-anterior force – least mobile segment	Patient prone, examiner applies a force in a downward direction to each lumbar vertebra.
Tenderness Piriformis muscle	
Tenderness Tensor Fascia Latae muscle	
Fibromyalgia tender points	
Tenderness Multifidus muscle at L4/5 level	
Tenderness Multifidus muscle at L5/S1 level	
Centralisation	With experienced (5 years) and novice therapists as patient performed repeated McKenzie movements.
Straight Leg Raise -reproduce dermatomal pain	
Straight Leg Raise – resistance angle	Reliability of angle measured.
Slump Test -angle feel pain	Reliability of therapists measuring angle at which patient feels pain.
Slump knee bend test neck extension	Positive if symptom diminishes with neck extension
McKenzie classification	Two examiners with more than 5 years experience training in this method evaluated the patients and needed to classify them as either having one of the following syndromes:postural, derangement or dysfunction. They also had to say if they have a lateral shift. Clinicians needed to agree on the classification type. For classification and not lateral shift identification 0.61-0.70 several studies

Appendix C – The Machine Learning Model Scripts

The following scripts were used to explore and analyse the kaggle-12 and UCI datasets.

Programming Environments

Two programming environments were used:

- PyCharm 2018.2.2 Community Edition (IDE)
- Jupyter Notebook, Server version 5.5.0. Current Kernel Information: Python 3.6.5 (Anaconda, Inc. | (default, Apr 29 2018 16:14:56) [GCC 7.2.0] Ipython 6.4.0

Computer used for builds

DELL Inspiron 13 7000, Intel core i5, 8th Gen, 8Gb RAM

Dataset versions (.csv)

original: the original Kaggle-12 with 310 samples

reduced: under sKaggleampled version of the original Kaggle-12. Contains 200 samples

augmented: over sampled version of the original Kaggle-12. Contains 420 samples

Additional notes

All scripts (.py and ipynb) have been written with Python interpreter version 3.5 in mind.

The .ipynb files require Jupyter notebook to be installed.

The scripts have been written so as to expect the dataset files to be in the same folder as the scripts.

Script name	Description
learning_curve.py	Generates learning curve graphs which plots different sample sizes (from original, reduced or augmented version of the Kaggle-12 dataset) against the accuracy scores for those sample sizes. Uses the learning curve method from the Python sklearn.model_selection module. *Instructions*: To run for the different versions of the datasets, need to uncomment code in main()
log_reg_model.py	Builds a model using the logistic regression algorithm. *Instructions*: The user can select which dataset version and which columns to be used in the model by specifying the flags e.g. '--dataset_type original,augmented,reduced –columns 0,1,2,3,4,5'. If do not use the flags the model will be built with default 12 features and the original dataset.

Script name	Description
	Column one of the dataset is denoted by'0'. Do not have spaces between the values used with the flags.
svm_linear_model.py	Builds a model using a Support Vector Machine with linear kernel. *Instructions:* See log_reg_model.py instructions with regards to how to run from the command line.
svm_rbf_model.py	Builds a model using a Support Vector Machine with radial basis function kernel. *Instructions:* See log_reg_model.py instructions with regards to how to run from the command line.
bw_svm_rbf_model.py	'bw' stands for 'black and white'. They have been included just to show that the author considered inserting black and white images instead of colour when realised that the dissertation is most likely to be printed in black and white. However, it is best to view the document electronically as well as on paper to see the graphs properly.
bw_log_reg_model.py	See above
kaggle_12.boxplots.ipynb	Draws box plots for all 12 features of the Kaggle dataset
kaggle_12_feature_ranges.iypnb	Outputs the class='normal' and class='abnormal' ranges for each of the 12 features in the Kaggle dataset.
kaggle_12_histogram.ipynb	Plots individual histograms for all the 12 features in the Kaggle dataset. The histogram have plots for back pain (class='abnormal') and 'no back pain' (class='normal') distributions. *Note:* The author recognises that the bin sizes are not the best size for showing off the distribution. Smaller bins have been drawn for pelvic incidence.
kaggle_individual_scatter.ipynb	Scatter graph plots for pairs of features that looked like that had a high correlation: degree_spondylolithesis/pelvic_incidence degree_ spondylolithesis/pelvic_tilt degree_ spondylolithesis/lumbar_lordosis_angle degree_ spondylolithesis/pelvic_slope degree_ spondylolithesis/pelvic_radius pelvic_incidence/lumbar_lordosis_angle pelvic_incidence/pelvic_tilt pelvic_incidence/sacral_slope
kaggle_12_kernel_density.ipynb	Outputs kernel density plots which gives a better idea (compared to the histograms) of the type of distribution the data has.
kaggle_12_pairwise_correlations.ipynb	Outputs Pearson's correlation values.
kaggle_12_pairwise_scatter.ipynb	Plots every possible combination of pair feature plots. The graphs can take a couple of minutes to generate.
kaggle_12_qqplot.ipynb	Plots the quantile-quantile plots for each of the 12 features in the Kaggle dataset.
bw_kaggle_12_histogram.ipynb	Same as kaggle_12_histogram.ipynb but draws graphs with patterns

Glossary

A	
Ankylosing spondylitis	A type of arthritis mainly affectng the lower spine.
C	
centralisation	The receding pattern of nerve pain from distal to proximal. Nerve pain can be present down legs and arms and as it gets better the end location of the pain moves further towards the centre of the body.
C parameter	Soft margin
Cervical tilt	An angle formed between the vertical line from the centre of the top surface of the 1st thoracic to the tip of the 2nd cervical vertebra.
Cohen's kappa coefficient (K)	Score used to evaluate the intra and inter reliability of clinician's measurements.
Cross validation (in machine learning)	A process used to test the generalisation ability of a model and its hyperparameters. Tests a model on various subsets of data.
CT	Computer tomography
D	
decision boundary	A line or surface that separates data into classes.
Degree of spondylolithesis	The percentage slip of one vertebra upon another. Usually occurs in the lumbar spine and more typically a top vertebra slips forward on the vertebra beneath it.
Direct tilt	Cannot find a definition for this even though it was part of the Kaggle dataset.
electrodiagnostics	Tests which check the integrity of the sensory and motor (muscle) nerves.
fibromyalgia	A condition where a person experiences widespread body pain for reasons unknown.
Gamma parameter	A parameter that is set when using a support vector machine radial basis kernel so as to separate non-linear data into their respective classes.
hyperplane	With regards to an SVM it is a plane that separates data into their classes in a higher dimensional space.
Intra-class correlation coefficient (ICC)	Score used to evaluate the intra and inter reliability of clinician's measurements
kappa coefficient (K)	See Cohen's kappa coefficient
likelihood ratio	With regards to evaluation techniques, likelihood ratio scores the accuracy of a test's ability to prove the presence of a condition.
Logistic regression	A binary classification technique that uses the logit function to find the probability of a sample belonging to a positive or negative class.

Lumbar lordosis angle	The angle of the concavity in the lumbar region.
Lumbar spinal stenosis	Narrowing of the foramina where the spinal cord or nerve roots exit/enter.
Lumbosacral radiculopathy	The 'radicular' term refers to nerve fibres that represent a level of the spine. Lumbosacral readiculopathy is when these nerve fibres are dysfunctional due to e.g. compression.
maximum likelihood estimation	A method used to find the parameters of a model that are most likely to produce the data observed (Brookes-Bartlett, n.d.)
McKenzie movements	Techniques applied to the spine where the "main difference to most other assessments is the use of repeated movements rather than a single movement" (The McKenzie Institute International, n.d.)
Machine learning	"Machine learning is the science of getting computers to act without being explicitly programmed." (Coursera, 2018)
Machine learning model	Usually an equation that has been learnt from applying an algorithm to data. The equation (model) is then used to generalise and, for example, predict what some unseen data represents.
Model (machine learning model)	See *machine learning model*
MRI	Magnetic Resonance Imaging
New York Criteria	Used for confirming the presence of Ankylosing Spondylitis. Uses a combination of clinical and radiographic findings (Brent, 2018)
Pelvic incidence	An angle formed between a perpendicular line drawn from the centre of the sacral end plate and a line drawn to the centre of the femoral head.
Pelvic radius	The distance of a line drawn between the centre of the femoral heads.
Pelvic slope	
Pelvic tilt	An angle formed between a line drawn from the centre of the sacral end plate and centre of the femoral head and a line vertical to the central femoral head.
Reject option	
Sacral slope	
Sacrum angle	The angle between the tangent of the sacral end plate and the horizontal line.
Scoliosis slope	"Inclination of the end, most flexed scoliosis vertebrae with respect to the horizontal line". (Negrini, 2012)
sensitivity	
Schober test	A test used in physical therapy to measure the active range of motion of the lumbar spine
specificity	

soft margin	
spondylolithesis	
Straight leg raise	
Support vector machine	
Support vectors	
SVM	
Thoracic slope	