Machine Learning Classification Models and Hypothesis Testing for Back Pain Data

Group 98: Laura Murphy: 21267700 ; Royston Pinto: 20210611

*Abstract*—This report compares the efficiency of various classification models, which are logistic regression, classification trees and bagging classification trees, as well as performing hypothesis testing. The binary classification problem presented with the data set is trying to correctly infer whether a patient has abnormal or normal back pain based on a selection of numerical clinical indicators. The hypothesis testing investigates the mean difference in one of the variables between the two types of back pain. This was carried out using a 2 sample t-test and the result showed a statistically significant difference in the mean of the chosen variable.

The comparison of the models is done using performance measures, which quantify the quality of the models. The accuracy of each model is used as a proxy for the optimal classifier, which is found to be the bagging classification tree. The relative importance of the predictor variables in the bagging model is established, which foremost includes the degree of spondylolisthesis as an important variable. The hypothesis testing focuses on the plausibility of the numerical variables being a statistically significant variable in one of the groups. The benefits of implementing this model in the health sector are discussed, which could potentially lead to vast improvements in the quality of health services in Ireland and worldwide.

Keywords—Backpain; Classification Model; Logistic Regression; Classification Trees; Bagging; Hypothesis Testing

# Introduction

Lower back pain can be initially classified into either abnormal or normal pain, with the problems being caused by parts of the complex, interconnected network of spinal muscles, nerves, bones, discs or tendons in the lumbar spine.

Although lower back pain is an extremely common ailment, its severity can vary considerably. Normal back pain is the more common type and can be caused by mechanical or physical harm done to the body. Abnormal pain can be linked to the damage done to the body’s central nervous system, which includes the spinal cord. Normal pain can often be helped by common painkillers, such as paracetamol and morphine, depending on the severity of pain. However, abnormal pain is usually only helped by specific medication and is left unaffected by painkillers.The differentiation between the type of pain is a primary key factor in determining the type of medication that would help ease a patients’ pain.

# Related Work

In today's time, lower back pain is very common and an ever increasing problem. Lower back pain mostly caused due deficiency in our diet of essential nutrients and lack of exercise. We have a lot of papers in our field having various datasets and papers published. It's been a well-known topic in various journals using various methods to get to certain conclusions. There has been a vast amount of related work done in this field in IEEE journals and so on [1]. Prediction of low back pain using artificial intelligence modelling by Nitish Aggarwal is a related paper as well as Multi-modal biomarkers of low back pain: A machine learning approach which is the 29th volume of the  NeuroImage:Clinical [2].

# Dataset & Exploratory Analysis

The data set is based on 310 individuals experiencing lower back pain. The data set contains a collection of 12 numerical variables for each patient as well as classifying their lower back pain as normal or abnormal. The variables are physical spine details including pelvic incidence, pelvic tilt, scoliosis slope and cervical tilt.

By exploring the data we found that, the pelvic incidence ranged from values of 26.15 to 129.83, having a mean of 60.5 and standard deviation of 17.2.  The pelvic tilt variable ranged from values of -6.55 to 49.4, with the maximum count of 79 ranging between 15.84 to 21.44. The cervical tilt had a mean of 11.9 and standard deviation of 2.89.  The degree of spondylolisthesis, which is an important variable that is explored further in this paper, ranged from values of -11.1- 419, where the maximum count of 205 were in the section of -11.06 to 31.90. The value of 419 is seen as an outlier in this set.

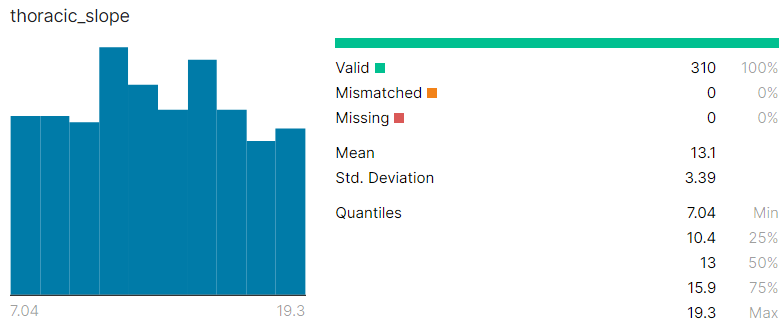
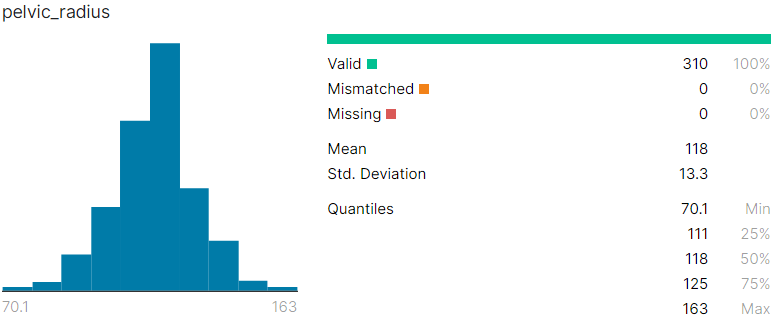
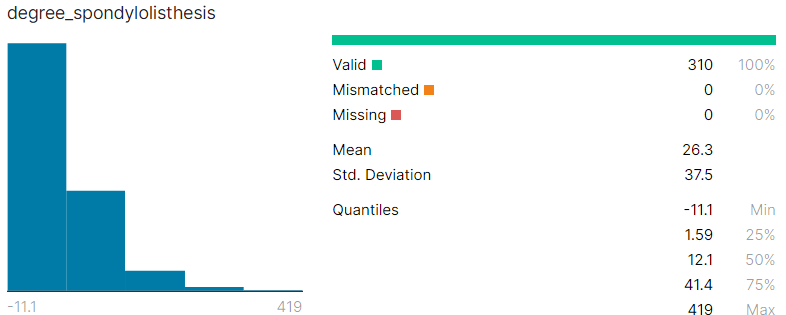


Fig. 1: Histograms of Degree\_spondylolisthesis, pelvic\_radius and throacic\_slope

# Hypothesis and Research question

One interesting variable in the dataset was chosen to investigate how statistically significant it is in differentiating between the 2 classes of back pain.

Spondylolisthesis is a condition that occurs when one vertebral body slips with respect to the adjacent vertebral body causing radicular or mechanical symptoms or pain, as explained in [3]. Degree\_spondylolisthesis is the numerical variable measured for each of the patients in the dataset and it describes the severity of this condition.

H0 - The mean degree of spondylolisthesis in patients with Normal back pain is not statistically different to the mean degree of spondylolisthesis in patients with Abnormal back pain.

HA - The mean degree of spondylolisthesis in patients with Normal back pain is statistically different to the mean degree of spondylolisthesis in patients with Abnormal back pain.

A 2 sample t-test was used to investigate the above hypothesis, in particular, searching for any differences between the means of the degree of spondylolisthesis.

A further research question was investigated which was to try and find an optimal classification method to predict lower back pain. If a sufficiently optimal classifier can be found, it would be able to correctly predict the lower back pain of a patient, given the relevant indicators.  If a patients’ pain could be classified before being seen by a medical professional, the time and ease of the service could be improved.

# Methods Used and Why

## Hypothesis Testing

*Function:* t.test()

An initial data exploration was performed on the variable before the hypothesis testing. From Figure 2, it appears that the Normal class has a lower degree of spondylolisthesis than the Abnormal class.

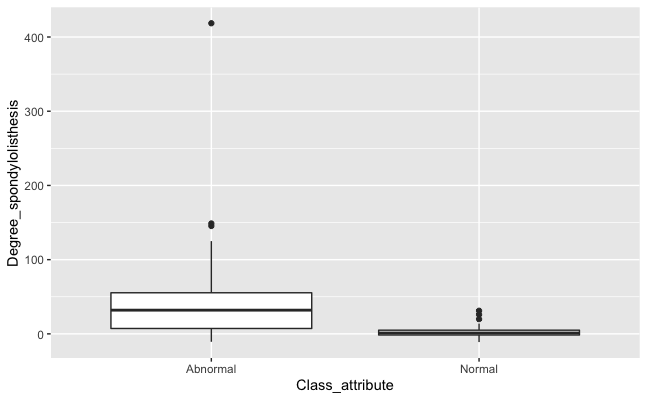


Fig. 2: Boxplots of Degree\_spondylolisthesis in Normal class and Abnormal class

The t-test was performed to see if the relationship observed in Figure 3 is statistically significant.

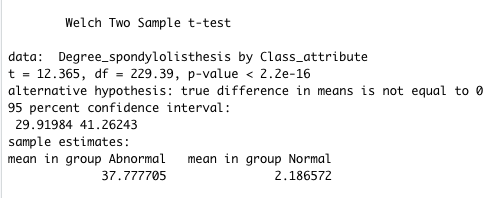


Fig. 3: R Output of 2 sample t-test

The results of the hypothesis test are discussed in the Results and Findings section.

The following classification algorithms are all attempting to correctly classify the lower back pain type (Normal or Abnormal) using the other numerical variables.

## Logistic Regression

*R Package: nnet, Function: multinom()*

Logistic regression is one of the most interpretable and oldest classification methods used for binary classification. This method attempts to estimate the relationship between the dependent variable, i.e. Class Attribute, and the explanatory variables. This estimation is done using Maximum Likelihood Estimation (MLE). MLE is based on maximising the likelihood function and in turn, determining which parameters are most likely to produce the observed data, according to [4].

For logistic regression to be used, the explanatory variables must be treated as categorical or continuous. The multinom() command is used for this data set, which is an extension of the binary logistic regression command, glm(). It treats the classification as a multi-class case with two classes instead of a strict binary classification.

## Classification Trees

*R Package: ROCR, Function: rpart()*

A classification tree splits the data into partitions and further splits those partitions into branches until no more appropriate splits can be made. These splits are made based on which variable splits give the greatest reduction in the Gini Index. The Gini Index is a statistical measure between 0 and 1 which rewards when all the observations belong to one category, i.e. closer to 0, and penalises when all the observations belong to different categories, i.e. closer to 1. Therefore, the variable which splits the data and the optimal splits of the variable which simultaneously manage to bring the Gini Index closer to 0, is chosen as the first split and so on. The tree is grown greedily until an optimal split is found at each node, according to [5].

## Bagging Classification Trees

*R Package: adabag, Function: bagging()*

Bagging, also known as bootstrap aggregating, is a method that can reduce the variability and increase the performance of classifiers, such as classification trees. Implementing this method with classification trees involves taking a sample with replacement, and training a classification tree on this sample. This is repeated with the chosen number of samples, and the average of the predictions of all the sample models was calculated. The default number of trees used in this function is 100. A bagging tree is created that is deemed more accurate than the original classification tree as it was designed over an average of models. It should be noted that asymptotically, 63.2% of unique values can appear in a bootstrap sample, stated in [6].

To create and test the classifiers, it is appropriate to split the data into a training set and a testing set. The training set is made up of a random sample of 70%, i.e. 217 observations, of the data set. This set is used in the training of the classifiers and is also used to compare the performance of the models. The remaining 30%, which contains 93 observations, is called the testing set, which is the sample of data used to provide an unbiased evaluation of the fit of the best model.

The predict() command is then used to attain each models’ prediction on the type of back pain that the testing observations have. A cross-tabulation of the true classification of back pain against the predicted classification is created for each model and at each iteration.

A generic example of the resulting table can be seen in Table 1.

|  | True Observations | |
| --- | --- | --- |
| Normal | Abnormal |
| Predicted Observations |  |  |
| Normal | True Positive(TP) | False Negative(FN) |
| Abnormal | False Positive(FP) | True Negative(TN) |

Table 1: Cross-Tabulation of True and Predicted Observations

From this table, performance measures of the models are calculated, based on the model’s predictions of the validation set. The performance measures most appropriate to differentiate the best classifier are as follows:

*Accuracy:* Defines how many are correctly classified?

(TP+TN)/(TP+FP+TN+FN)

*Sensitivity:*Defines of those that are truly positive, how many are predicted as positive?

(TP)/(TP+FP)

*Specificity:* Defines of those that are truly negative, how many are predicted as negative?

(TN)/(TN+FP)

The summary statistics are computed for the above measures on every classification method, and the average was taken as a suitable estimate for the performance measures.

The accuracy is the leading performance measure in distinguishing the efficiency of the models. Therefore, the model that is chosen as the optimal classifier has the highest accuracy as well as high sensitivity and specificity.

After the optimal classification model is selected, the model is then trained again on the full training set, which contains 70% of the observations. This trained model is then run on the test set, and an evaluation of the output is carried out. The evaluation involves discussing the output of the final model and the performance measures of the model in predicting the test set.

The importanceplot() command is also used to view which variables are of most importance in determining the fit of the optimal model, detailed further by Kuhn and Johnson [7]. The relative importance takes into account the gain of the Gini index given by a variable in a tree. A large value indicates that the variable contributes more to the predictive power of the model.

# Results and Findings

## Results of Hypothesis Testing

In Figure 2, the results of the hypothesis test are shown.  The results show a p-value < .01 supporting the alternative hypothesis that “true difference in means is not equal to 0”; essentially it states there is a statistical difference between the means in the Normal back pain class and the Abnormal back pain class.

## Results of Classification Models

In choosing the appropriate logistic regression model, it is important to note that the phenomenon of separability occurs in this dataset. This means that one or some predictor outcomes are associated with only one class outcome, i.e. high pelvic tilt could result in the observation being classified as Abnormal, regardless of the other predictor variables.

The traditional classification tree is the second best model, with an accuracy of 0.8061. Furthermore, this tree is the least capable of correctly classifying abnormal pain, only correctly predicting 66.05% of the abnormal patients.

The bagging classification tree is the best classification model, with the highest performance measures in all 15 iterations.

Table 2 summarises the measures which assess the performance of the classification methods. These figures below are an average of the measures calculated from each of the 15 iterations. They distinguish the efficacy of the models in correctly predicting the type of back pain.

| Classification Method | Accuracy | Sensitivity | Specificity |
| --- | --- | --- | --- |
| *Logistic Regression* | 0.8409 | 0.8776 | 0.7740 |
| Classification Tree | 0.8061 | 0.8792 | 0.6605 |
| Bagging Classification Tree | 0.9425 | 0.9566 | 0.9155 |

#### Table 2: Performance Measures for Each Method based on the Average of the 15 Iterations

An overall summary of the importance of each predictor in the trained bagging classification tree is displayed using a variable relative importance plot seen in Figure 4. The predictor with the highest relative importance is Degree\_spondylolisthesis.

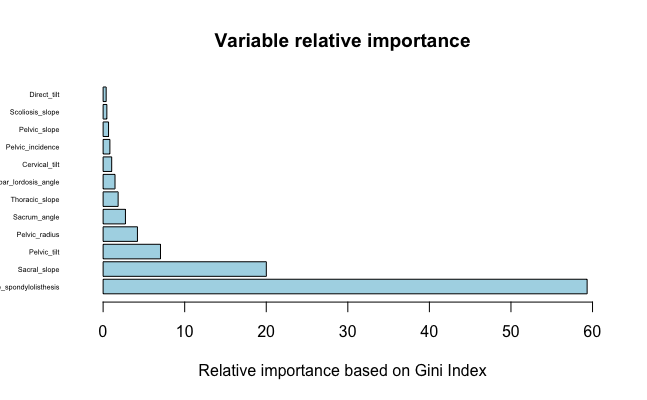


Figure 4: Plot of Relative Importance of Each Variable in Bagging Algorithm

# Conclusion

At the 1% significance level, there is sufficient evidence to support that there is a significant difference in the mean degree of spondylolisthesis in Normal and Abnormal back pain.

Furthermore, in pursuance of finding an algorithm to successfully classify back pain given the available data, a comparison of reputable classification methods is carried out. By successfully executing this comparison, the optimal classifier is determined to be the bagging classification tree. The variable with the highest predictive quality is found to be the degree of spondylolisthesis, similar to what is discovered in the hypothesis testing. Therefore, given the relevant information, the bagging model can be used to predict the back pain of a patient, with high accuracy, with the degree of spondylolisthesis being a statistically significant variable in determining the type of back pain.

This result sees the potential use of this model in optimising the service and care of patients. The implementation of this model in medical facilities, particularly Accident and Emergency departments, could greatly increase the quality of medical care given. This could be achieved by measuring the spinal details such as pelvic tilt and cervical tilt on patients claiming back pain, upon entry to an A&E department. This initial examination would pay particular attention to the significant variables used in classifying the pain. The results of the tests could be run through the model, and further advice and

treatment could be given based on the highly accurate prediction, such as the prescription of painkillers if abnormal pain is predicted. A reduction in wait times, a reduction in short-term pain of patients, and specialist appointments are all likely developments from this model being used.

# References

[1]: Nitish Aggarwal. Prediction of low back pain using artificial intelligence modelling [Internet]. Available from:

<https://jmai.amegroups.com/article/view/6003/html>

[2]: Lamichhane B, Jayasekera D. Multi-modal biomarkers of low back pain: A machine learning approach, 29th volume NeuroImage:Clinical [Internet]. Available from:

<https://www.sciencedirect.com/science/article/pii/S2213158220303673>

[3]: Steven Tenny, Christopher C. Gillis. Spondylolisthesis; 2021[updated 2021 July 25]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK430767/>

[4]: Avinash Navlani. Understanding Logistic Regression in Python[Internet]. America: Data Camp; 2018[updated 2018 September 7]. Available from:

<https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

[5]: Frontline Solvers. Classification Tree[Internet]. America: Frontline Solvers. Available from:

<https://www.solver.com/classification-tree>

[6]: Enhance Data Science. Machine Learning Explained:Bagging[Internet]. America: Enhance Data Science; 2017[updated 2017 June 18]. Available from:

<http://enhancedatascience.com/2017/06/28/machine-learning-explained-bagging/>

[7]: Kuhn M, Johnson K. Applied Predictive Modelling. New York: Springer; 2013