

School of Computing

FACULTY OF ENGINEERING



UNIVERSITY OF LEEDS

**Application of Machine Learning to Routine Interpretation of Blood
Test Results**

Sebastian Thomas

**Submitted in accordance with the requirements for the degree of
MSc Advanced Computer Science (Artificial Intelligence)**

2021/2022

The candidate confirms that the following have been submitted:

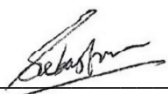
Items	Format	Recipient(s) and Date
Deliverable 1	Report	SSO (24/08/22)
Deliverable 2	GitHub link	Supervisor, assessor (24/08/22)

Type of Project: Empirical Investigation

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student)

_____

Summary

Routine interpretation of blood test results is useful in diagnosing conditions like hyper/hypocalcaemia. Blood test results can be used to monitor the levels of calcium and other molecules that contribute to the regulation of calcium concentrations in the blood. The application of machine learning in this clinical field could help facilitate making diagnoses for consultants.

After evaluating a variety of machine learning models, regression trees proved to be the most accurate with an accuracy score of 0.93. The use of a pruned regression tree model improved this accuracy score to 0.94 which is the best model achieved in this project. To determine the effectiveness and validity of this model, it was compared with a consultant given decision tree. There were many similarities between both trees and proved that the results achieved in this are valid which means this project was a success.

Acknowledgements

I would like to thank Prof. Anthony Cohn for providing consistent guidance and support throughout the project duration. He was always available to help me if I needed guidance or something reviewed.

I would also like to thank my assessor Dr Yanlong Huang for his valuable feedback and suggestions.

Table of Contents

Summary	iii
Acknowledgements	iv
Table of Contents.....	v
Chapter 1 Introduction.....	1
1.1 Project Aim.....	1
1.2 Objectives	1
1.3 Deliverables.....	2
1.4 Ethical, legal, and social issues	2
Chapter 2 Background Research.....	3
2.1 Literature Survey	3
2.1.1 Support Vector Machine Based Classification Method for Hyper/Hypocalcemia Diagnosis	3
2.1.2 Pro-FHH: A Risk Equation to Facilitate the Diagnosis of Parathyroid- Related Hypercalcemia	3
2.1.3 Decision Trees and Pruning.....	4
2.1.4 Improving Diagnostic Recognition of Primary Hyperparathyroidism with Machine Learning	4
2.2 Methods and Techniques	4
2.2.1 Algorithms.....	4
2.2.2 Technology	5
2.3 Choice of Methods.....	5
2.3.1 Chosen algorithms	5
2.3.2 Chosen technology	5
Chapter 3 Datasets and Experimental Design	7
3.1 Dataset.....	7
3.2 Experimental Setup	9
3.2.1 Processing Data	9
3.2.2 Experiment 1.....	10
3.2.3 Experiment 2.....	10
3.2.4 Experiment 3.....	10
3.2.5 Experiment 4.....	11
Chapter 4 Results of the Empirical Investigation	12
4.1 Evaluation of Models	12

4.2 Regression Tree Predictions	13
4.3 Ensemble Model Predictions	15
4.4 Pruned Regression Tree Predictions	18
Chapter 5 Validation of Results	22
5.1 Comparison with Consultant's Decision Tree.....	22
5.2 Metrics.....	23
Chapter 6 Conclusion	24
List of References	25
Appendix A External Materials.....	27
Appendix B Ethical Issues Addressed	28
Appendix C Decision Trees.....	29

Chapter 1

Introduction

Blood tests have a variety of uses and it is one of the most common types of medical test (NHS, 2019). 'Blood tests are a standard part of routine and preventive healthcare. A doctor will often order a blood test before or following a physical examination. A doctor may also order blood tests to evaluate specific conditions.' Calcium is an important mineral for several biological processes including neurotransmission, muscle contraction, hormone secretion and the clotting cascade (Moody, 2021). Regulation of blood calcium concentrations is important for these biological processes. Blood test results can be used to monitor the levels of calcium and other molecules that contribute to the regulation of calcium concentrations in the blood. Doctors can diagnose patients using these blood test results.

1.1 Project Aim

The aim of this project is to develop and compare machine learning algorithms from a large patient data set provided by LTHT. Comparison of the best model obtained will be compared with consultant grade interpretations to determine the effectiveness. If successful, this project would demonstrate how the incorporation of machine learning and AI into the clinical field can harmonise reporting while reducing human input; a finding yet to be fully documented.

1.2 Objectives

- Read and analyse background literature related to interpretation of blood test results and about decision support systems for clinical diagnosis, and hypo/hypercalcaemia in particular
- Split dataset into training, validation and test sets
- Pre-process the data for training
- Train a few machine learning models to interpret blood test results
- Evaluate results
- Implement a machine learning ensemble model
- Visualise the decision tree(s)
- Finish code and write final project report

1.3 Deliverables

- MSc Project Report
- Jupyter notebook with complete code and showing the output of all cells

1.4 Ethical, legal, and social issues

This project does not use any personal data of the patients and the data cannot be used to trace back to the patients.

Chapter 2

Background Research

This chapter provides the literature survey and choice of methods and techniques. Literature survey includes all the reading done for this project to understand the background and evaluation of similar projects which have been done. Methods and techniques section will focus on the tools needed and solution ideas.

2.1 Literature Survey

2.1.1 Support Vector Machine Based Classification Method for Hyper/Hypocalcemia Diagnosis

The full title of the paper studied is 'A Support Vector Machine Based Classification Method for Hyper/Hypocalcemia Diagnosis' (Wang, Jin and Wang, 2021). The project described in this paper is similar because this project looks at hyper/hypocalcaemia too. In this this paper a support vector machine (SVM) algorithm is used for classification. This SVM's parameters are then optimised using particle swarm optimisation algorithm (Wang, Jin and Wang, 2021). This optimised SVM model achieved an accuracy of 99.96% which was higher compared to traditional SVM based diagnosis method (Wang, Jin and Wang, 2021). This is important because the high accuracy in diagnosis of hyper/hypocalcaemia means it could be perfect for the project.

2.1.2 Pro-FHH: A Risk Equation to Facilitate the Diagnosis of Parathyroid-Related Hypercalcemia

This paper focuses on familial hypocalciuric hypercalcemia (FHH) and parathyroid-related hypercalcemia. Primary hyperparathyroidism (PHPT) is an endocrine disease and most patients who have this disease also have mild hypercalcaemia (Bertocchio et al., 2018). These patients will have slightly high parathyroid hormone (PTH) concentration or even normal PTH concentration (Bertocchio et al., 2018). The treatment for this disease is parathyroidectomy. FHH patients have similar calcium, magnesium and PTH concentrations so I can be hard to distinguish them from PHPT patients especially at normal levels of PTH (Bertocchio et al., 2018). Using calcium to creatine clearance ratio (CCCR) can help distinguish between the two groups (Bertocchio et al., 2018). So, this aim of this paper is to protect FHH patients from undergoing parathyroidectomy. Two different logistic regression models were used in this project. One model is supervised and the other is unsupervised. They achieved area under the receiving operator characteristic curve (AUROC) score of

0.862 and 0.961 respectively (Bertocchio et al., 2018). These are good AUROC scores so these models can be considered for this project too.

2.1.3 Decision Trees and Pruning

Chapter 6 of 'Data Mining: Practical Machine Learning Tools and Techniques' book was studied to understand pruning of decision trees. Pruning prevents overfitting the data and there are two ways to do this. Post pruning takes a fully grown decision tree and discards unreliable parts, and pre pruning stops growing a branch when information becomes unreliable (Witten and Frank, n.d.). This chapter of the book also describes cost-complexity pruning (CCP) which can provide a more compact tree. This may be required because some decision tree algorithms like C4.5 does not prune enough when post pruning (Witten and Frank, n.d.).

2.1.4 Improving Diagnostic Recognition of Primary Hyperparathyroidism with Machine Learning

In this paper the project is similar but only focuses on PHPT. In this project 10-fold cross validation is used to compare machine learning models. After testing on multiple models, Bayesian network models gave the highest accuracy, correctly classifying 95.2% of all PHPT patients (Somnay et al., 2017). Moreover, AdaBoost was used to improve the accuracy of the Bayesian network model to 97.2% (Somnay et al., 2017). These high accuracies are encouraging, and a similar approach can be used in the project.

2.2 Methods and Techniques

2.2.1 Algorithms

There are many machine learning algorithms that could be used for the project. Firstly, in section 2.1.1 the paper studied showed that the SVM model was successful in the project. 'The purpose of SVM is to establish an optimal decision hyperplane, which maximizes the distance between the two classes of samples nearest to the hyperplane on both sides of the plane, so as to provide good generalization ability for classification problems' (Wang, Jin and Wang, 2021).

Decision tree models could be used as a model for this project too. 'Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features' (scikit learn, 2009). It basically works like a flowchart and makes decisions based on previous experience (w3schools, n.d.).

Another machine learning algorithm that could be used for the project is logistic regression. This was used in one of the projects in background research and was successful. Logistic regression is used to solve classification problems, so it is suitable for this project.

In addition, K-nearest neighbours (KNN) is an algorithm that can be used to solve classification problems. 'It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points' (w3schools, n.d.).

Linear Discriminant Analysis (LDA) is a machine learning algorithm that can be used for multi-class classification. It can be used to reduce the dimensions. LDA is 'a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule' (scikit-learn, n.d.).

Finally, another algorithm that is interesting is Naïve Bayes (NB) algorithm. NB can be used for multi-class classification problems. As mentioned in 2.1.4, Bayesian network models perform really well at classifying PHPT so it could do the same in this project too.

2.2.2 Technology

Python is a programming language that is commonly used for machine learning. Python has many useful libraries like scikit-learn and pandas. Pandas can be used to load in data as a data frame and this can easily be manipulated. Matplotlib library is useful too because it allows visualisation of the data. There are many IDEs that could be used for Python. For example, Jupyter Notebook and Spyder. Apart from Python another option is Waikato Environment for Knowledge Analysis (WEKA). WEKA is a machine learning software in Java. 'Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization' (Waikato, 2019).

2.3 Choice of Methods

2.3.1 Chosen algorithms

For this project all of the algorithms mentioned in section 2.2.1 will be used. Just like the project in 2.1.4 all the algorithms will be compared using K Fold stratified cross validation and the best model would be used for further experiments. Also, this would satisfy part of the aim of this project which is to develop and compare machine learning algorithms.

2.3.2 Chosen technology

Previous knowledge of Python and its libraries mean that this is the best option for this project. In addition, Python is user friendly and there is a large community of developers, so it is expected that there are plenty of resources such as tutorials and forums to learn from.

Similarly, previous experience with Jupyter notebook means this is the best option. However, this will be done in Google Colab so that there is enough space for the data and computing power.

Chapter 3

Datasets and Experimental Design

This chapter looks at the dataset and the experiments that will be carried out.

3.1 Dataset

The patient dataset used in this project is from Leeds Teaching Hospitals NHS Trust (LTHT). LTHT clinical biochemists review all Calcium and PTH results and add interpretative comments based on five key blood test results, calcium, PTH, Vitamin D, magnesium and eGFR. The data is provided as an excel file. This is shown in Figure 1. In this figure it is clear that the dataset doesn't contain any personal information relating to the patient that can be used to identify them. All the data is continuous apart from the interpretation which is the target variable. The different codes for interpretation are a diagnosis outcome and these are given in Table 1.

	A	B	C	D	E	F	G
1	Patient	PTH	Ca	Vit D	Mg	eGFR	Interpretation
2	1	1.1	1.9	72	0.1	80	A1
3	2	15.4	2.85	90	0.9	66	I
4	3	0.7	1.7	89	1	90	A2
5	4	2.3	1.9	60	0.1	80	B1
6	5	12.2	1.7	32	1	90	C1
7	6	0.5	2.92	95	1.3	71	G
8	7	0.5	2.1	63	0.2	88	A1
9	8	6.4	1.7	89	1.3	90	B2
10	9	0.5	2.9	75	1.4	71	G
11	10	1.4	1.8	94	0.2	90	A1
12	11	12.3	2	43	1.4	71	C1
13	12	3	1.8	58	0.2	90	B1
14	13	2.9	2.77	56	0.8	71	H
15	14	0.7	1.9	60	0.2	90	A1
16	15	1.1	2.61	84	0.9	71	G
17	16	7.4	1.9	60	0.2	90	B1
18	17	1.3	2.91	84	0.9	71	I
19	18	1.9	2.1	63	0.2	90	B1
20	19	11.2	2.1	24	1.4	71	C1
21	20	0.8	2	63	0.3	77	A1
22	21	10.8	1.8	80	0.7	86	C2
23	22	1.4	2.79	84	0.9	71	G
24	23	0.8	2	79	0.3	77	A1
25	24	16.3	1.8	32	0.7	90	C1
26	25	12.3	2.15	79	0.9	71	C2
27	26	4.2	2	79	0.3	77	B1
28	27	1.4	2.2	80	0.9	71	D
29	28	0.7	2.5	84	1.1	90	D
30	29	7.1	2	63	0.3	77	B1
31	30	5.3	2.91	84	0.9	71	H
32	31	4.6	1.8	55	0.9	90	B2
33	32	5.3	2.1	65	0.3	77	B1
34	33	0.7	3.12	75	1.4	86	G
35	34	4	2.19	55	0.3	77	B1

Figure 1: Dataset

Table 1: Codes for Interpretation in Dataset (as provided by LTHT)

Code	Description
A1	'?Hypocalcaemia secondary to magnesium deficiency (blunted PTH response). See http://nww.lhp.leedsth.nhs.uk/common/guidelines/detail.aspx?ID=2105 for guidance on Mg replacement before repeat Ca.'
A2	'Hypocalcaemia with low PTH suggestive of primary hypoparathyroidism, suggest referral to Endocrinology.'
A3	'Hypocalcaemia with low PTH suggest repeat with Mg'
B1	'?Hypocalcaemia secondary to magnesium deficiency (blunted PTH response). See http://nww.lhp.leedsth.nhs.uk/common/guidelines/detail.aspx?ID=2105 for guidance on Mg replacement before repeat Ca.'
B2	'Hypocalcaemia with inappropriately normal PTH suggestive of primary hypoparathyroidism, suggest referral to Endocrinology.' however if borderline low calcium suggest repeat in the first instance.
B3	Hypocalcaemia with inappropriately normal PTH suggest repeat with Mg.
C1	'Secondary hyperparathyroidism due to vitamin D deficiency/insufficiency
C2	?historical vitamin D deficiency/?Mg.'
C3	?Vitamin D status
D	'Normal calcium with low PTH unlikely to be significant.'
E	Normal profile
F1	Secondary hyperparathyroidism due to vitamin D deficiency/insufficiency
F2	?normocalcaemic primary hyperparathyroidism, suggest repeat in the first instance before considering referral to Endocrinology.
F3	'Calcium within reference range and borderline raised PTH likely not significant. Unless primary hyperparathyroidism is suspected no follow

	up required.
G	'Non-PTH dependant hypercalcaemia suggest further investigations'.
H	'Hypercalcaemia with PTH in the reference range consistent with primary hyperparathyroidism, suggest referral to Endocrinology.'
I	'Hypercalcaemia with raised PTH consistent with primary hyperparathyroidism, suggest referral to Endocrinology'

3.2 Experimental Setup

3.2.1 Processing Data

Pandas library was used to load the excel file in as a DataFrame. The dataset contains many null values. For vitamin D there are 10 null values and for magnesium there are 20 null values. However, all these null values belong to a certain class therefore removing the data completely will remove some classes from the data. To keep these classes the null values should be replaced with -99. This is one way of handling missing values in the dataset.

In addition to this, outliers need to be removed from the dataset. There are a few outliers in the dataset which can be seen on the box plot in figure 2. The most obvious outlier is for magnesium. This was removed by getting rid of data that is greater than 0.99 quantile.

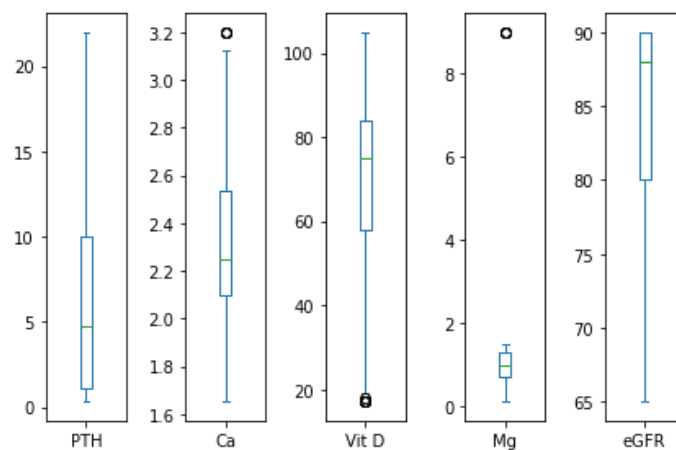


Figure 2: Box Plot

There are a number of duplicate records in the dataset too. These 46 records were dropped from the dataset, leaving just 350 records in the dataset. The distribution of this data across

the classes can be seen in figure 3. The dataset is not perfectly balanced, so the least populated class has just 8 records and most populated class has 26 records.

Interpretation	
A1	19
A2	20
A3	8
B1	21
B2	21
B3	12
C1	25
C2	20
C3	9
D	26
E	23
F1	24
F2	24
F3	24
G	26
H	25
I	23

Figure 3: Class Distribution

3.2.2 Experiment 1

The first experiment is to compare all the machine learning models mentioned in section 2.2.1 using stratified 7-Fold cross validation. 7-Fold is used because of the amount of data in the A3 class. Stratified cross validation is used because the dataset is not balanced. This method splits the data into 7 subsets and each subset in turn is used for testing and the remaining for training. The scores from this experiment should then be compared.

3.2.3 Experiment 2

The next step is to split the dataset into 8:2 ratio with the training set containing the majority and the test set with the minority. Train the best model discovered in experiment 1 on the training set and use the model to predict on the test set. The model should be evaluated using a confusion matrix. This can be done using the seaborn library. Metrics such as accuracy, precision, recall and F1 score should be used. This can be done using scikit-learn library to produce a classification report. The best model from experiment 1 was a regression tree. This tree should be displayed so that all the rules can be seen. The feature importances of this model should be plotted on a graph.

3.2.4 Experiment 3

This experiment will try an ensemble approach which is one of the objectives of this project. The use of ensemble networks is comparable to a group of consultants agreeing on a classification. In experiment 2 one regression tree produced good results so a random forest should give even better results. Random forest classifier should be trained on the training set and used to predict on the test set. The model should be evaluated in the same way as for Experiment 2 with a confusion matrix and classification report. At least one of the trees

should be displayed to show how complex the tree is. The feature importances of this model should be plotted on a graph too.

3.2.5 Experiment 4

Lastly, a pruned decision tree model should be investigated. With a greater depth, decision trees are prone to overfitting. A maximum depth could be set so a less complex tree is formed. This is called pre pruning and it may cause the model to stop prematurely. A better option is to use post pruning. As mentioned in 2.1.3, cost complexity pruning could give a more compact tree. The best tree using this method should be used to predict on the test set and the model should be evaluated using a confusion matrix and classification report. This tree should be displayed too so that the complexity can be compared with the unpruned tree. The feature importances of this model should be plotted on a graph too.

Chapter 4

Results of the Empirical Investigation

This chapter provides an evaluation of the experiments.

4.1 Evaluation of Models

In this investigation six models were compared to obtain the best model that should be used to predict on the test set. Stratified 7-Fold Cross Validation was used in this experiment to maximise the performance on each class.

Table 2: Model Accuracy

Algorithm	Accuracy Score	Variance
Logistic Regression (LR)	0.65	0.065
Linear Discriminant Analysis (LDA)	0.83	0.074
K-Nearest Neighbours (KNN)	0.29	0.072
Classification and Regression Trees (CART)	0.93	0.026
Gaussian Naïve Bayes (NB)	0.89	0.037
Support Vector Machines (SVM)	0.41	0.074

Table 2 shows that the CART algorithm achieved an accuracy of 0.93 which means it is the best to use for the classification problem. This high accuracy score was produced by a single regression tree so an ensemble approach may give a higher score. The Bayes model performed quite well too, producing an accuracy of 0.89. This model was expected to give a good result because it performed really well in other similar projects which are mentioned in chapter 2. KNN model gave the worst performance and only gave an accuracy score of 0.29. This could be because of the imbalanced dataset. The KNN algorithm is known to have problems with imbalanced datasets. A significance test could have been done to check if best model is actually the best, but the gap is big enough to mean it is.

4.2 Regression Tree Predictions

The best model to use for this classification problem is the regression tree model. This was trained and used to make predictions on the test set. This model achieved an accuracy score of 0.91, so it performed really well. The only disappointment with the results is that the model did not classify any of the test data correctly for the A3 class. This is because this class didn't have a lot of data like other classes. It only has 8 records, so it is not enough for the model to learn. A way to solve this problem and improve accuracy would be to collect more data for the underrepresented classes. Apart from this, in figure 4 the diagonal colouring on the confusion matrix shows the success of the model in classification.

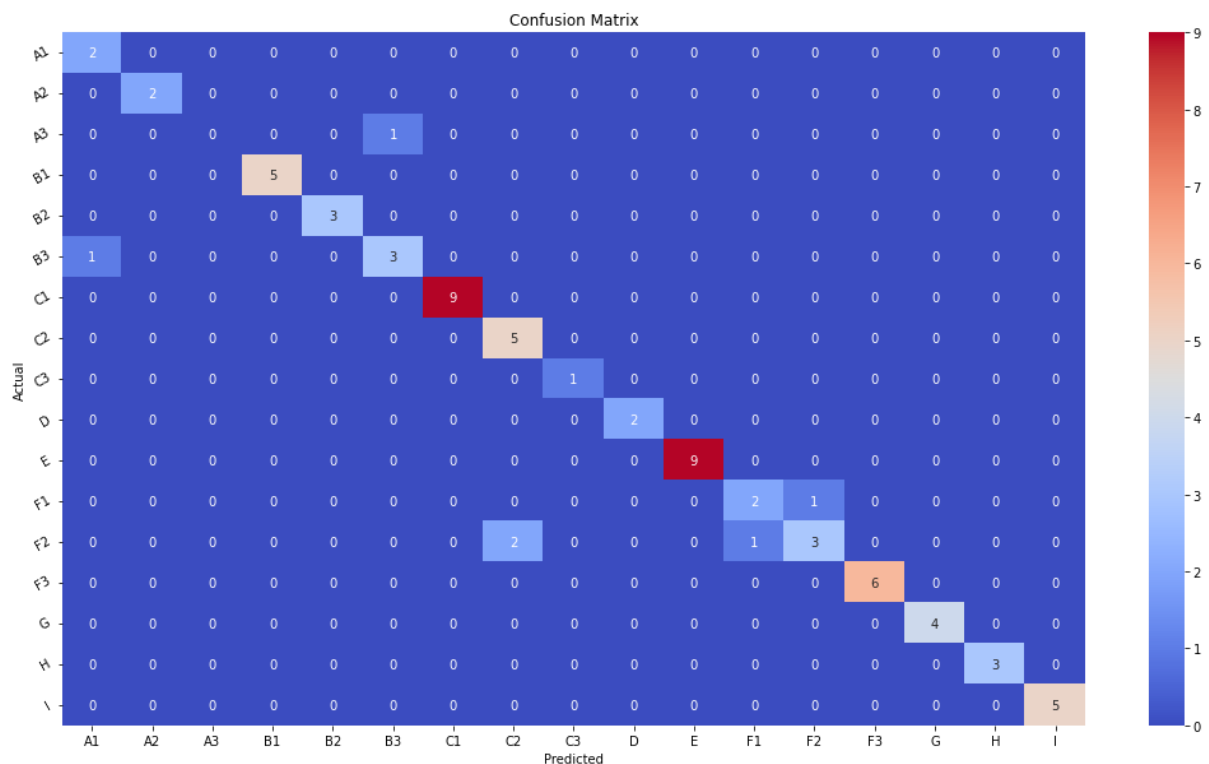


Figure 4: Confusion Matrix for Regression Tree Classification

In figure 5 the classification report shows the F1 score, and eleven classes have a score of 1.00. Class A3 has an F1 score of 0.00 and classes F1 and F2 also did not perform well with just 0.67 and 0.60 F1 scores. These metrics are useful because some classes have more data than others.

	precision	recall	f1-score	support
A1	0.67	1.00	0.80	2
A2	1.00	1.00	1.00	2
A3	0.00	0.00	0.00	1
B1	1.00	1.00	1.00	5
B2	1.00	1.00	1.00	3
B3	0.75	0.75	0.75	4
C1	1.00	1.00	1.00	9
C2	0.71	1.00	0.83	5
C3	1.00	1.00	1.00	1
D	1.00	1.00	1.00	2
E	1.00	1.00	1.00	9
F1	0.67	0.67	0.67	3
F2	0.75	0.50	0.60	6
F3	1.00	1.00	1.00	6
G	1.00	1.00	1.00	4
H	1.00	1.00	1.00	3
I	1.00	1.00	1.00	5
accuracy			0.91	70
macro avg	0.86	0.88	0.86	70
weighted avg	0.91	0.91	0.91	70

Figure 5: Classification Report for Regression Tree

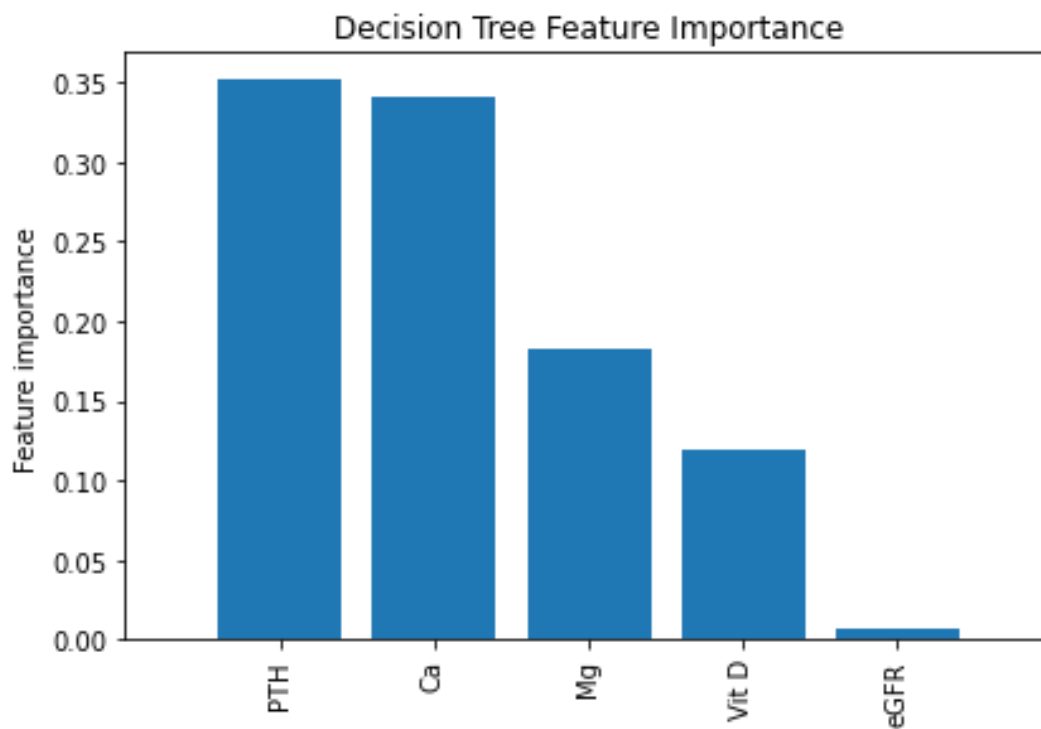


Figure 6: Regression Tree Feature Importance

of the project by correctly predicting most of the samples. There have been only five misclassifications, so it has performed better than a single regression tree.

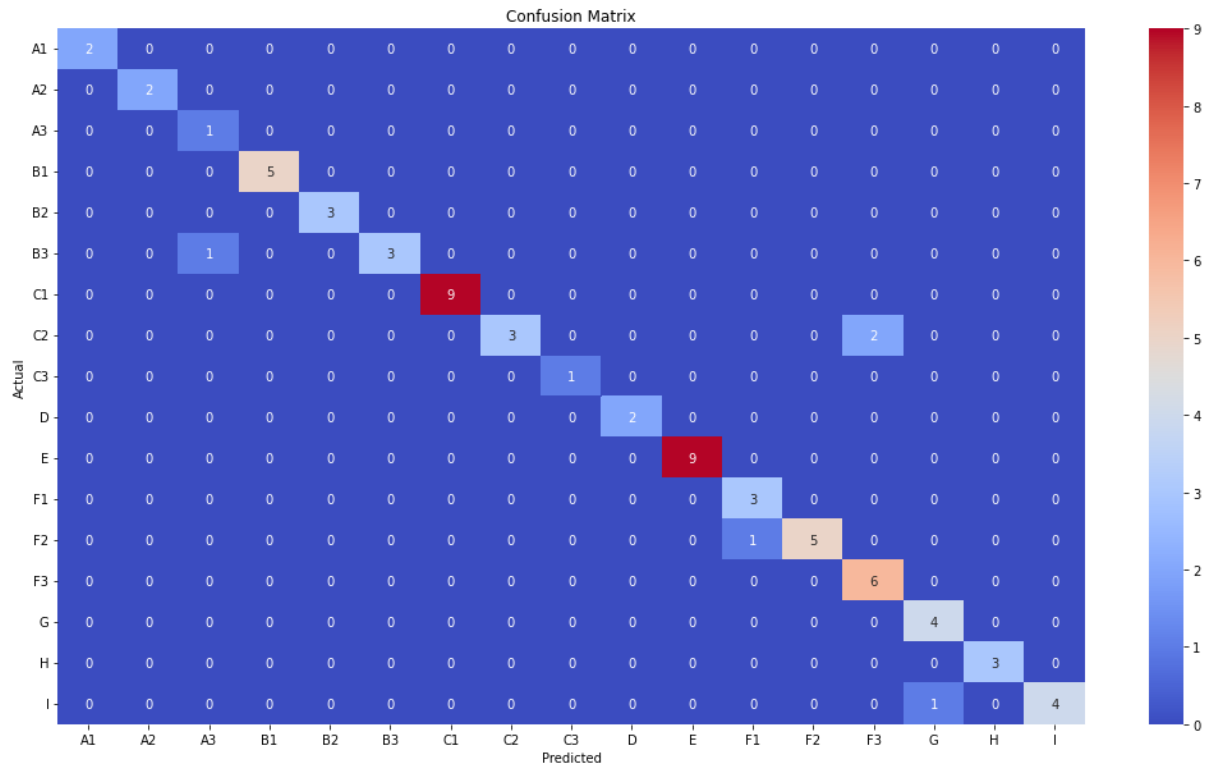


Figure 8: Confusion Matrix for Random Forest Classification

	precision	recall	f1-score	support
A1	1.00	1.00	1.00	2
A2	1.00	1.00	1.00	2
A3	0.50	1.00	0.67	1
B1	1.00	1.00	1.00	5
B2	1.00	1.00	1.00	3
B3	1.00	0.75	0.86	4
C1	1.00	1.00	1.00	9
C2	1.00	0.60	0.75	5
C3	1.00	1.00	1.00	1
D	1.00	1.00	1.00	2
E	1.00	1.00	1.00	9
F1	0.75	1.00	0.86	3
F2	1.00	0.83	0.91	6
F3	0.75	1.00	0.86	6
G	0.80	1.00	0.89	4
H	1.00	1.00	1.00	3
I	1.00	0.80	0.89	5
accuracy			0.93	70
macro avg	0.93	0.94	0.92	70
weighted avg	0.95	0.93	0.93	70

Figure 9: Classification Report for Random Forest

In figure 9 the statistics of what is shown in figure 8 confusion matrix is provided. Every class has at least one correct classification so there are no F1 scores with 0.00. In addition, in this experiment the weighted average of precision is 0.95 which is much higher than with the regression tree.

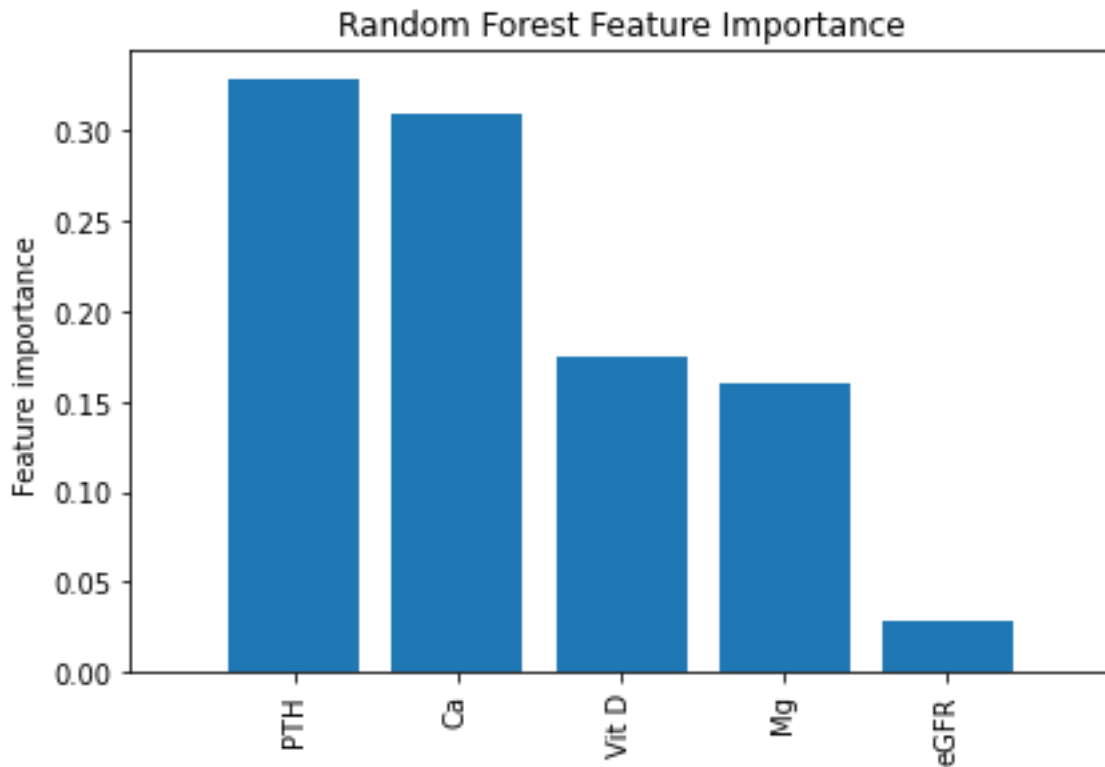


Figure 10: Random Forest Feature Importance

Figure 10 shows that PTH and calcium have the most importance just like in figure 6 with an importance score of over 0.30. However, for random forest vitamin D is given more importance than magnesium whereas in figure 6 it was the opposite. In addition, the random forest has given slightly more importance to eGFR compared to the single regression tree.

In figure 11 which shows tree 0 and tree 15, it can be seen that these trees in the random forest have a large depth and many decision rules because they are fully grown regression trees. This does not cause an overfitting problem like with single regression trees because aggregation of the multiple trees reduces overfitting and error caused by bias. A larger image of this figure can be found in appendix C.

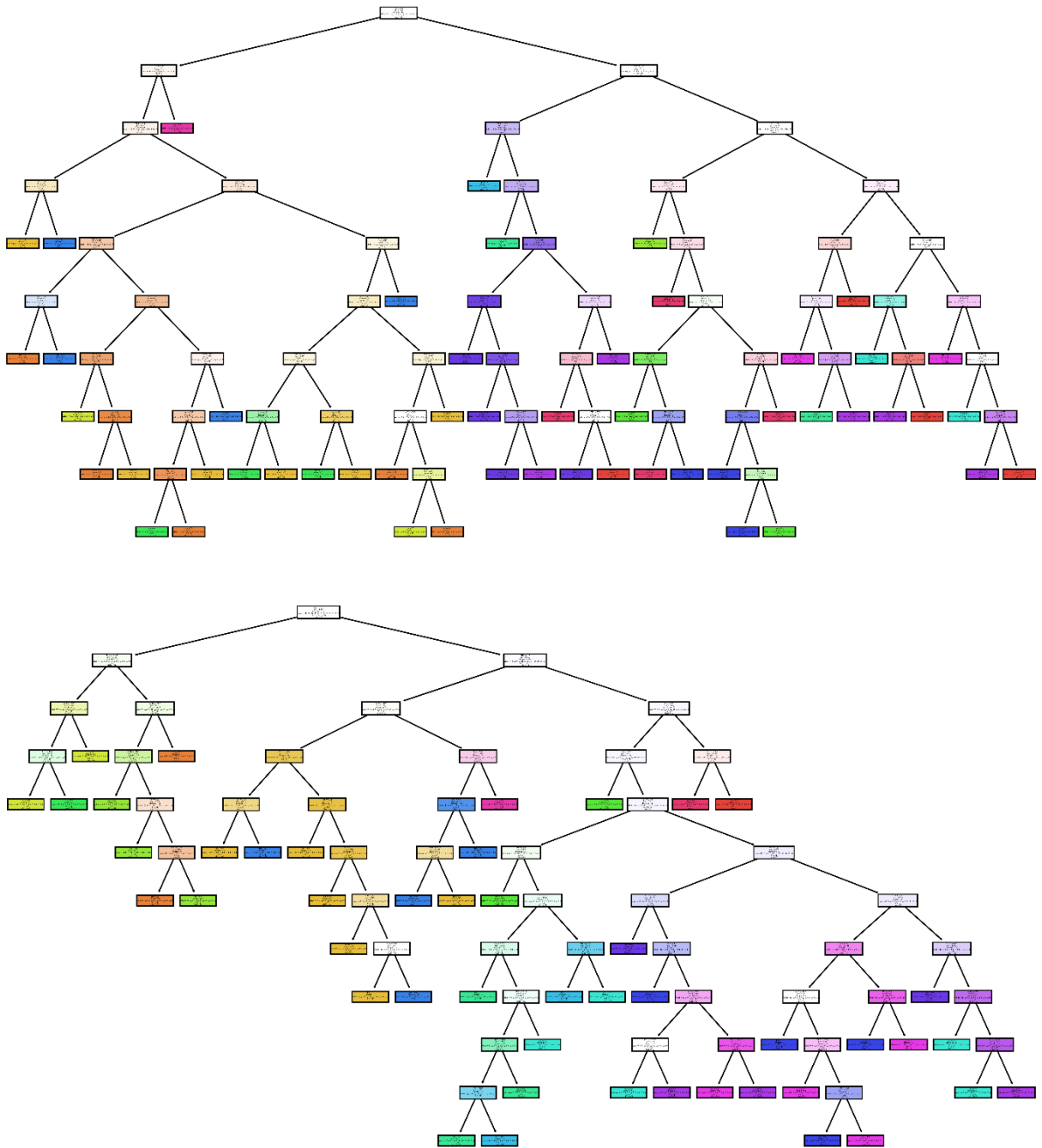


Figure 11: Trees in Random Forest

4.4 Pruned Regression Tree Predictions

Pruned regression trees limit overfitting by reducing the complexity of the regression tree so a higher accuracy score is expected. This is exactly what happened and an accuracy score of 0.94 was achieved by the model. But just like the unpruned regression tree, on figure 12 the pruned tree did not classify any samples correctly for A3 class. This is because this class was the most underrepresented class in the training set with only 8 samples. Only 4 samples

in the test set were misclassified so this model performed extremely well and has achieved the aim of the project.

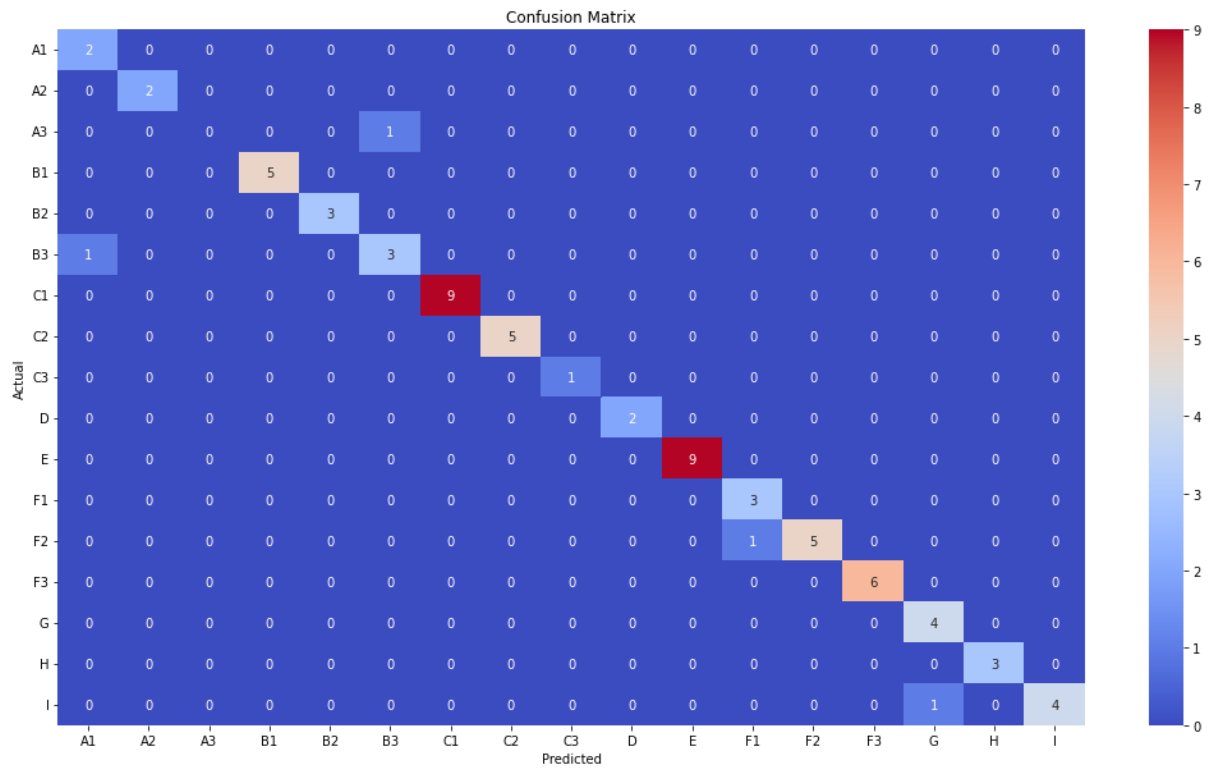


Figure 12: Confusion Matrix for Pruned Regression Tree

	precision	recall	f1-score	support
A1	0.67	1.00	0.80	2
A2	1.00	1.00	1.00	2
A3	0.00	0.00	0.00	1
B1	1.00	1.00	1.00	5
B2	1.00	1.00	1.00	3
B3	0.75	0.75	0.75	4
C1	1.00	1.00	1.00	9
C2	1.00	1.00	1.00	5
C3	1.00	1.00	1.00	1
D	1.00	1.00	1.00	2
E	1.00	1.00	1.00	9
F1	0.75	1.00	0.86	3
F2	1.00	0.83	0.91	6
F3	1.00	1.00	1.00	6
G	0.80	1.00	0.89	4
H	1.00	1.00	1.00	3
I	1.00	0.80	0.89	5
accuracy			0.94	70
macro avg	0.88	0.90	0.89	70
weighted avg	0.94	0.94	0.94	70

Figure 13: Classification Report for Pruned Regression Tree

In figure 13 the classification report shows that the precision macro average and weighted average are both lower than the precision scores for random forest model. However, the score for recall compared to the random forest model is lower too which is good. The F1 score takes the harmonic mean of precision and recall and the score for this model is 0.94, which is 0.01 higher than for the random forest model.

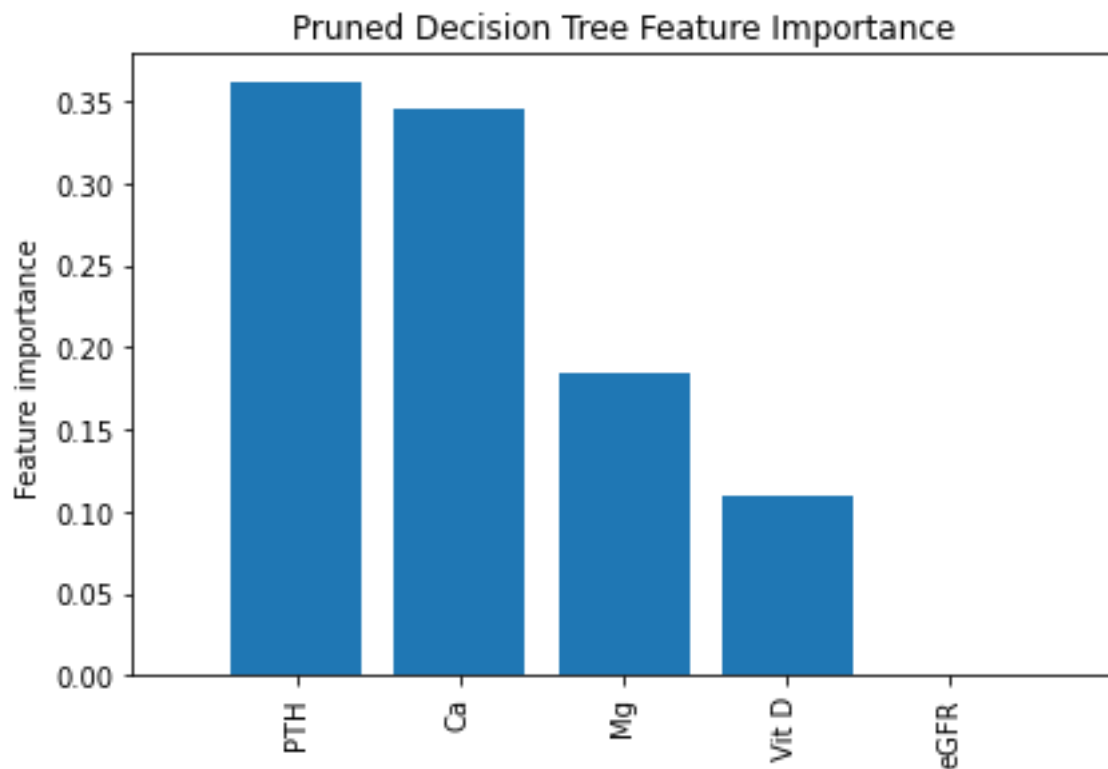


Figure 14: Pruned Regression Tree Feature Importance

Figure 14 shows the feature importance graph of the pruned regression tree. It is similar to the graph for the unpruned regression tree except for the eGFR feature. eGFR is completely removed by the pruned regression tree so it has 0.00 feature importance. This shows that eGFR is completely irrelevant in classification for this project. This is why there was nothing related to eGFR in chapter 2 for background research in relation to hypercalcaemia and hypocalcaemia.

In figure 15 the pruned tree is much simpler than the unpruned tree, and it only has a depth of 7 with just 21 rules. This reduction in complexity reduces overfitting on the model. In addition, figure 15 shows that there is no rule showing eGFR because it is given 0.00 importance by the model. A larger image of this figure can be found in appendix C.

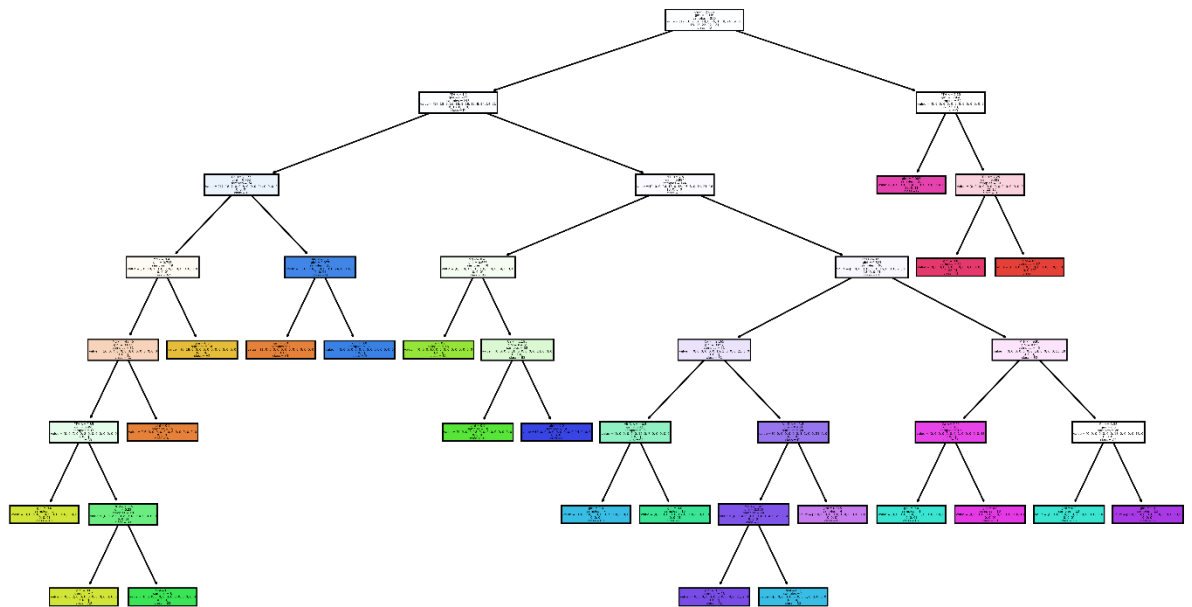


Figure 15: Pruned Regression Tree

Chapter 5

Validation of Results

This chapter will explain the validity of results by comparison of the decision tree rules against a decision tree given by a consultant. Also, the use of different metrics to obtain valid results will be discussed.

5.1 Comparison with Consultant's Decision Tree

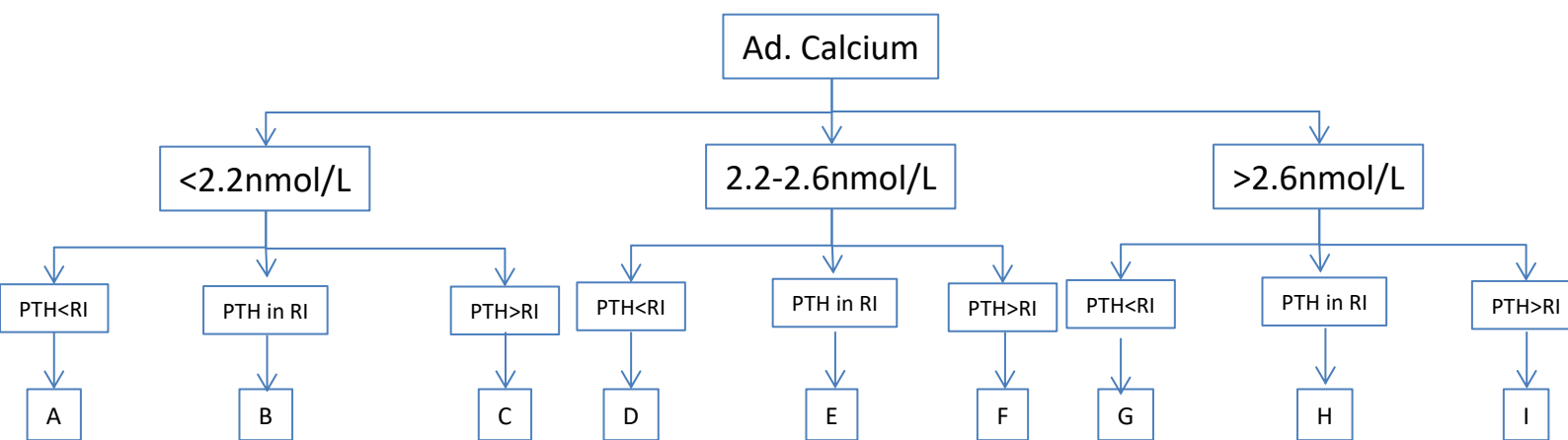


Figure 16: Consultant Decision Tree

In Figure 15 the root node checks whether calcium is less than or greater than 2.605. Checking for a similar calcium concentration of 2.6 happens in the consultant's decision tree too. Also, in the pruned tree there are many rules which check if concentration of calcium is greater than or less than 2.2. It is not exactly 2.2 but very close to 2.2 so this is similar to the consultant's tree too. In addition, there is one rule for PTH, checking if it is higher than or lower than 9.95. This is similar to what the consultant checks which is PTH concentration of 10 shown in figure 17 for class F. In the pruned decision tree continuing after this rule to the leaf nodes show that 2 out of 4 of the leaf nodes are class F. Therefore, this also proves the validity of the results. Furthermore, in the pruned tree there is checking for vitamin D level of 69.0 this is close to 75 mentioned in figure 17 which the consultant checks for. Lastly, there is checking of magnesium at concentration level of 0.45 and 0.5. This is similar to 0.4 which the consultant checks for and the leaf nodes after this have B classes and A class. So, the results are valid because it matches with the consultant's interpretations. The only difference is that there are some negative numbers for magnesium and vitamin D which is caused by the replacing of null values with -99. All of these similarities prove that the results are valid.

A:
-If Mg<0.4 Coded comment 'Hypocalcaemia secondary to magnesium deficiency (blunted PTH response). See <http://nww.lhp.leedsth.nhs.uk/common/guidelines/detail.aspx?ID=2105> for guidance on Mg replacement before repeat Ca.'
-If Mg >0.4 Coded comment 'Hypocalcaemia with low PTH suggestive of primary hypoparathyroidism, suggest referral to Endocrinology.'
B:
-If Mg <0.4 Coded comment 'Hypocalcaemia secondary to magnesium deficiency (blunted PTH response). See <http://nww.lhp.leedsth.nhs.uk/common/guidelines/detail.aspx?ID=2105> for guidance on Mg replacement before repeat Ca.'
-If Mg in >0.4 Coded comment 'Hypocalcaemia with inappropriately normal PTH suggestive of primary hypoparathyroidism, suggest referral to Endocrinology.' however if borderline low calcium suggest repeat in the first instance.
C:
-Vit D <75 Coded comment 'Secondary hyperparathyroidism due to vitamin D deficiency/insufficiency
-Vit D >75 [?historical](#) vitamin D deficiency/?Mg.'
D:
Coded comment, 'Normal calcium with low PTH unlikely to be significant.' If calcium low end of normal suggest repeat in 1-2months.
E:
Normal profile, however If PTH in top half of reference range and suspect primary hyperparathyroidism (EG osteoporosis/renal stones) then refer???
F:
-If Vitamin D <75 Coded comment, Secondary hyperparathyroidism due to vitamin D deficiency/insufficiency
-If Vitamin D >75 and PTH> 10 coded comment, [?normocalcaemic](#) primary hyperparathyroidism, suggest repeat in the first instance before considering referral to Endocrinology.
-If Vitamin D >75 and PTH <10 coded comment, 'Calcium within reference range and borderline raised PTH likely not significant. Unless primary hyperparathyroidism is suspected no follow up required.
G:
Coded comment 'Non-PTH dependant hypercalcaemia suggest further investigations'.
H:
Coded comment, 'Hypercalcaemia with PTH in the reference range consistent with primary hyperparathyroidism, suggest referral to Endocrinology.'
I:
Coded comment, 'Hypercalcaemia with raised PTH consistent with primary hyperparathyroidism, suggest referral to Endocrinology'

Figure 17: Consultant's Rules and Comments

5.2 Metrics

To evaluate the models scikit-learn's classification report was used. Apart from accuracy this includes other metrics which are precision, recall and F1 score. Accuracy is a skewed measure of performance for datasets that are highly imbalanced. F1 score takes the harmonic mean of precision and recall. How this is calculated is detailed in figure 18. In addition, confusion matrices are used which also show precision and recall. The use of these metrics makes the results valid.

$$PRE = \frac{TP}{TP + FP}$$
$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$
$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC}$$

Figure 18: Precision, Recall and F1 Score (Raschka, 2022)

Chapter 6

Conclusions and Future Work

This chapter provides a conclusion of the project and possibilities for future work.

6.1 Conclusions

This project has been successful in achieving the aim of the objective which is to compare machine learning algorithms and use best model to compare against consultant grade interpretations to determine the effectiveness. All of the objectives have been achieved and the best machine learning model produced an accuracy of 0.94. This accuracy can be improved more with a larger balanced dataset so that classification of class A3 is more precise.

The best model obtained was a pruned regression tree. This is a simple model with not too many rules and has many similarities to the consultant's decision tree. This demonstrates that machine learning and AI can be incorporated into the clinical field to harmonise reporting while reducing human input.

6.2 Future work

There are many opportunities for future work related to this project. Firstly, if more features such as CCCR, age, gender and phosphate are provided then more conditions can be diagnosed. For example, PHPT and FHH could be distinguished between.

Also, not having enough data for some classes caused the F1 score to be low for a few classes. Having a larger balanced dataset could improve the F1 score for classes.

In addition, other metrics such as AUROC (Area Under the Receiver Operating Characteristic) and Matthews Correlation Coefficient (MCC) could be used as an alternative way to evaluate the models.

Furthermore, other ensemble approaches could be considered such as boosting. The paper studied in 2.1.4 used AdaBoost to improve the accuracy the Bayesian network to 97.2%. Something similar could be done as future work.

Finally, machine learning algorithms can be optimised to give better performance. For example, the paper studied in section 2.1.1 used particle swarm optimisation algorithm to optimise SVM model's parameters. This achieved a very high accuracy, so it is worth investigating this further.

List of References

- Moody, B. (2021). *Calcium Regulation - Vitamin D - PTH*. [online] TeachMePhysiology. Available at: <https://teachmephysiology.com/biochemistry/electrolytes/calcium-regulation/>.
- NHS (2019). Overview - Blood tests. [online] NHS. Available at: <https://www.nhs.uk/conditions/blood-tests/>.
- S. Wang, Q. Jin and J. Wang, "A Support Vector Machine Based Classification Method for Hyper/Hypocalcemia Diagnosis," 2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS), 2021, pp. 1-5, doi: 10.1109/SAFEPROCESS52771.2021.9693645.
- Bertocchio JP, Tafflet M, Koumakis E, Maruani G, Vargas-Poussou R, Silve C, Nissen PH, Baron S, Prot-Bertoye C, Courbebaisse M, Souberbielle JC, Rejnmark L, Cormier C, Houillier P. Pro-FHH: A Risk Equation to Facilitate the Diagnosis of Parathyroid-Related Hypercalcemia. *J Clin Endocrinol Metab*. 2018 Jul 1;103(7):2534-2542. doi: 10.1210/je.2017-02773. PMID: 29727008.
- Witten, I. and Frank, E., n.d. *Data Mining*. San Diego: Elsevier Science & Technology Books.
- w3schools (n.d.). *Python Machine Learning Decision Tree*. [online] Available at: https://www.w3schools.com/python/python_ml_decision_tree.asp.
- scikit learn (2009). 1.10. *Decision Trees* — *scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/tree.html>.
- w3schools. (n.d.). *Python Machine Learning - K-nearest neighbors (KNN)*. [online] Available at: https://www.w3schools.com/python/python_ml_knn.asp [Accessed 23 Aug. 2022].
- scikit-learn. (n.d.). *sklearn.discriminant_analysis.LinearDiscriminantAnalysis* — *scikit-learn 0.24.1 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html.
- Somnay, Y.R., Craven, M., McCoy, K.L., Carty, S.E., Wang, T.S., Greenberg, C.C. and Schneider, D.F. (2017). Improving diagnostic recognition of primary hyperparathyroidism with machine learning. *Surgery*, 161(4), pp.1113–1121. doi:10.1016/j.surg.2016.09.044.
- Waikato. (2019). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [online] Available at: <https://www.cs.waikato.ac.nz/ml/weka/>.

Jason Brownlee (2019). Your First Machine Learning Project in Python Step-By-Step. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>.

Raschka, S., 2022. [online] Sebastianraschka.com. Available at: <https://sebastianraschka.com/pdf/lecture-notes/stat451fs21/13_feat-sele__slides.pdf> [Accessed 23 August 2022].

Raschka, S. (2022). *What is the best validation metric for multi-class classification?* [online] Dr. Sebastian Raschka. Available at: <https://sebastianraschka.com/faq/docs/multiclass-metric.html> [Accessed 24 Aug. 2022].

Appendix A

External Materials

```
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))
#models.append(('RF', RandomForestClassifier()))

# evaluation of each model
results = []
names = []
for name, model in models:
    kfold = StratifiedKFold(n_splits=7, random_state=1, shuffle=True)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```

Figure 19: Models and Cross Validation (Jason Brownlee, 2019)

```
indices = np.argsort(importances[::-1])

plt.ylabel('Feature importance')
plt.bar(range(X_train.shape[1]),
        importances[indices],
        align='center')

feat_labels = dataset.columns[0:5]
plt.xticks(range(X_train.shape[1]),
           feat_labels[indices], rotation=90)

plt.xlim([-1, X_train.shape[1]])

plt.tight_layout()
plt.title("Random Forest Feature Importance")
#plt.savefig('feature-importance.pdf', dpi=300)
plt.show()
```

Figure 20: Feature Importance Code (Raschka, 2022)

Appendix B

Ethical Issues Addressed

This project does not have any ethical issues because there is no personal data used that can be used to trace back to the patients.

Appendix C

Decision Trees

The link below contains the images of decision trees.

https://leeds365-my.sharepoint.com/:f:/g/personal/sc21st_leeds_ac_uk/ErXlcF4QaHNEh3YDr6nZqsgBmZ4EDOojVvmEN8hplP4XhA?e=IFKayr