The Ten Thousand Patient Dataset:

A Study of the Diagnosis of Pulmonary Embolisms

Dr D.C.Horgan BSc(Hons),MSc,MA,PhD,CPhys,MinstP, FRAS

1 Introduction

This article describe the data analysis undertaken on a dataset of 10,000 Patient records. The objective of the analysis was to identify patients suffering from a Pulmonary Embolism as defined in the ICD10 codes I26-I28. The analysis of the dataset was undertaken using a cluster, hosted at CoCalc which hosts a range of projects.The text documents were converted to csv format and uploaded into separate Pandas DataFrames. Column names were produced for each of the columns in the pandas DataFrame.

2 Exploratory Data Analysis

An exploratory data analysis wa undertaken using Pandas DataFrames. The head command was to verify the formatting of the pandas DataFrame. Using pandas groupBy and chaining the frequency of each diagnosis was calculated and horizontal barcharts produced using matlibplot as shown in fig. 1

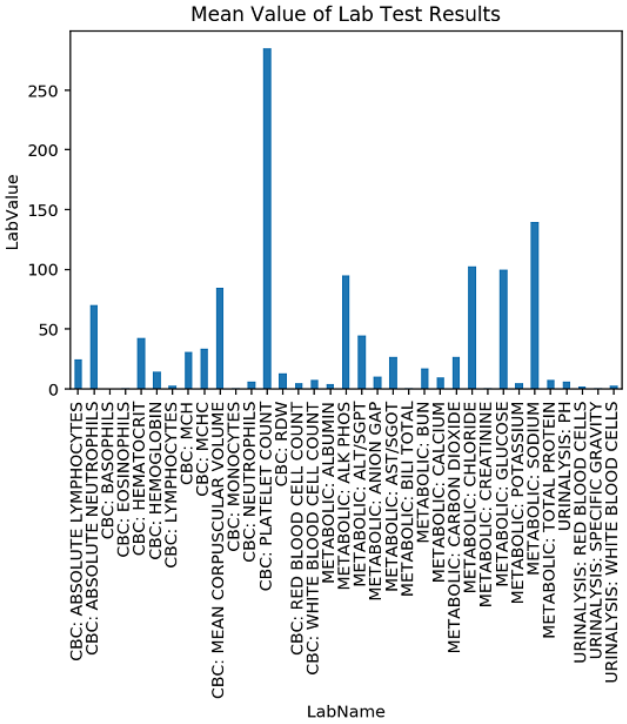

Figure 1:

This analysis was also performed using the PrimaryDiagnosisCode, which was an international standards organisation (ISO) ICD10 code which is use to label medical conditions around the world this is shown in fig.2:

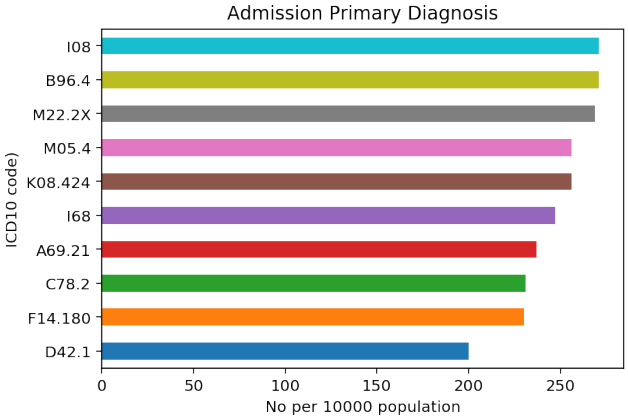

Figure 2:

Since a majority of the dataset consisted of laboratory data values a thorough examination of these was made.Using groupBy and other commands as shown in fig. 3

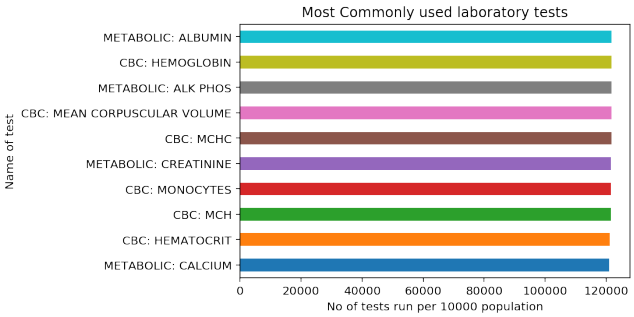

Figure 3:

The mean laboratory results were calculated for both the sample population, the I26-28 code group and for the (sample - ICD10 Pulmonary Embolism) population. The values for the sample population are shown in fig.4:

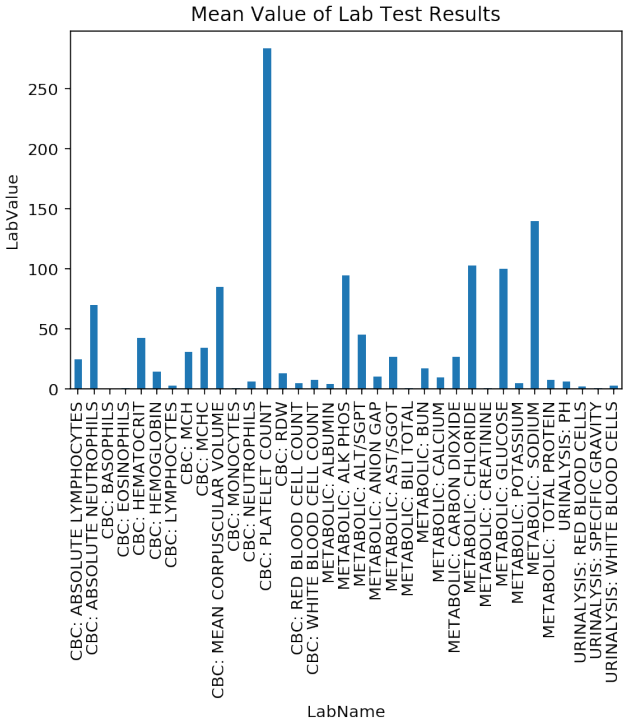

Figure 4:

and for the the ICD10 Pulmonary Embolism population in **??**

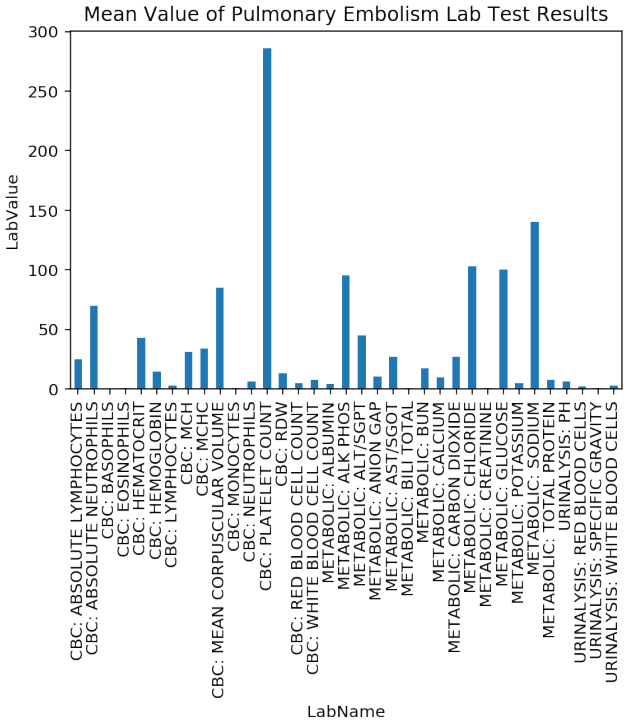

Figure 5:

Comparing the Lab test results from both the ample population and the (sample - ICD10 Pulmonary Embolism)population shows no noticeable difference between the two groups as seen in fig.6

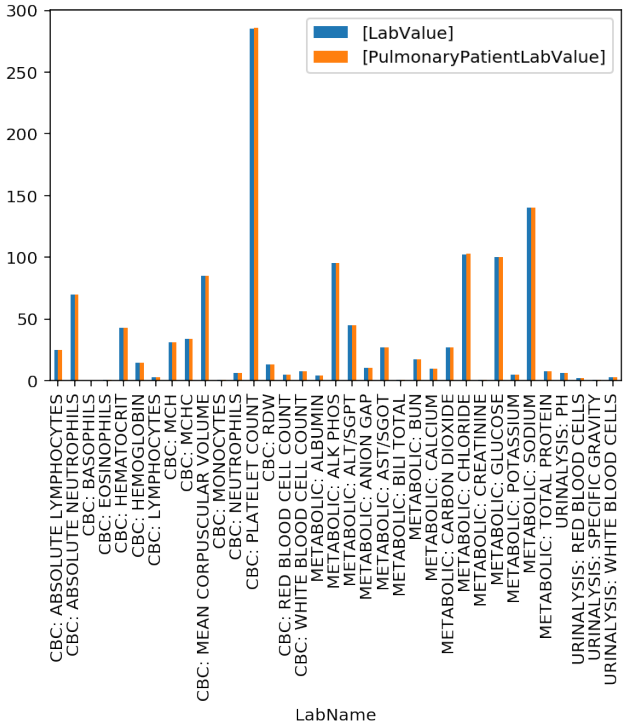

Figure 6:

This observation was confirmed by a statistical analysis using the T-test which had a T statistic of -0.0019 and a p value of 0.9984,indicating that the laboratory test results were not significantly different to each other.
This process was repeated using the F2 ICD10 codes for a neurological disorder. Again there was no obvious difference in the laboratory test results as shown in fig.7

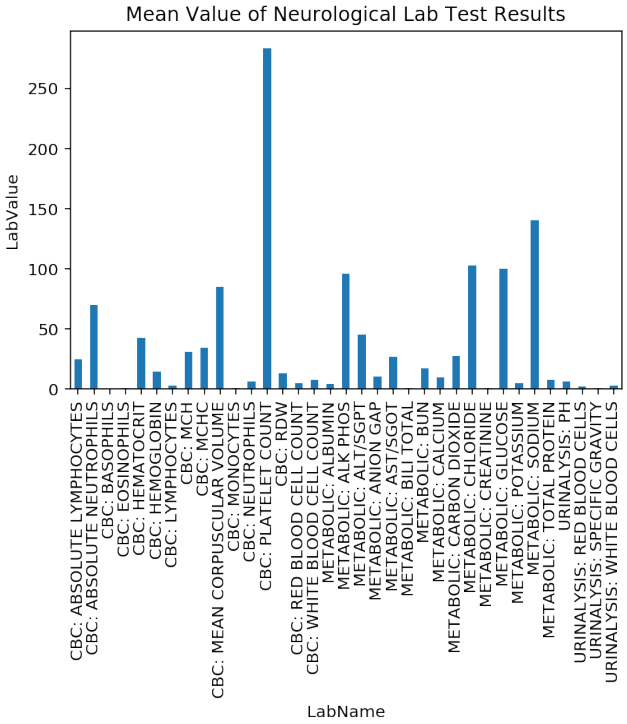

Figure 7:

This was confirmed using the T-Test which gave the T statistic: -0.0017and p value=0.9985
Three batches of Laboratory Test results (neurological, pulmonary Embolism and control group) were analysed and compared as shown in fig.8

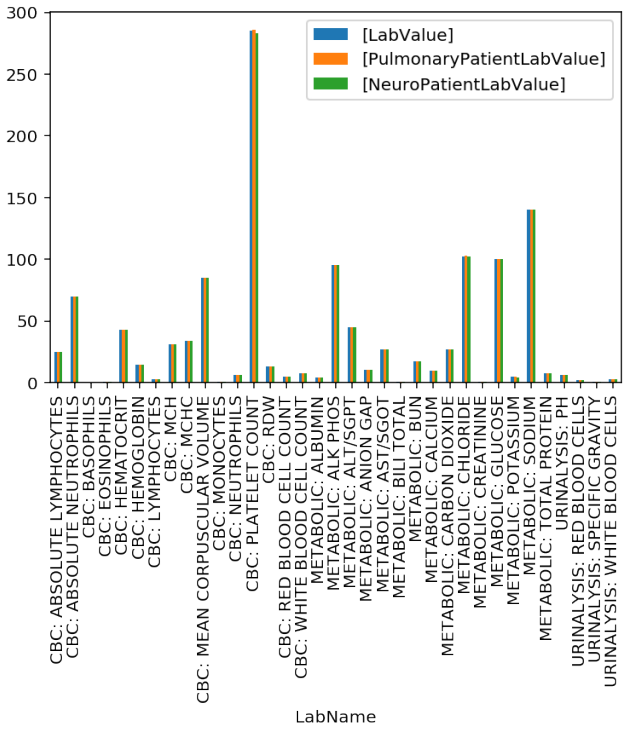

Figure 8:

There was no visible difference between these groups and there was also no statistical difference between them as measured by the statistical T-Test of 0.0017 and pvalue of 0.9985

3 Data Analysis

3.1 Preprocessing

The datasets were ingested into Pandas DataFrames and concatenated and merged to produce a Pandas DataFrame of features - principally the LabNames. The PrimaryDiagnosisCodes being used as the target.The categorical data in LabName and PrimaryDiagnosisCodes preprocessed using scikit one-hot and then standardised.The preprocessed dataset was then split using scikit xyspitter-train into X-train, X-test, y-train and y-test sections.

4 Classification and Regression Analyses

4.1 ML Models

The following models were used:

DecisionTreeClassifier

RandomForestClassifier

LogisticRegression

SupportVectorMachine

KNeighborsClassifier

BBayes-model(GaussianNB)

4.1.1 DecisionTreeClassifier

An analysis was undertaken using scikit-kit learn. In this scikit-learn DecisionTreeClassifier was used to analyse if the ICD10 codes I26-28 which are relevant to Pulmonary Embolism could be differentially classified using the values of the laboratory readings.

The Pandas Dataframes were examined and the main features selected and confirmed using Logistic regression. The codes in the PrimaryDiagnosisCodes DataFrame were then one-hot encoded. The LabNames and gender were also one-shot encoded and the LabValues left as numerical values. The data was then split into training, test and validation parts using np.split. then trained using scikit-learn DecisionTreeAnalysis. This technique was chosen because it was important that the trained machine Learning Model produced be transparent and explicable to users. After training the model was used to predict the ICD10 codes of the Patient.

A graph of the decision tree classification was produced which showed the classification path to each of the PrimaryDiagnosisCodes used in the sample. The overall size of the trained DecisionTree model was extensive as shown in fig.9 and fig.10

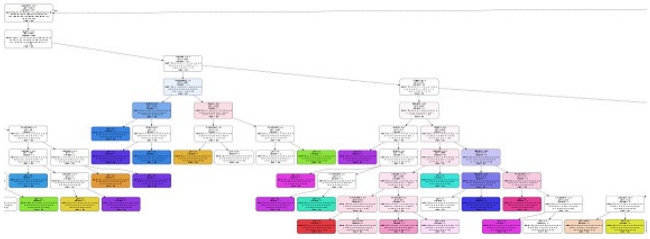

Figure 9:

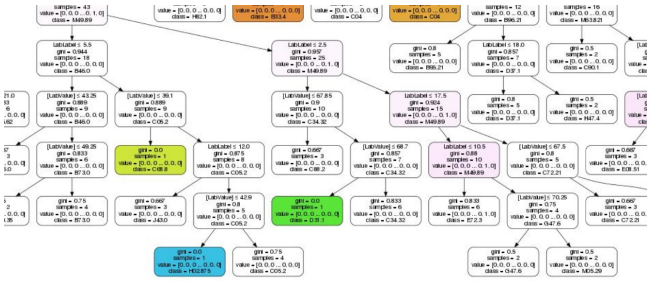

Figure 10:

The analysis using scikit-kit learn DecisionTreeClassifier used to see if the ICD10 codes I26 - I28 relevant to Pulmonary Embolism could be deferentially classified was quite successful and produced a Decision Tree Model which could be used within an application.

4.1.2 RandomForestClassifier

The scikit-learn Ramdom Forest Classifier was used to train the data and identify important features.

Features sorted by their score

0.8449 - LabValue

0.075 - year

0.0508 - PatientPopulationPercentageBelowPoverty

0.0195 PatientGender

0.0026 - PrimaryDiagnosisCode

0.0023 - White

0.0021 - Asian

0.0015 - Unknown

0.0011 - AfricanAmerican

This is shown graphically in fig.11

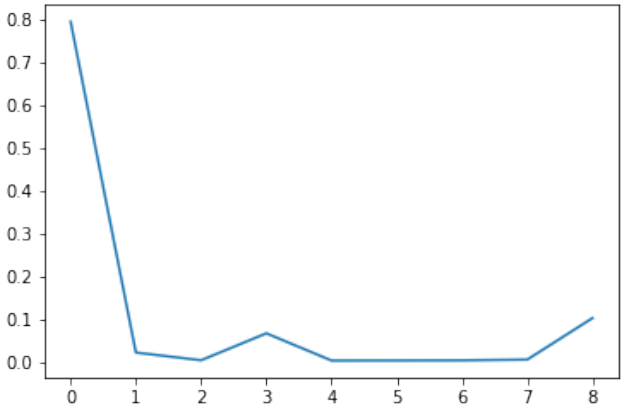

Figure 11:

The confusion matix for the random tree clasifier shown as an array and heatmap in fig.12

38	1147
673	16841

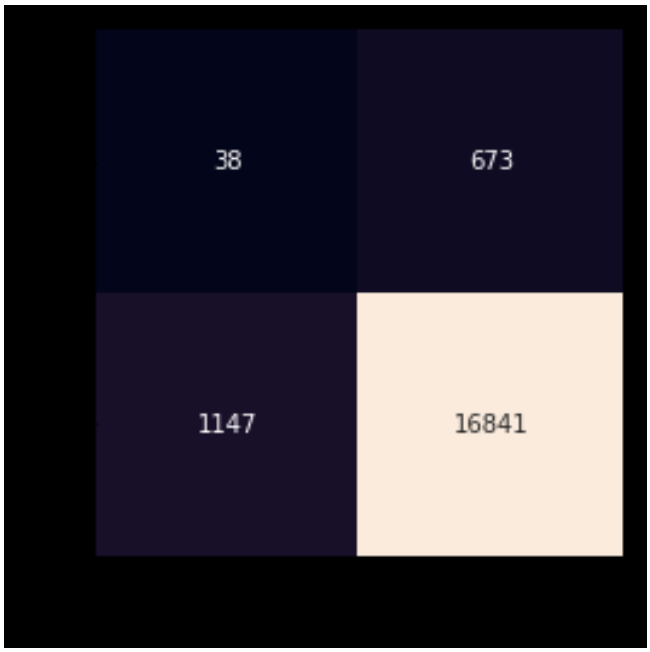

Figure 12:

4.1.3 LogisticRegression

Scikit Logistic regression was used and the training accuracy in terms of finding Pulmonary Embolism was 0.9373 whilst the test accuracy was 0.9366

The Logistic Regression confusion matrix is show below:

0	1185
0	17514

and as a heatmap in fig.13

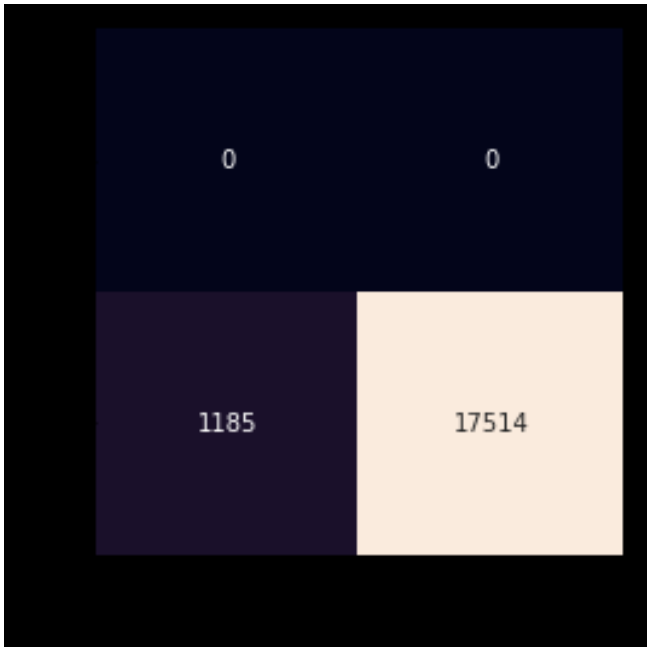

Figure 13:

metrics for classification:

precision	recall	f1-score	support
0 0.00	0.00	0.00	1185
1 0.94	1.00	0.97	17514

4.1.4 DecisionTreeClassifier

The training accuracy using binary classification of ICD10codes into Pulmonary Embolism(PE) and Not Pulmonary Embolism (not PE) was = 0.9827.

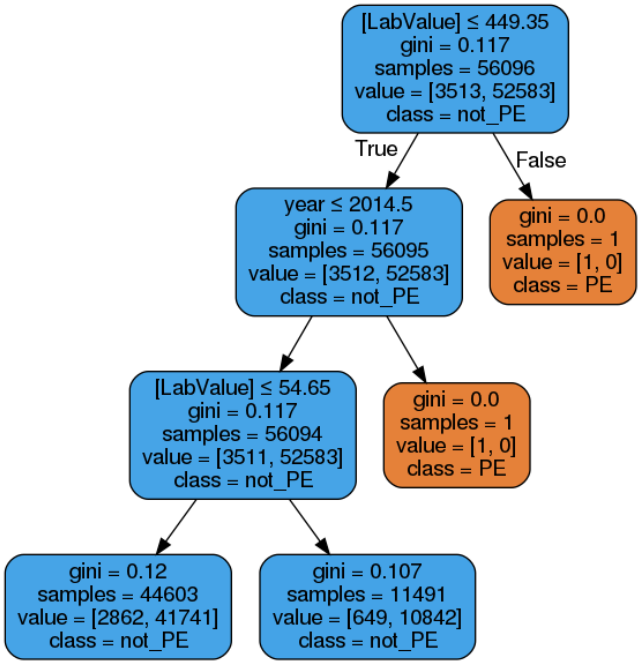

Figure 14:

4.1.5 SupportVectorMachine

The confusion matrix for the SVM model is:

0	1185
0	17514

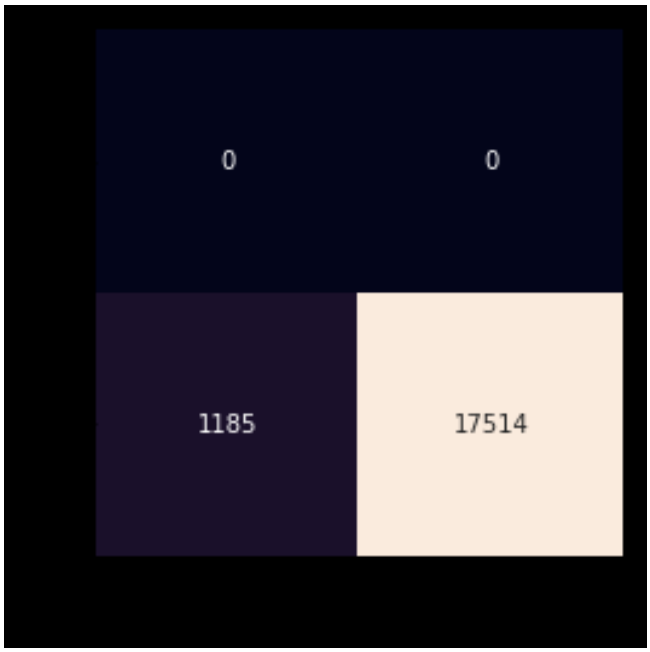

Figure 15:

4.1.6 KNeighborsClassifier

The training accuracy of the KNeighborsClassifier model with n=3 was 0.9405.

The KNeighborsClassifier confusion matrix and heatmap is shown below fig.16:

10	1175
245	17269

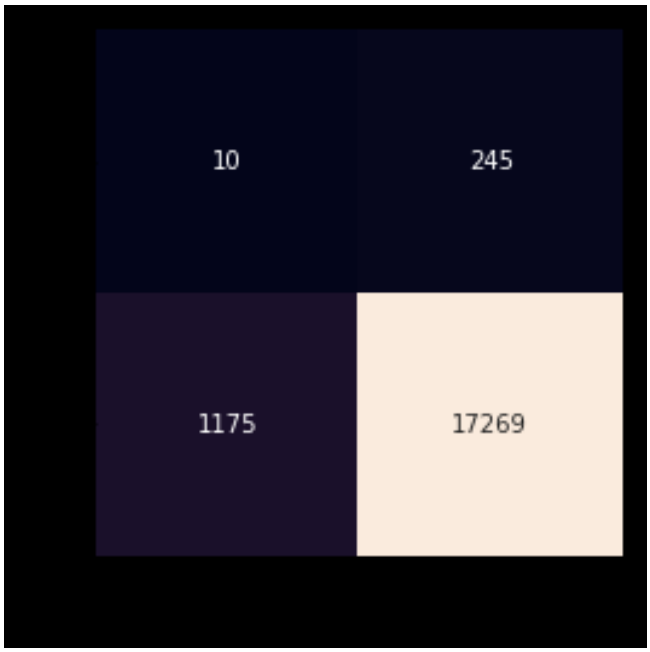

Figure 16:

4.1.7 Bayes-model(GaussianNB)

The training accuracy for bayes-model was 0.9373. The confusion matrix and heatmap for the Bayes model is shown below:

0	1185
0	17514

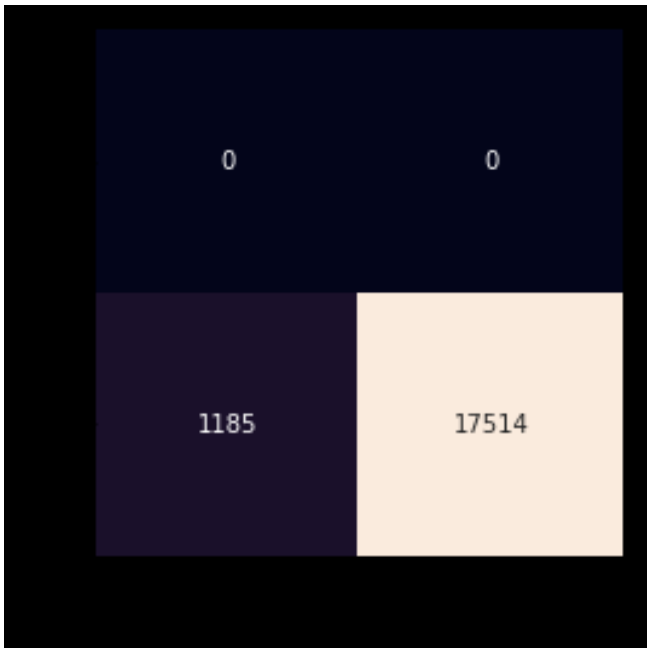

Figure 17:

4.1.8 comparison of classifiers

measure	Decision Tree	Random forest	svc model	knn neighbours	Bayes
accuracy	0.936574	0.902669	0.936628	0.924060	0.936628
precision	0.936624	0.936235	0.936628	0.936294	0.936628
recall	0.999943	0.961574	1.000000	0.986011	0.967277
f1	0.967248	0.948735	0.967277	0.960509	0.967277

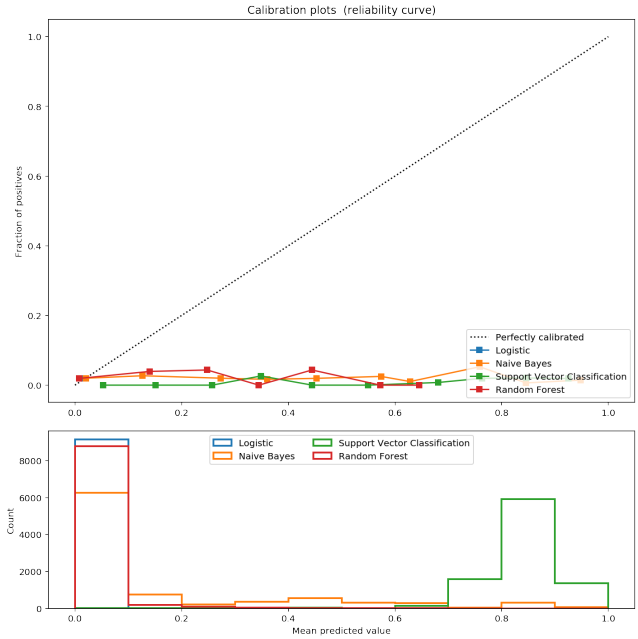

Figure 18:

5 Correlation Analysis of ICD10 codes

The exploratory data analysis indicated that the tests commonly used at admission would not in themselves provide a basis for diagnosing Pulmonary embolism. The Classification analysis indicated that factors such as age, income, gender and ethnicity in combination with the LabValues could have diagnostic value and in particular that a Decision Tree Classifier could produce a good level of prediction.

This analysis looks at other disorders which may be associated with Pulmonary Embolism and which therefore might be a useful sign that a Pulmonary Embolism should also be considered.

The disorders looked at were;
The following ICD10 codes were investigated:

I21 Acute myocardial infarction

I23 current complications following ST elevation myocardial infarction

I24 acute ischemic heart diseases

I25 Chronic ischemic heart disease

I26 Pulmonary embolism

I27 pulmonary heart diseases

I28 diseases of pulmonary vessels

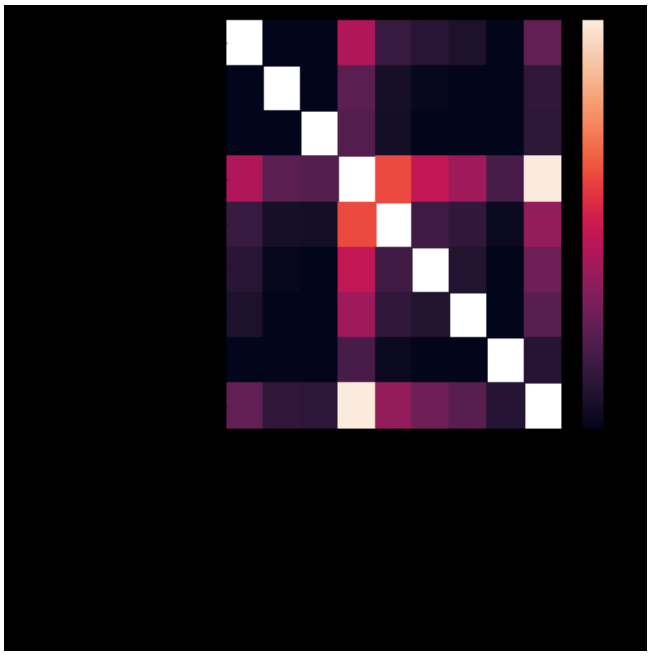

Figure 19:

I82 venous embolism and thrombosis

I97 Intraoperative and postprocedural complications and disorders of circulatory system

6 Conclusions

Investigating associated illnesses has been quite successful,
and we can say that:

Pulmonary embolism is strongly associated with Chronic ischemic heart disease.

Chronic ischemic heart disease is moderately associated with myocardial infarction

Chronic ischemic heart disease is moderately associated with pulmonary hypertension

Chronic ischemic heart disease is weakly associated with Acute coronary thrombosis

An actionable insight from this is that a patient presenting with Chronic ischemic heart disease, primary pulmonary hypertension, Acute coronary thrombosis or a myocardial infarction should be considered at risk from Pulmonary Embolism as well and that further investigation perhaps using a D-dimer test would be beneficial.

7 Appendices

Appendix A: A description of the Dataset
The dataset consists of four tables:

PatientCorePopulatedTable

PatientID - a unique ID representing a patient.

PatientGender - Male/Female.

PatientDateOfBirth - Date Of Birth.

PatientRace - African American, Asian, White.

PatientMaritalStatus - Single, Married, Divorced, Separated, Widowed.

PatientLanguage - English, Icelandic, Spanish.

PatientPopulationPercentageBelowPoverty - given in percent

AdmissionsCorePopulatedTable

PatientID - a unique ID representing a patient.

AdmissionID - an admission ID for the patient.

AdmissionStartDate - start date.

AdmissionEndDate - end date.

AdmissionsDiagnosesCorePopulatedTable

PatientID - a unique ID representing a patient.

AdmissionID - an admission ID for the patient.

PrimaryDiagnosisCode - ICD10 code for admission's primary diagnosis.

PrimaryDiagnosisDescription - admission's primary diagnosis description.

LabsCorePopulatedTable

PatientID - a unique ID representing a patient.

AdmissionID - an admission ID for the patient.

LabName - lab's name, including:

CBC: WHITE BLOOD CELL COUNT

CBC: RED BLOOD CELL COUNT

CBC: HEMOGLOBIN

CBC: HEMATOCRIT

CBC: MEAN CORPUSCULAR VOLUME

CBC: MCH

CBC: MCHC

CBC: RDW

CBC: PLATELET COUNT

CBC: ABSOLUTE NEUTROPHILS

CBC: ABSOLUTE LYMPHOCYTES

CBC: NEUTROPHILS

CBC: LYMPHOCYTES

CBC: MONOCYTES

CBC: EOSINOPHILS

CBC: BASOPHILS

METABOLIC: SODIUM

METABOLIC: POTASSIUM

METABOLIC: CHLORIDE

METABOLIC: CARBON DIOXIDE

METABOLIC: ANION GAP

METABOLIC: GLUCOSE

METABOLIC: BUN

METABOLIC: CREATININE

METABOLIC: TOTAL PROTEIN

METABOLIC: ALBUMIN

METABOLIC: CALCIUM

METABOLIC: BILI TOTAL

METABOLIC: AST/SGOT

METABOLIC: ALT/SGPT

METABOLIC: ALK PHOS

URINALYSIS: SPECIFIC GRAVITY

URINALYSIS: PH

URINALYSIS: RED BLOOD CELLS

URINALYSIS: WHITE BLOOD CELLS

LabValue - lab's value

LabUnits - lab's units.

LabDateTime - date.