**INSTITUTE OF HEALTH INFORMATICS**

Graduate Programme in Health Informatics

# Assessed Coursework Submission

| | |
|---|---|
| **Student candidate number:** | ***** |
| **Module:** | CHME0016: Machine Learning in Healthcare and Biomedicine |
| **Date due:** | Monday, 13th April 2019, 12:00 midday |
| **Word count:** (excluding references, diagrams and appendices) | 2395<br><br>ipython notebook can be found at the following link: https://github.com/DarrylBourke/ML----Assignment-Submission |
| **Disability or other medical condition** for which UCL has granted special examination arrangements: | |
| **Formative feedback:** | Please address in formative feedback:<br>Yes please.<br>Please ignore in formative feedback: |

**MLHB Assignment - ILPD (Indian Liver Patient Dataset).**

ipython notebook can be found at the following link:
https://github.com/DarrylBourke/ML----Assignment-Submission

**A Comparison of Machine Learning Approaches to Predict Liver Disease.**

**Introduction**

Liver disease can be caused by alcohol consumption, obesity, viral infection, heredity illness or an immuno-response deficiency. Patients with liver disease including cirrhosis of the liver can have symptoms such as weakness, appetite loss, feeling sick and jaundice (NHS online A, 2019).

Annually in North America and Europe, there are >70000 hospital visits due to acute liver failure and cirrhosis (Harrison M, 2018). Liver cirrhosis causes 1.8% of total deaths in Europe and rates of liver disease are increasing in the UK (Blachier, M et al., 2013). The rate of cirrhosis amongst UK males being 14.9 and females 7.6 per 100000 population (Blachier, M et al., 2013), with an overall prevalence of between 4-5% (Pimpin, L et al., 2018). Among Indian people =>15 years old, liver cirrhosis accounted for the deaths of 45.8 of males and 14.7 of females per 100000, (Global Health Observatory data repository, 2019).

Liver disease is diagnosed using blood tests, viral testing, biopsies, immunology tests, endoscopic procedures and imaging techniques (British Liver Trust A, 2019). Imaging techniques used may include ultrasound, Computed tomography (CT) and magnetic resonance imaging (MRI), (Procopet et al., 2017). Blood tests are the most common and are the focus of the dataset used in this analysis. Liver function blood tests check for normal levels of enzymes and proteins in order to make an assessment and include Alanine aminotransferase (Sgpt), aspartate aminotransferase (Sgot), alkaline phosphatase (Alkphos), bilirubin (TB) and albumin (ALB), (British Liver Trust B, 2019). Normal range for Sgpt is <35 IU/L, Sgot is <40 IU/L where a high level of either can indicate liver disease. Alkphos should be <104 IU/L, TB is <24 umol/L and ALB is between 34 to 50 g/L for normal liver function, (Thriva health hub, 2019). These features have been used for predictive modelling in literature sources, (Ramana, B V, 2012).

The aim of the following analysis was to compare the predictive performance of two machine learning models, support vector classification and random forest classification, with regard to liver disease.

**Methodology**

Preliminaries and data cleansing

This was carried out using Python 2.7, scikit-learn, numpy, pandas, matplotlib, seaborn and stats packages running on a MacBook Air 2015 (macOS High Sierra) with the output presented as an ipython notebook.

The ILPD (Indian Liver Patient Dataset) dataset was uploaded from the UCI Machine-Learning Repository, and comprised 570 patients (406 liver patients and 164 liver disease free patients) aged 4 to 90+ years (mean age 44.8 years) of whom 71.23% were male.

The data included ten features and one class label (Selector). Gender was recoded from categorical to binary, while the Selector was recoded to binary. The data was checked for duplicates.

There were four missing data values for A/G Ratio, which were replaced with the mean of all A/G Ratio values. Mean values were used to replace null values. Label encoding and one hot encoding were used to complete the conversion of nominal categorical features (Gender) to a numerical format for the same reason.

Data Exploration

Histograms were used to visualise the distribution of each feature in relation to each other. Box-plots were used to show the variance, mean and confidence intervals for each feature, while providing information on potential outliers.

A correlation-matrix was plotted to provide details on feature correlation for feature importance information. Additionally, statistically significant p-values were used to determine feature importance, and a random forest regressor object was generated to calculate and plot the feature importance.

71.23% were positive for liver disease, representing a data imbalance ratio of 2.48 : 1. Plots were also used to explore age and gender distributions.

Potential data leakage was dealt with by removing duplicate records. Potential outliers were investigated using an elliptical envelope detector command with outliers indicated in an array as -1 per instance. Based on feature importance analysis, Alkphos, Sgot, Sgpt and Age features were rescaled using the log of the values. These were then plotted in two feature space plots for explorative purposes. The latter plot used features selected as per literature and supported by the feature importance analysis and the former plot used features selected as suggested by their statistical significance (p-values). No outlier data points were excluded.

Training and Testing the Dataset

The dataset was split in to 75% (427 instances) for training and 25% (143 instances) for testing purposes. Standardisation was applied to the dataset using the scikit-learn standard scalar function. This function centred features at the mean 0 and giving them a standard deviation of 1.

Hyper-parameter optimisation

Hyper-parameter optimisation was performed on both chosen supervised learning algorithms. The operation uses a GridSearchCV function and a grid of different parameter variables as defined by the user to test for the optimal parameter combination. Aside from a set random state to improve repeatability, the parameters vary between classifiers, as can parameter ranges evaluated by the grid search. To optimise the computation time, the parameter ranges were limited (the number was adjusted through trial and error). All

features were included in the main analysis and PCA was used in a further sensitivity analysis step. The output was a list of the optimised parameters and an accuracy score.

Random Forest Classification (RFC)

Random forest classification, sees the average output of a collection of decision trees used to select an optimised list of parameters that can be applied to the training and testing sets. Alongside a user defined list of parameters, randomised training data was used for the decision trees generation, with each attempt using a different randomised training data sample.

Features represented by nodes in the structure were split into leaves depending upon them meeting a particular threshold. As the number of splits increased, the decision for each subsequent split became more computationally demanding. Therefore, to reduce the computational complexity, features of lower importance could be removed using Principle Component Analysis (PCA), which could limit the potential for overfitting. The aggregated output was given an accuracy score representing the effectiveness of classification fit. A key benefit of RFCs is the reduced likelihood of overfitting and thus increased bias than when using single decision trees.

Support Vector Classification (SVC)

Support vector classification creates a hyperplane separating features by their binary classification. Depending on the hyperplane fit, there can be over-fitting or under-fitting of the hyperplane. As with the RFC, a list of user defined parameters was applied to a GridSearch to determine the optimal parameters for the training and testing sets. An accuracy score for effectiveness of fit was also returned.

The SVC model in this report was tested for both a linear kernel or a radial basis function (RBF) kernel. The linear kernel was best suited to where there was a straight hyperplane, while the RBF kernel can curve to better fit the feature space. Fitting of the RBF was strongly affected by the gamma and the C parameter. Gamma parameter values of 1.0 gave curved decision boundary fits, over lower gamma parameters <1.0 (straighter hyperplane). Higher gamma parameter values (>100) led to overfitting. High gamma values use nearby data points, which result in a closer fit to the hyperplane, as is the opposite for lower gamma values. With increasing C value the decision boundary fitting the data more closely. However, with optimal C values the output can be generalizable by avoiding over-fitting.

Evaluation of classifiers

Using the 75% training and 25% testing data split, an initial testing set prediction was performed on both the following models. After which, the GridSearchCV function was applied to get an accuracy score and best fit parameters. A further GridSearchCV function was applied for the Area under the curve (AUC) scores for both the training and testing sets.

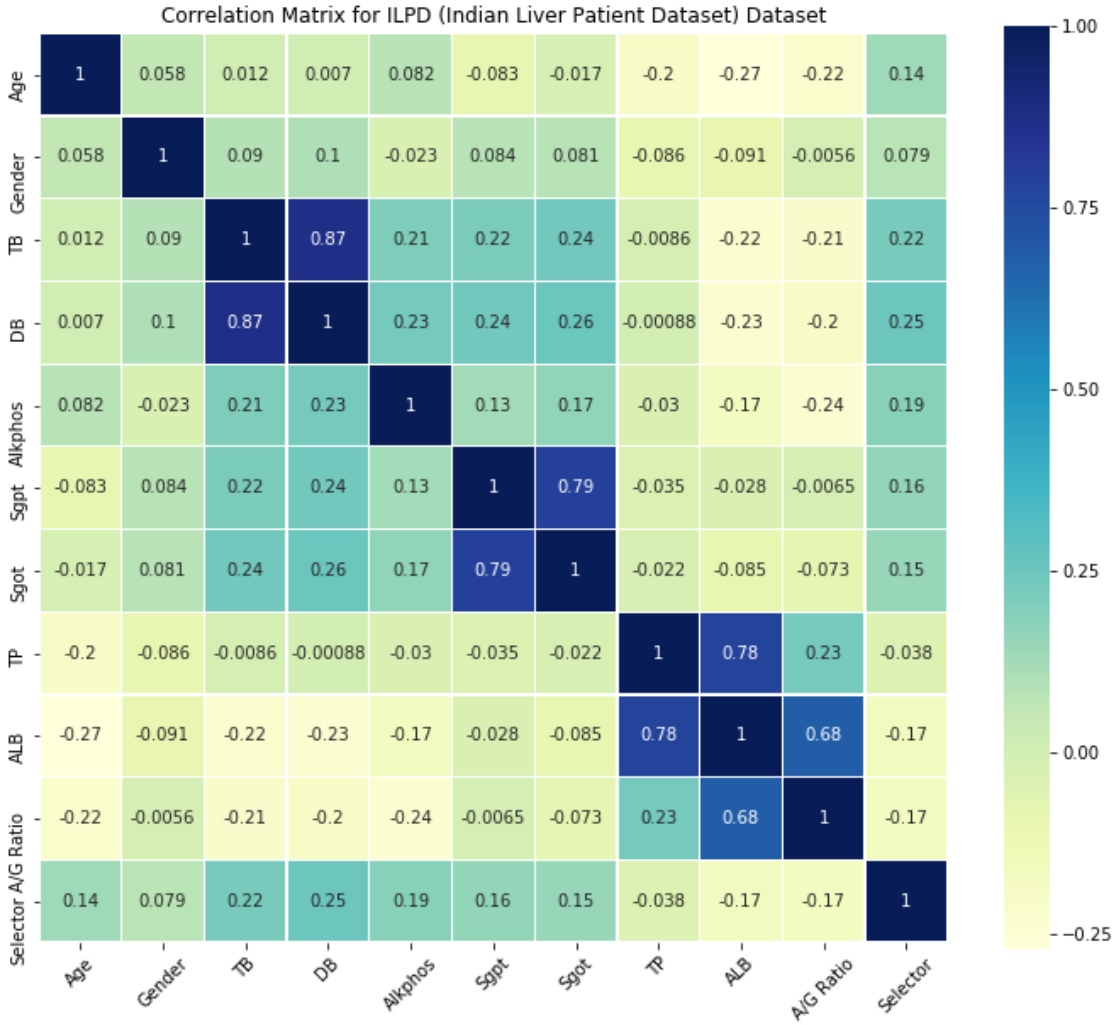
Optimised parameters were then applied to each classifier for a confusion matrix, accuracy, precision, recall and F1 scores. This was carried out on the training and testing datasets. A receiver operating characteristic (ROC) curve was produced using the training and testing AUC metrics. K-Fold cross validation was applied to the testing dataset for accuracy and AUC comparison purposes. Hyper-parameter sensitivity analysis was performed by varying the

parameter metrics one at a time for their respective accuracy, precision, recall and F1 scores. Further sensitivity analysis was carried out using principal components analysis (PCA) on both the models.

**Results**

Feature selection could be viewed in a number of ways, those features used in existing literature, those based on their statistical significance, by using the confusion matrix (Fig 1.) or from the feature importance plotting (Fig 2.). As the three of the top four features from the feature importance plotting also appear in literature sources and two have statistically significant p-values, the Alkphos, Sgot and Sgpt features were selected to visualise the feature space in Fig 3. Dimensionality reduction was carried out as a sensitivity analysis using PCA as per Fig 2.

*Fig 1. Confusion matrix showing the highest correlations between the dataset attributes.*






*****

*Fig 2. Plot of the feature importance showing the most (Alkphos) and least (Gender) important feature.*

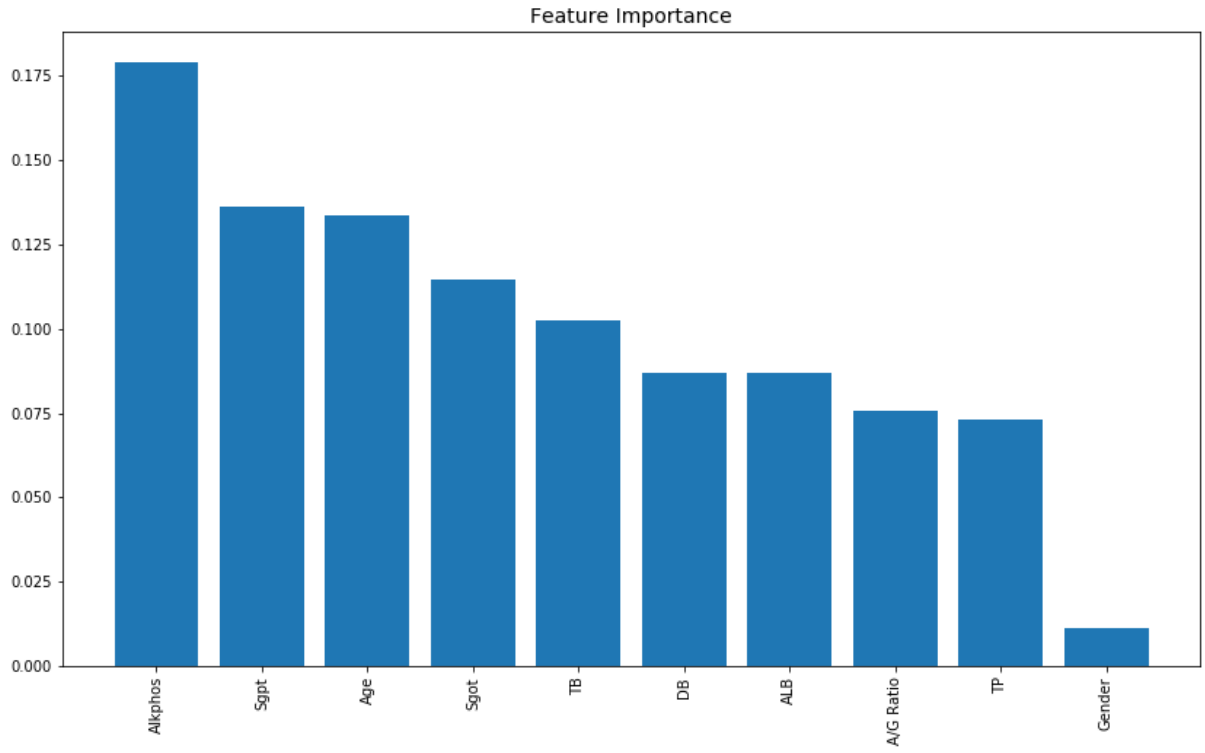
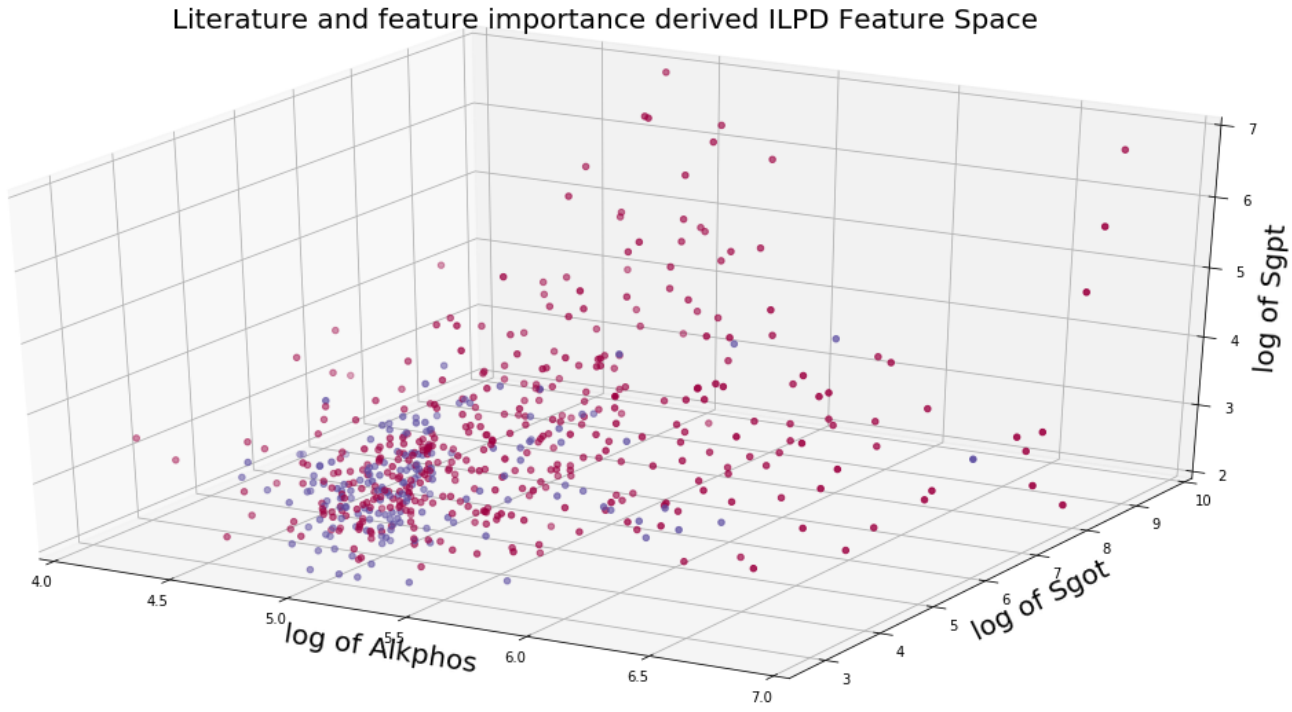



*Fig 3. Visualisation of the ILPD feature space of the liver patients (red) and the non-liver patients (blue). The features selected were taken from literature and as suggested by the feature importance plot in (Fig 3.).*






*****

In Fig 4a, the RFC ROC curves show an overly-fit optimised training AUC metric of 0.999, with a moderately fit optimised testing AUC metric of 0.72. There is an improvement of fit using PCA, with AUCs of 0.982 for training and 0.729 for testing.

*Fig 4a. Receiver Operating Characteristic (ROC) plots for the optimised random forest classification training and testing datasets.*

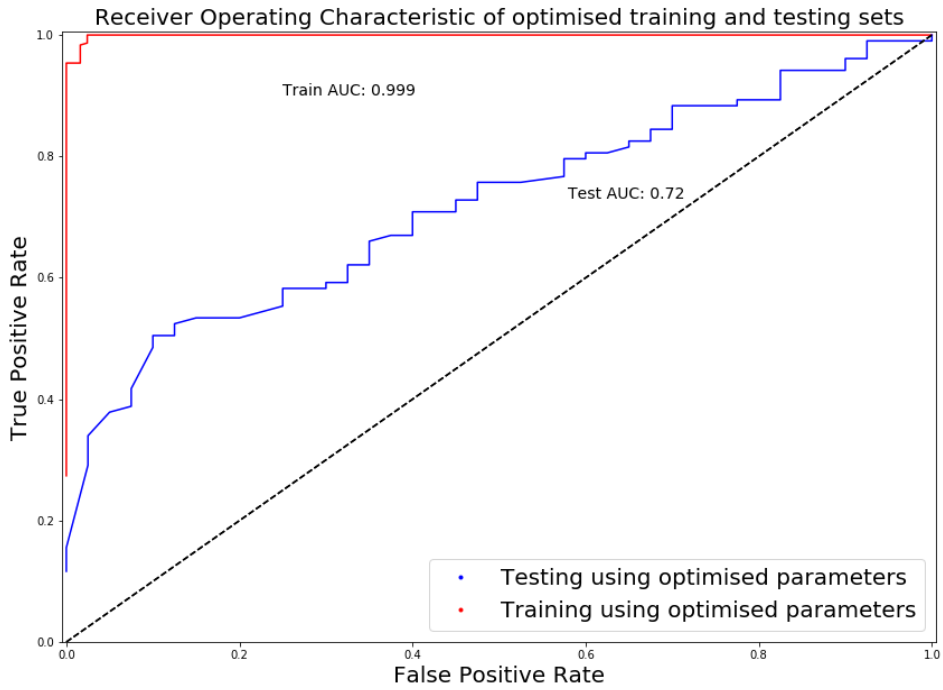
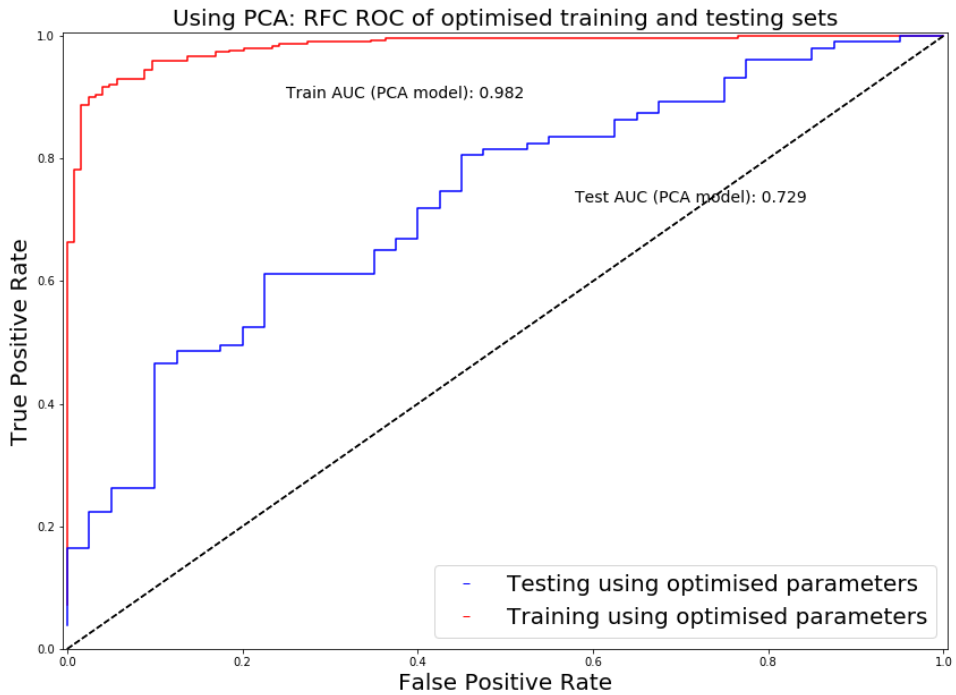



*Fig 4b. Receiver Operating Characteristic (ROC) plots for the optimised random forest classification training and testing datasets using PCA for feature selection.*






*****

For the SVC model, a highly fitted parameter optimised training set (0.986) can be observed in the SVC ROC curve in Fig 5, alongside a poorly fitting testing set (0.557).

*Fig 5a. Receiver Operating Characteristic (ROC) plots for the optimised support vector classification training and testing datasets.*

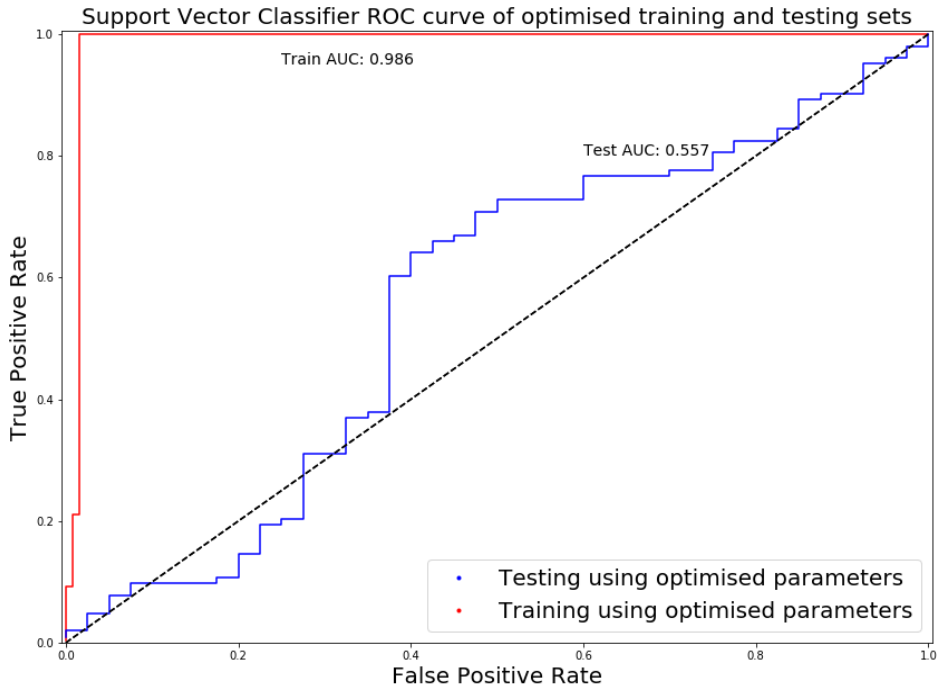
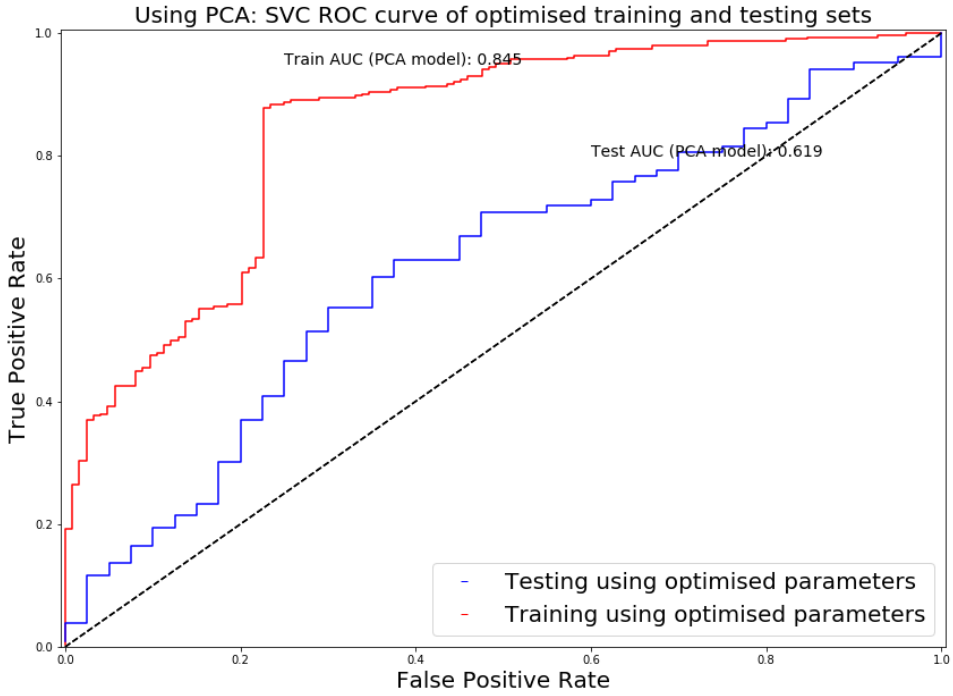



*Fig 5b. ROC plots of the optimised PCA support vector classification training and testing datasets.*






*****

SVC and the RFC models performed roughly the same for the training datasets in either the optimised or the non-optimised models. SVC performed better than the RFC model the non-optimised testing AUC model, however, crucially the optimised RFC testing AUC (0.72), see Table 1.

*Table 1. Shows the area under the curve (AUC) accuracy performance for both the random forest and support vector classifiers in relation to the training and testing datasets.*

| | **Random Forest Classifier** | | **Support Vector Classifier** | |
|---|---|---|---|---|
| | | **with PCA** | | **with PCA** |
| **Training AUC** | 0.736 | 0.753 | 0.774 | 0.773 |
| **Testing AUC** | **0.713** | 0.672 | **0.695** | **0.695** |
| **Optimised Training AUC** | 0.999 | 0.982 | 0.986 | 0.845 |
| **Optimised Testing AUC** | 0.720 | **0.729** | 0.557 | **0.619** |

In Table 2, SVC outperformed the RFC model prior to optimisation with a mean accuracy of 0 .713. While the optimised training set showed little difference between the RFC and the SVC models. However, once the optimised parameters had been applied to the testing set, the R FC model mean accuracy performance improved (0.685) over the initial testing (0.657) and w as better than the SVC optimised testing set (0.636). The K-Fold analysis showed the opposit e with the SVC mean accuracy at 0.685 (+/- 11.21%), as opposed to the RFC mean accuracy o f 0.621 (+/- 15.63%). The K-Fold standard deviation for both models show a significant varia nce suggesting over-fitting. For both models, the test accuracy for optimised PCA approach were best at 0.727 (RFC) and 0.699 (SVC). Over-fitting was again suggested in the KFold valid ation step with standard deviations for both models of +/- 7.20% and +/- 15.63%.

*Table 2. The table shows the results of initial testing dataset, training dataset, testing dataset and K-Fold cross validation performance metrics for both the random forest and support vector classifiers. PCA n_estimator details can be found in Fig 5.*

| | **Random Forrest Classifier** | | **Support Vector Classifier** | |
|---|---|---|---|---|
| | **Optimised** | **Optimised with PCA** | **Optimised** | **Optimised with PCA** |
| **Initial testing** | | | | |
| precision | 0.67 | - | 0.52 | - |
| recall | 0.66 | - | 0.71 | - |
| F1-score | 0.66 | - | 0.60 | - |
| Mean accuracy | 0.657 | - | **0.713** | **-** |
| **Training set** | | | | |
| precision | 0.99 | 0.93 | 0.97 | 0.78 |
| recall | 0.99 | 0.93 | 0.96 | 0.73 |
| F1-score | 0.99 | 0.93 | 0.96 | 0.63 |
| Mean accuracy | **0.991** | 0.925 | **0.965** | 0.728 |
| **Testing set** | | | | |
| precision | 0.66 | 0.66 | 0.61 | 0.51 |
| recall | 0.69 | 0.69 | 0.64 | 0.70 |
| F1-score | 0.67 | 0.67 | 0.62 | 0.59 |
| Mean accuracy | 0.685 | **0.727** | 0.636 | **0.699** |
| **K-Fold** | | | | |
| Mean accuracy | 0.621<br>Std:<br>+/- 15.63% | **0.643**<br>Std:<br>+/- 10.18% | 0.685<br>Std:<br>+/- 11.21% | **0.70**<br>Std:<br>+/- 7.20% |



*****

In Table 3, RFC hyper-parameter sensitivity analysis showed max depth had decreasing precision with increasing recall for the was expected as precision is trade off with recall. The best F1-score was where the precision and recall scores were the same. There was little variation in the F1-score, however with increasing depth from 2 to 100, the mean accuracy also increased from 0.622 to 0.713, respectively. Max feature demonstrated that the number of features have a bearing on the performance, with increasing mean accuracy of 0.643 to 0.678 for max features of 3 and 10 respectively. Both the min_sample_leaf and min_sample_split parameters had minimal variation from the optimised values.

*Table 3. Hyper-parameter sensitivity evaluation Random Forest Classifier using testing data sets.*

| **Hyper-parameter testing: Random Forest Classifier** | **precision** | **recall** | **F1-score** | **Mean Accuracy** |
|---|---|---|---|---|
| **Max depth** | | | | |
| 2 | **0.74** | 0.62 | 0.64 | 0.622 |
| 5 | 0.72 | 0.66 | 0.67 | 0.657 |
| 7 | 0.72 | 0.67 | 0.69 | 0.671 |
| 9 | 0.70 | 0.70 | **0.70** | 0.699 |
| 10 | 0.67 | 0.68 | 0.68 | 0.678 |
| 20 | 0.67 | 0.71 | 0.68 | 0.706 |
| 100 | 0.68 | **0.71** | 0.69 | **0.713** |
| **Max feature (With max depth =12)** | | | | |
| 3 | 0.64 | 0.64 | 0.64 | 0.643 |
| 10 | 0.67 | 0.68 | 0.68 | 0.678 |
| **Min sample leaf (With max depth =12)** | | | | |
| 5 | 0.67 | 0.61 | 0.63 | 0.608 |
| **Min sample split (With max depth =12)** | | | | |
| 4 | 0.69 | 0.69 | 0.69 | 0.692 |

In Table 4, SVC hyper-parameter sensitivity evaluation showed that with gamma=1, an increasing C value resulted in decreasing mean accuracy from 0.72 to 0.664. The highest F1-score was found where the precision and recall were near equal. With C=10 and gamma increasing from 0.0001 to 10, there was little difference for precision, recall, F1-score or mean accuracy. Changes to class weight with C=10 and gamma=1 showed no variation in the mean accuracy.

*Table 4. Hyper-parameter sensitivity evaluation SVC using testing data sets.*

| **Hyper-parameter testing: Support Vector Classifier** | **precision** | **recall** | **F1-score** | **Mean Accuracy** |
|---|---|---|---|---|
| **C (With Gamma=1)** | | | | |
| 0.1 | 0.52 | **0.72** | 0.60 | **0.720** |
| 1.0 | 0.56 | 0.69 | 0.60 | 0.692 |
| 50 | 0.66 | 0.67 | **0.67** | 0.671 |
| 100 | 0.65 | 0.66 | 0.66 | 0.664 |
| 300 | **0.65** | 0.66 | 0.66 | 0.664 |
| **Gamma (With C =10)** | | | | |
| 0.0001 | 0.52 | 0.72 | 0.60 | 0.720 |
| 0.001 | 0.52 | 0.72 | 0.60 | 0.720 |
| 0.01 | 0.52 | 0.72 | 0.60 | 0.720 |
| 0.1 | 0.51 | 0.70 | 0.59 | 0.699 |
| 10 | 0.63 | 0.71 | 0.62 | 0.713 |
| **Class weight (With C =10 and Gamma=1)** | | | | |
| 2 | 0.62 | 0.64 | 0.63 | 0.643 |
| 3 | 0.61 | 0.64 | 0.62 | 0.636 |
| 5 | 0.61 | 0.64 | 0.62 | 0.636 |
| 10 | 0.61 | 0.64 | 0.62 | 0.636 |

In Table 5, using PCA as a sensitivity analysis, PCA optimised accuracy scores were found to be 0.987 (training) and 0.751 (testing) using 7 features for RFC and 0.845 (training) and 0.619 (testing) using 6 features for SVC. The related 10 fold analysis gave 0.6848 std: +/- 0.112 for the best RFC set and 0.70 std: +/- 0.072 for the best SVC set.

Table 5. RFC and SVC sensitivity analysis of optimised training and testing sets using PCA

| **PCA (n_components)** | **Random Forest Classifier using PCA** | | **Support Vector Classifier using PCA** | |
|---|---|---|---|---|
| | Training | Testing | Training | Testing |
| **Optimised AUC** | | | | |
| **2** | 0.951 | 0.742 | 0.739 | 0.618 |
| **3** | 0.962 | 0.740 | 0.931 | 0.507 |
| **4** | 0.974 | 0.708 | 1.0 | 0.502 |
| **5** | 0.979 | 0.718 | 1.0 | 0.540 |
| **6** | 0.988 | 0.719 | **0.728** | **0.699** |
| **7** | **0.987** | **0.751** | 0.980 | 0.522 |
| **8** | 0.987 | 0.728 | 0.980 | 0.537 |
| **9** | 0.989 | 0.717 | 0.980 | 0.542 |
| **10** | 0.990 | 0.716 | 0.986 | 0.556 |
| **KFold for optimal n_estimates** | 0.6848 std: +/- 0.112 | | 0.70 std: +/- 0.072 | |

**Discussion and Conclusion**

The optimised ROC curves suggest that the training datasets were over-fit and not generalizable enough to accommodate the testing datasets. Assessing the models using too many parameter types, may be another potential cause of the over-fit. In order to limit overfitting, principal component analysis (PCA) modelling was implemented providing improved testing optimised RFC testing fit at 75.1%. Tpot or a boosting function could be used to better explore hyper-parameter modelling in order to achieve improved ROC curve fitting.

In the literature, accuracy of 71.5026% test sample fitting has been reported for RFC models, (Pahareeya, J, 2014). This was improved to 83.9677% with the application of 100% overfitting, (Pahareeya, J, 2014). In this analysis, the best RFC accuracy attained was 0.751 for the optimised PCA model.

Previously reported accuracy for support vector classification was 82.5342% (RBF kernel) and 81.6781% (linear kernel) (Kadu, G, 2018). In the current analysis, the RBF kernel was found to give the best accuracy of 0.699 to the optimised SVC model using PCA.

Time to completion calculations were used to indicate computational complexity of the models used. Model execution times for an SVC model of 3210ms have been reported for SVC, (Vijayarani, S, 2015), which is in contrast to 40.4ms for the optimised SVC testing and 28.3ms for the optimised PCA SVC testing carried in this work.

Implementation of the models would need to consider stake-holder consultation to explore the clinical utility, such as setting thresholds for clinical risk scoring. In addition, the benefits of predictive algorithms should outweigh any risk to patients. External validation would also be needed if it were to be used in the UK as the data is based on an Indian dataset.

The inherent computational complexity would need suitable hardware to accommodate speed of processing and use or complex steps, such as regularisation (limits over-fitting), Vollmer, S, 2019).

For model improvement, a metrics dashboard may be used for the ongoing monitoring of the algorithms to inform decisions on potential upgrades. If deployed, standardisation of data input would be needed to ensure consistency, such as consistent feature selection, data input and correct use of units of measure. Conversion of the models to a suitable user interface for ease of daily implementation may also be considered.

The best accuracies were achieved using PCA (0.751 (RFC) and 0.699 (SVM)), demonstrating the usefulness of the approach in the optimisation of a predictive machine learning algorithm.

**References**


- Ramana, B V, Prasad Babu, M S and Venkateswarlu N B, 'A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis', International Journal of Computer Science Issues, ISSN :1694-0784, May 2012.
- Ramana, B V, Prasad Babu, M S and Venkateswarlu, N B, 'A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis', International Journal of Database Management Systems (IJDMS), Vol.3, No.2, ISSN: 0975-5705, PP 101-114, May 2011.
- Harrison, M F. 'The Misunderstood Coagulopathy of Liver Disease: A Review for the Acute Setting', West J Emerg Med. 2018 Sep; 19(5): 863–871. doi: 10.5811/westjem.2018.7.37893
- NHS online, 2019. 'Liver disease', https://www.nhs.uk/conditions/liver-disease/ . Last viewed on 10.05.19
- Blachier, M, Leleu, H, Peck-Radosavljevic, M, Valla, D-C, Roudot-Thoraval, F. 'The burden of liver disease in Europe: A review of available epidemiological data', Journal of Hepatology, Vol.58, Issue 3, PP. 593-608, March 2013.
- Pimpin, L, Cortez-Pinto H, Negro F, Corbould E, Lazarus J V, Webber, L, Sheron, N, and the members of the EASL HEPAHEALTH Steering Committee, 'Burden of liver disease in Europe: Epidemiology and analysis of risk factors to identify prevention policies', Vol.69, Issue 3, PP. 718-735, September 2018.
- British Liver Trust A, 2019, 'Tests and Screening', https://www.britishlivertrust.org.uk/liver-information/tests-and-screening/ . Last viewed on 10.05.19
- Procopet, B, Berzigotti, A. 'Diagnosis of cirrhosis and portal hypertension: imaging, non-invasive markers of fibrosis and liver biopsy', *Gastroenterology Report,* Vol. 5, Issue 2, PP. 79-89, May 2017. Doi: https://doi.org/10.1093/gastro/gox012
- British Liver Trust B, 2019, 'Liver blood tests (formerly known as liver function tests, or LFTs)', https://www.britishlivertrust.org.uk/liver-information/tests-and-screening/liver-blood-tests-formerly-lfts/. Last viewed on 10.05.19.
- Thriva health hub, 2019. 'How to test your liver function', https://thriva.co/hub/liver-function-test . Last viewed on 10.05.19.
- Global Health Observatory data repository, 2019. 'Liver cirrhosis (15+), age-standardized death rates by country', http://apps.who.int/gho/data/node.main.A1092 . Last viewed on 10.05.19.
- Vollmer, S, Mateen, B A, Bohner, G, Király, F J, Ghani, R, Jonsson, P, Cumbers, S, Adrian Jonas, A, Katherine S.L. McAllister K S L, Myles, P, Granger, D, Birse, M, Branson, R, Moons, K G M, Collins, G S, Ioannidis, J P A, Holmes, C, Hemingway, H. 'Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness', https://arxiv.org/pdf/1812.10404.pdf . Last viewed on 11.05.19.
- Pahareeya, J, Vohra, R, Makhijani, J, Patsariya, S, 'Liver Patient Classification using Intelligence Techniques'. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 2, ISSN: 2277 128X. February 2014.
- Kadu, G, Raut, R, Gawande, S S, 'Diagnosis of liver abnormalities using Support Vector Machine', International Journal for Research Trends and Innovation, Vol. 3, Issue 7, ISSN: 2456-3315, 2018.




*****

- Vijayarani, S, Dhayanand, S. 'Liver Disease Prediction using SVM and Naïve Bayes Algorithms', International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 4, April 2015.