# Data Mining and Decision Systems 600092 Assigned Coursework Report

Student ID: 201601628

Date: 07 October 2019

Due Date: 12 December 2019

**Report must be within 8 page maximum. Strict page limits will be enforced. Any extra pages will be ignored and no marks awarded for any work on these. Exclusions to this limit are the front page, the references section, and any appendices. Please keep to the given section headings and format; subsections are permitted.**

# Methodology

**Overview**

A context-adapted CRISP-DM methodology was followed, with reduced focus on the 'business understanding' and ' deployment' phases in the absence of further client interaction. This document outlines the processes followed whilst adhering to this framework, the results achieved, and a discussion of findings.

1. **Business Understanding**

This phase sought to understand and summarise the problem ahead by considering the domain (cardiovascular healthcare), personal domain knowledge, and additional research.

Particular attention was given to medical terminology, leading to the following assumptions and findings:

| **Assumption** | **Finding** |
|---|---|
| Patients with an indication of "a-f" (arterial fibrillation) should also be recorded as having an arrhythmia, since a-f is a type of arrhythmia (Arrhythmia, 2019). | The data did not conform to this at all.<br>It was also found that there are many arrhythmias with varying severities, so the assumption changed to view the 'Arrythmia' feature as, "one more severe than a-f" (Categories of Arrhythmias \| Texas Heart Institute, 2019). |
| Ipsi and Contra relate to lesions on the same side (as something) and opposite side (of something). Considering the context, if the 'something' is a stroke, it might make sense for Ipsi and Contra to have a sum of 100%. | Less than 1/5 of the data conformed to this (*see ln-63-64 in code).* |

In terms of performance metrics, sensitivity and specificity carry particular weight in the healthcare domain, since resources are too limited to waste on unnecessary cases and misclassifications may result in loss of life. In addition, the F1 Score can provide more meaningful insight than accuracy into overall model performance. Since it is weighted and is based off the sensitivity and specificity, it better handles imbalanced data (Huilgol, 2019).

2. **Data Understanding**

<u>Conformity to Data Dictionary</u>

The data dictionary was viewed with scepticism and treated as an *idealistic* overview. To compare the data, a python dictionary was created to describe each of features as given by the data dictionary (*see ln-12 in code)*.

This code looked for ways in which the *actual* data deterred from expectations and outputted the findings (*see ln 13:14 in code).*

The key findings were:

- The session column which was mentioned, separate to the dictionary, was not present. Because the task is not using time-series data or predicting progression

rates, and it is natural for a person's symptoms and risk factor to change visit to visit, this was not too much of a concern.
- Contra values were string representations of numbers. These were converted to numbers (*see ln-19 in code)*.
- Indication had differently formatted categories (e.g. “Asx” and “ASx") for the same class. Further domain research suggested there was no difference, so they were merged (*see ln-18 in code)* (Brothers et al., 2015).
- The Random feature was supposed to be unique, but there were 298 duplicate values (*see ln-15:17 in code)*.
- The Label feature had two records classified as “Unknown”. These were imputed in one data frame and dropped in another (*see ln-20:22 in code)*.

An interest was taken in the possibility that the missing session feature could be encoded *within* random or id, or it that id and random were mislabelled. To explore this idea, records with duplicated ‘Random’ values were inspected for multiple changes to history and diabetes. If these values changed multiple times, it would indicate that the Random feature isn’t the patient id, since once a patient has history or diabetes it shouldn’t revert. None of these contradictions appeared, however, supporting the possibility that ‘Random’ could be the patient id and ‘Id’ might be the session id (*see ln-15:17 in code)*.

Checking for Duplicates

Checks for duplicates included code that looked for records with **all but one** attribute the same, and **all but two** attributes the same (*see ln-23:24 in code*). The most meaningful finding from these checks was the apparent homogeneity of the data set; when the ‘id’ and ‘random’ columns were excluded from the checks, two thirds of all records were ‘duplicated’.

Aside from the this, a single record was found when ‘id’ and ‘contra’ were ignored, and they appeared only 2 records apart by index (*see ln-25 in code*). Assuming human input, it’s possible that values in this column were mistakenly entered. However, carrying the consideration surrounding ‘Random’ and ‘Id’, it appears that this could be 2 visits from the same patient. As a single datapoint, a note was made of its existence but the effect of removing it was expected to be negligible.

Checking for Missing Data

Checks for missing data were made increasingly thorough after falling afoul to ‘hidden’ missing values that were empty string literals. These were originally found after visualising the indication column, where “nan” appeared”. To remedy this, the normal checks (*isna()/isnull()*) were used alongside regular expressions that also looked for values that were blank strings or some case-insensitive version of ‘nan’ - and the analyses were re-run (*see ln-9 in code)*.

18 records with missing data were found (see *ln-26 in code*). Recalling the homogeneity that the dataset displayed in the search for duplicates, a method for imputing values based on the nearest neighbours was created, since most records had hundreds of instances when random and id were ignored (*see ln-4 & ln-27 in code)*.

Distribution and Outliers

These checks focused on visualisations and statistics of each feature to find class imbalances and to understand the distribution of the target class (*see ln-31:62 in code)*.

They revealed the existence of a risk-only cluster that appeared when distributing any feature against Id (*see ln-34:35 in code & Figure-1 in Results)*. Several causes are plausible: since the datapoints are so sparse in that region, it appears as though some 'NoRisk' cases could be omitted. Alternatively, 'Id' may indeed be the patient identifier, and that cluster could have been part of a study group for a clinical trial, considering the domain.

Regardless of *how,* the effect of this cluster was evaluated by removing it but there were no significant changes to feature correlations or data distribution (*see ln-36:38)*.

In addition, several instances of massively imbalanced features were found; history and diabetes, namely (see ln-52 in code & Figure-2 in Results). It was decided that these features were likely better-off removed from the training data due to the biases they would introduce, potentially causing the model to faulter with future data.

## 3. Data Preparation

This data utilised the acquired understanding to generate several datasets, cleaned and transformed according to various hypotheses.

Data Frames (Data Variations)

In order to be able to apply transformations to all variations of the dataset without error, data frames were collated in dictionary so that they could be iterated over (*see ln-65:72 in code).*

Transformation

*(see ln-73:75 in code)*

Most of the data was converted to a binary representation using '*pd.DataFrame.replace()* ', since most of the feature categories were binary (e.g. 'yes', 'no'). Alternatively, these features could have been one-hot encoded with *'get_dummies()'*, but this would have the effect of increasing the feature space unnecessarily.

Regarding normalisation, Ipsi and Contra were the only numeric features of interest (ignoring random and id) and were normalised by a simple division of 100 since they are *supposed* to be percentages (ranging 0-100). This has the added benefit of making the model robust to test data since, for example, min-max normalisation would assume the absolute min and max to be that of the training sample.

Based on the exploration of the indication feature *(see ln-41 in code)*, two methods of encoding were used. One binarized the feature based on evidence that patients with 'CVA' or 'TIA' indications may have a slightly reduced risk relative to 'ASX' and 'A-F'. A possible reason for this pattern may be that the latter are precursors to the former. For example, 'ASX' may increase the chance of 'TIA', but after having that mini-stroke ('TIA') surgery or lifestyle changes may occur that reduce risk. The other method used the more conventional, '*get_dummies()*' to one-hot encode the 4 values.

Feature Selection
Feature selection comprised of manual selection (based on the data understanding and correlation heatmaps) and a random forest classifier that weighted each feature to support or dispute the manual selection *(see Figure-3 in Results & ln-78 in code)*.

Based on correlations and imbalances, the manual selection proposed the removal of id, random, history, and diabetes – with IHD and indication regarded as potentials too. The classifier weights matched the manual selection closely, with the only difference being the value of the diabetes feature. It is assumed that this is down to its overly good class split *(see ln-42 in code)*. It was still disregarded because of the bias it would likely introduce in future test data.

Sampling (TTS)
Data-driven stratification was used to create a conventional with a 70:30 split that balanced distribution of classes in both the train and test sets. The random seed was modified until the visualisations presented a similar distribution in both the train and test sets (*see ln-82 in code)*.

Additionally, stratified k-fold was used to create 5 balanced splits in a similar way, to validate model performances across the whole distribution of the data (*see ln-83 in code)*. Five was chosen as the number of folds so that the training data was still substantial (~300 records).

An example of contra train-test splits for both methods is included in Figure-5 of the Results*.*

## 4. Modelling

The modelling phase was approached in a stepwise manner to attempt to select the optimal training data and model. Models were contained similarly to datasets (in a dictionary) for use in a wrapper method that automated testing and reporting for any number of models and datasets (*see ln-11).*

1) Baseline Model (Data Selection)
   The baseline model a simple logistic regression classifier that was tested across all data variations to try and find the optimal data configuration to continue with (*85:89).*

2) Additional Models with Selected Data (Model Selection)
   A variety of models were then created with their default hyperparameters (except for a modest max-depth in random forest and decision trees) (*90:92).*

3) Hyperparameter Tuning (Selected Model/s)
   The first hyperparameter explored was the max-depth, as it is common for both selected models. A simple loop created classifiers for depth +-5 from the baseline (*93:95).*

A table of all model metrics and final-model visualisations are presented in the Results section.

# Results

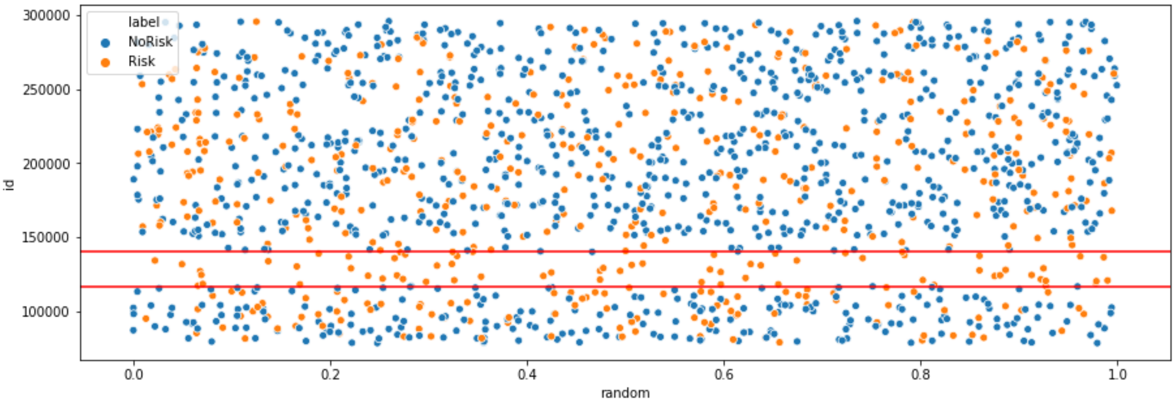
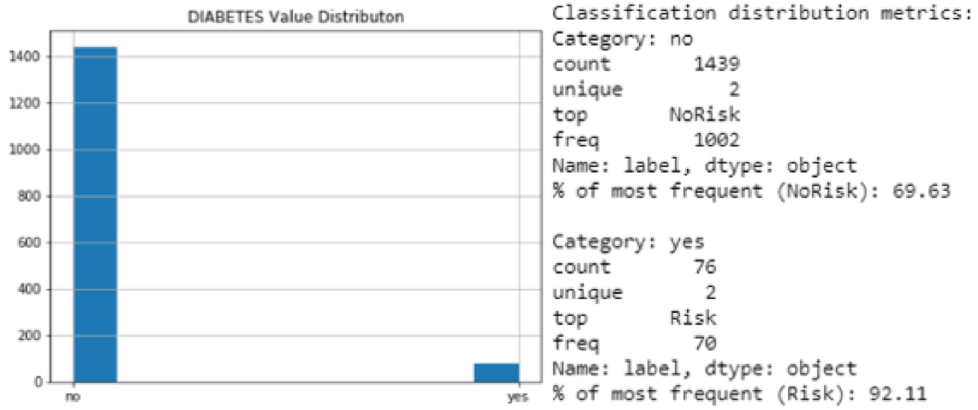


*Figure 1: Id cluster.*



```
Classification distribution metrics:
Category: no
count        1439
unique          2
top        NoRisk
freq         1002
Name: label, dtype: object
% of most frequent (NoRisk): 69.63

Category: yes
count          76
unique          2
top          Risk
freq           70
Name: label, dtype: object
% of most frequent (Risk): 92.11
```

*Figure 2: Diabetes feature imbalance.*

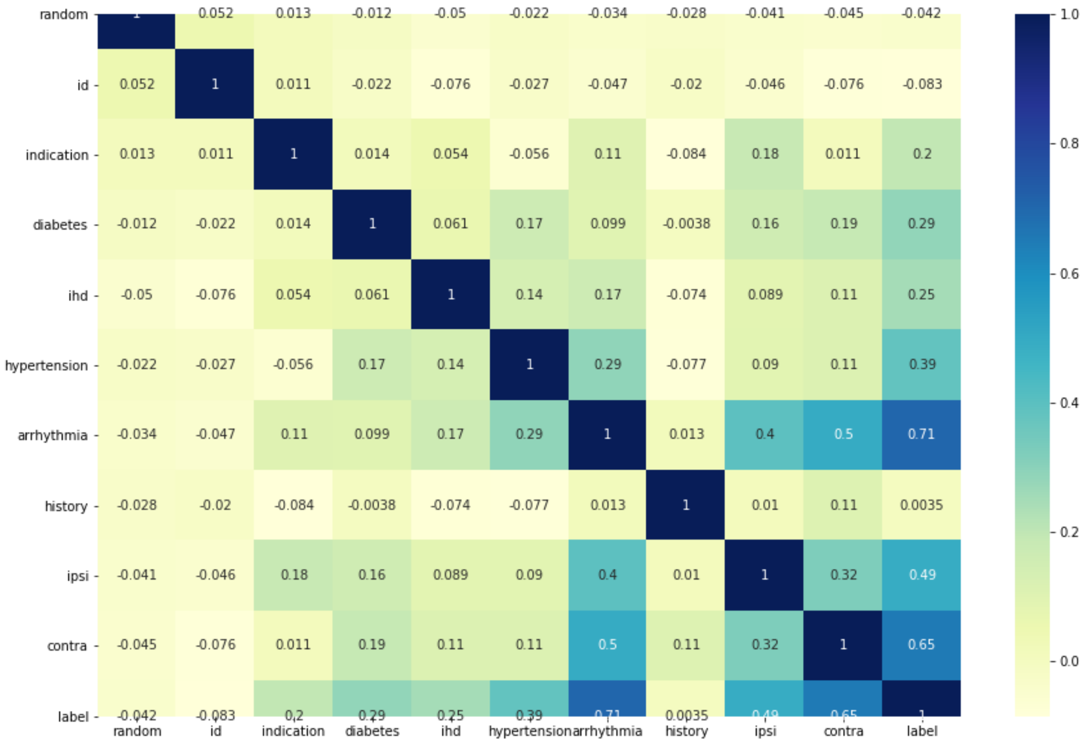
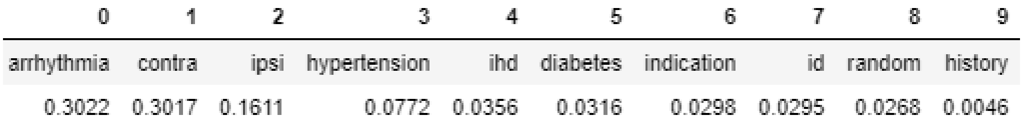


*Figure 3: Feature correlations heatmap.*



| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| arrhythmia | contra | ipsi | hypertension | ihd | diabetes | indication | id | random | history |
| 0.3022 | 0.3017 | 0.1611 | 0.0772 | 0.0356 | 0.0316 | 0.0298 | 0.0295 | 0.0268 | 0.0046 |

*Figure 4: Ordered, random forest feature weights.*

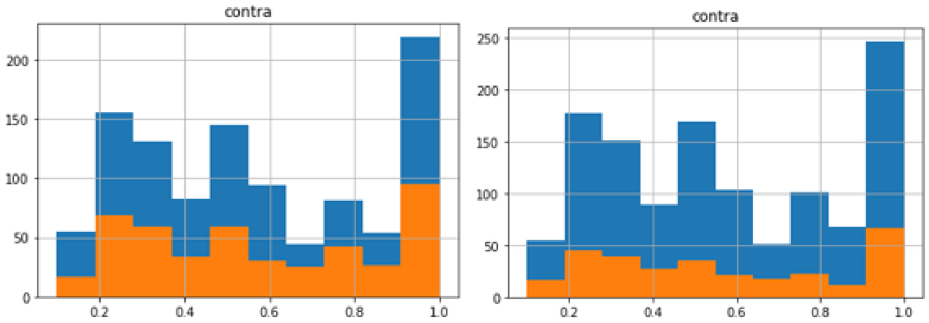
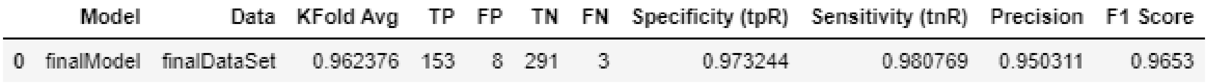


*Figure 5: Example of stratified train test sampling (conventional 70:30 split on left and k-fold on right).*



| | Model | Data | KFold Avg | TP | FP | TN | FN | Specificity (tpR) | Sensitivity (tnR) | Precision | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | finalModel | finalDataSet | 0.962376 | 153 | 8 | 291 | 3 | 0.973244 | 0.980769 | 0.950311 | 0.9653 |

*Figure 6: Final model and dataset metrics.*

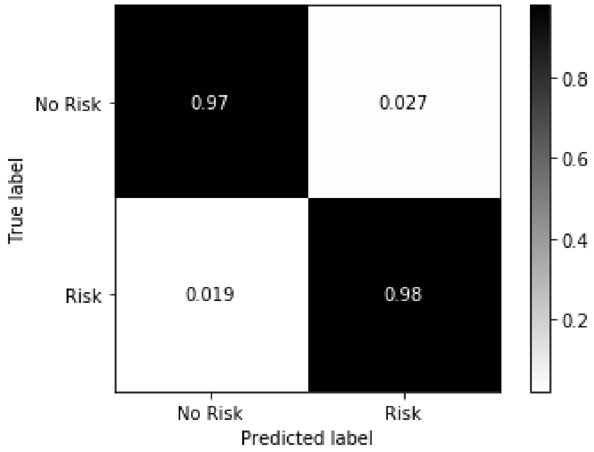


*Figure 7: Confusion matrix for final model.*

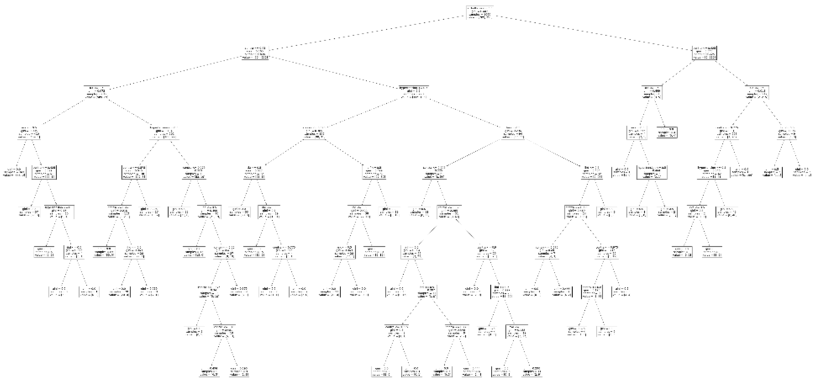
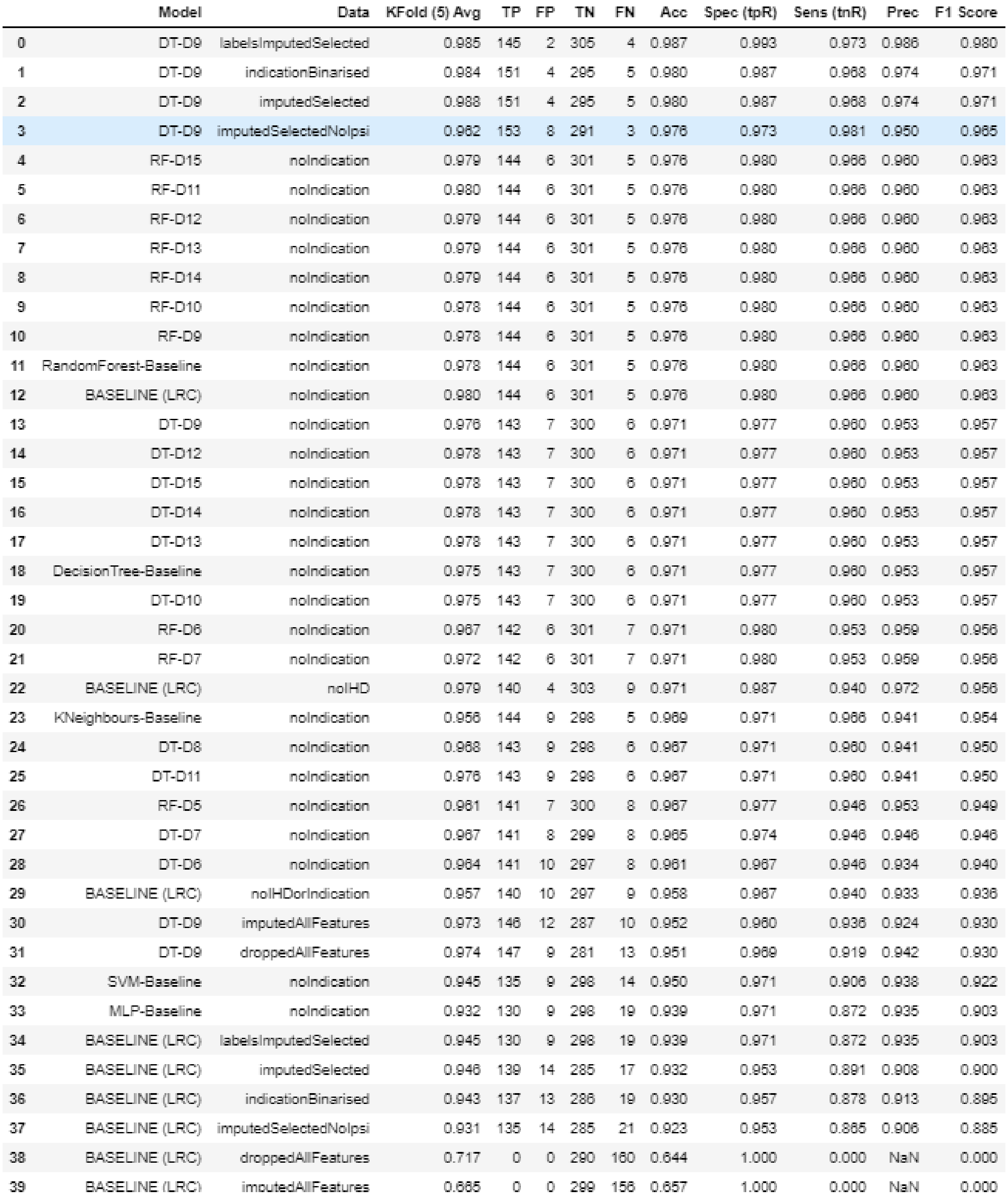


*Figure 8: Final model, plotted (decision tree) – link to file: https://ibb.co/d2FSQ90.*



| | Model | Data | KFold (5) Avg | TP | FP | TN | FN | Acc | Spec (tpR) | Sens (tnR) | Prec | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | DT-D9 | labelsImputedSelected | 0.985 | 145 | 2 | 305 | 4 | 0.987 | 0.993 | 0.973 | 0.986 | 0.980 |
| 1 | DT-D9 | indicationBinarised | 0.984 | 151 | 4 | 295 | 5 | 0.980 | 0.987 | 0.968 | 0.974 | 0.971 |
| 2 | DT-D9 | imputedSelected | 0.988 | 151 | 4 | 295 | 5 | 0.980 | 0.987 | 0.968 | 0.974 | 0.971 |
| 3 | DT-D9 | imputedSelectedNoIpsi | 0.962 | 153 | 8 | 291 | 3 | 0.976 | 0.973 | 0.981 | 0.950 | 0.965 |
| 4 | RF-D15 | noIndication | 0.979 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 5 | RF-D11 | noIndication | 0.980 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 6 | RF-D12 | noIndication | 0.979 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 7 | RF-D13 | noIndication | 0.979 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 8 | RF-D14 | noIndication | 0.979 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 9 | RF-D10 | noIndication | 0.978 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 10 | RF-D9 | noIndication | 0.978 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 11 | RandomForest-Baseline | noIndication | 0.978 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 12 | BASELINE (LRC) | noIndication | 0.980 | 144 | 6 | 301 | 5 | 0.976 | 0.980 | 0.966 | 0.960 | 0.963 |
| 13 | DT-D9 | noIndication | 0.976 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 14 | DT-D12 | noIndication | 0.978 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 15 | DT-D15 | noIndication | 0.978 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 16 | DT-D14 | noIndication | 0.978 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 17 | DT-D13 | noIndication | 0.978 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 18 | DecisionTree-Baseline | noIndication | 0.975 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 19 | DT-D10 | noIndication | 0.975 | 143 | 7 | 300 | 6 | 0.971 | 0.977 | 0.960 | 0.953 | 0.957 |
| 20 | RF-D6 | noIndication | 0.967 | 142 | 6 | 301 | 7 | 0.971 | 0.980 | 0.953 | 0.959 | 0.956 |
| 21 | RF-D7 | noIndication | 0.972 | 142 | 6 | 301 | 7 | 0.971 | 0.980 | 0.953 | 0.959 | 0.956 |
| 22 | BASELINE (LRC) | noIHD | 0.979 | 140 | 4 | 303 | 9 | 0.971 | 0.987 | 0.940 | 0.972 | 0.956 |
| 23 | KNeighbours-Baseline | noIndication | 0.956 | 144 | 9 | 298 | 5 | 0.969 | 0.971 | 0.966 | 0.941 | 0.954 |
| 24 | DT-D8 | noIndication | 0.968 | 143 | 9 | 298 | 6 | 0.967 | 0.971 | 0.960 | 0.941 | 0.950 |
| 25 | DT-D11 | noIndication | 0.976 | 143 | 9 | 298 | 6 | 0.967 | 0.971 | 0.960 | 0.941 | 0.950 |
| 26 | RF-D5 | noIndication | 0.961 | 141 | 7 | 300 | 8 | 0.967 | 0.977 | 0.946 | 0.953 | 0.949 |
| 27 | DT-D7 | noIndication | 0.967 | 141 | 8 | 299 | 8 | 0.965 | 0.974 | 0.946 | 0.946 | 0.946 |
| 28 | DT-D6 | noIndication | 0.964 | 141 | 10 | 297 | 8 | 0.961 | 0.967 | 0.946 | 0.934 | 0.940 |
| 29 | BASELINE (LRC) | noIHDorIndication | 0.957 | 140 | 10 | 297 | 9 | 0.958 | 0.967 | 0.940 | 0.933 | 0.936 |
| 30 | DT-D9 | imputedAllFeatures | 0.973 | 146 | 12 | 287 | 10 | 0.952 | 0.960 | 0.936 | 0.924 | 0.930 |
| 31 | DT-D9 | droppedAllFeatures | 0.974 | 147 | 9 | 281 | 13 | 0.951 | 0.969 | 0.919 | 0.942 | 0.930 |
| 32 | SVM-Baseline | noIndication | 0.945 | 135 | 9 | 298 | 14 | 0.950 | 0.971 | 0.906 | 0.938 | 0.922 |
| 33 | MLP-Baseline | noIndication | 0.932 | 130 | 9 | 298 | 19 | 0.939 | 0.971 | 0.872 | 0.935 | 0.903 |
| 34 | BASELINE (LRC) | labelsImputedSelected | 0.945 | 130 | 9 | 298 | 19 | 0.939 | 0.971 | 0.872 | 0.935 | 0.903 |
| 35 | BASELINE (LRC) | imputedSelected | 0.946 | 139 | 14 | 285 | 17 | 0.932 | 0.953 | 0.891 | 0.908 | 0.900 |
| 36 | BASELINE (LRC) | indicationBinarised | 0.943 | 137 | 13 | 286 | 19 | 0.930 | 0.957 | 0.878 | 0.913 | 0.895 |
| 37 | BASELINE (LRC) | imputedSelectedNoIpsi | 0.931 | 135 | 14 | 285 | 21 | 0.923 | 0.953 | 0.865 | 0.906 | 0.885 |
| 38 | BASELINE (LRC) | droppedAllFeatures | 0.717 | 0 | 0 | 290 | 160 | 0.644 | 1.000 | 0.000 | NaN | 0.000 |
| 39 | BASELINE (LRC) | imputedAllFeatures | 0.665 | 0 | 0 | 299 | 156 | 0.657 | 1.000 | 0.000 | NaN | 0.000 |

Figure 9: All models and their metrics (chosen model is highlighted).

| Data Name | Description |
|---|---|
| droppedData | Missing values removed, with select features. |
| labelsImputedSelected | Missing values and labels imputed with select features. |
| imputedSelected | Missing values imputed with select features. |
| indicationBinarised | Missing values imputed with select features, with indication binarized. |
| ImputedSelectedNoIpsi | Missing values imputed with select features, with Ipsi additionally removed. |
| imputedAllFeatures | Missing values imputed with ALL features. |
| droppedAllFeatures | Missing values removed with ALL features. |
| noIndication | Missing values and labels imputed with select features-indication removed. |
| noIHD | Missing values and labels imputed with select features-IHD removed. |
| noIHDorIndication | Missing values and labels imputed with select features-IHD & indication removed. |

Table 1: Data variations.

# Evaluation & Discussion

## 5. Evaluation

Baseline Model (Data Selection)

The baseline model performed shockingly well, which is likely because of the homogeneity of the data (and therefore simplicity of the problem)– with the exception being the datasets including all features (*see ln-87 in code & Figure-9 in Results)*.

The labelsImputedSelected dataset was deemed optimal despite it coming second place to droppedData in F1-score, because it included additional datapoints. Referring back to the data understanding: removing 'indication' reduced the number of false negatives, whilst removing 'ihd' reduced the number of false positives. Considering the context of the problem, reducing the number of false negatives was prioritised and the dataset carried forward was the "noIndication" variation (*see Table 1 in results)*.

Additional Models with Selected Data (Model Selection)

All models performed desirably, again pointing to an overly simplistic problem (*see ln-92 in code)*. Whilst the KNeighbours classifier performed almost as well as the decision tree and random forest, the latter performed better in cross validation suggesting they are more robust. Furthermore, the nature of decision trees makes them transparent and explainable, which is hugely beneficial in this domain.

Hyperparameter Tuning (Selected Model/s)

The baseline depths seemed almost optimal in both models and changes to the depth (+-5) provided little or no improvement.

Comparing the best random forest and best decision tree, it can be seen there is very little difference in any metrics. For that reason, the decision tree with depth 9 (DT-D9) was selected for being simpler (cheaper/more efficient) model - with the added benefit that it is more transparent and explainable as an individual tree.

Finally, the chosen model was run against all data again, to ensure the optimal data configuration was chosen. Variation in F1 scores was quite small, (0.979-0.956) and suggesting that changes to the data were mostly insignificant beyond removing features.

With such slight differences in the performance metrics and **very** close quantities of false positives and negatives, the opportunity was taken to select the model trained without Ipsi (recalling its imbalance and outliers) since future if the representation is as poor as has been assumed it will affect the robustness of the final model on future test data.

## 6. Deployment

The proposed model for deployment is the decision tree identified in Figures 6,7, and 8: model, "DT-D9", trained on data, "imputedSelectedNoIpsi".

With an almost equal sensitivity and specificity score, as well as having one of the smaller feature sets (ihd, hypertension, arrhythmia, contra, indication) the model is one of the most portable and robust to overfitting - as demonstrated by the similar scores across accuracy, f1 and cross validation. With a relatively small max depth, the model should be easily explainable and more readily approved by the FDA if the decision process is validated by domain experts.

# References

Arrhythmia. (2019) nhs.uk. Available online: https://www.nhs.uk/conditions/arrhythmia/ [Accessed 20/12/2019].

Brothers, T., Ricotta, J., Gillespie, D., Geraghty, P., Kenwood, C., Siami, F., Ricotta, J. & White, R. (2015) Contemporary results of carotid endarterectomy in "normal-risk" patients from the Society for Vascular Surgery Vascular Registry. *Journal of Vascular Surgery*, 62(4), 923-928.

Categories of Arrhythmias | Texas Heart Institute. (2019) Texas Heart Institute. Available online: https://www.texasheart.org/heart-health/heart-information-center/topics/categories-of-arrhythmias/ [Accessed 20/12/2019].

Huilgol, P. (2019) *Accuracy vs. F1-Score*. Medium. Available online: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2 [Accessed 20/12/2019].

Upload Image — Free Image Hosting. (2019) ImgBB. Available online: https://imgbb.com/ [Accessed 20/12/2019].