Data Mining and Decision Systems

600092

Assigned Coursework Report

Student ID: 201760879

SHAFI SHAFI

Contents

1 Methodology

1.1 Introduction

This project will be tackling a classification problem to produce optimal results for a patient's risk. To help deal with this project, the CRISP-DM methodology was adopted and followed. The CRISP-DM methodology consist of six phases 'Business Understanding, Data understanding, Data Preparation, Data Modelling, Evaluation, Deployment'. Figure 1 shows a sample of data provided in the Appendix. And Figure 2 shows – The given 'Data Description for the Legacy Data Provided'.

1.2 Data Understanding

First step was to understand what was needed to be accomplished from the business perspective, which in our case was to 'filter, clean, and transform the medical data provided appropriately, such that it can be used to produce optimal classification for patient risk'. After we assessed the problem a plan was put in place which was clean then filter and transform the data appropriately so it could be used in models. Before doing that the data needed to be understood. The data's strength and limitation were needed to be grasped, for example, whether all the data provided was needed, if there are any null values should they be dropped or predicted. After looking at type of data provided, it was established that there were to categories for the data, numerical and categorical data as you can see in figure 2. For this project the library's used were pandas, NumPy, and sklearn.

1.3 Data Preparations

The pandas library was used to read in the provided data. From analysing the data earlier, it was established that the column Random represents 'Real number of help in randomly sorting the data records' and the column 'Id' represents 'Anonymous patient record identifier: Should be unique values unless patient has multiple sessions'. It was established that the data in these two columns do not help us with the classification problem so both the columns were dropped. Checking the datatypes for the columns and checking if they are unique we spot that they contain null values, a value called 'Unknown' and white space ' '. A spelling error was spotted in the indication column ('ASx','Asx'). The White space was spotted in the 'Contra' column. As you can see in figure 2 indication should only have four not five values. The value called 'Unknown' and the white space were dealt with, by creating a

variable called missing values at the top of the note book and adding potential null values which could be in the data set but not be identified as null values, such as a 'Unknown' and ' ' white space. The variable named missing values is then inputted in as the new definition of null values when being read in. To remove the null values for the data, initially it was done by filling the null values with the median values of the columns. However, because the data set is reasonably sized, the decision was made to remove all rows containing any null values. Another reason why they were dropped is because this is a medical data set, predicting a value could cause false classification and could be fatal e.g. a patient not being at risk when they actually are (False negative) could result in earlier death in the real world. For the misspelt value in the indication column the value 'Asx' was changed to 'ASx'. This was done because there was significantly more 'ASx' than 'Asx' and it could clearly be seen this is a spelling error, so the assumption was made not to consider it a null value.

1.4 Numerical Attributes

Drawing a histogram for 'Contra' and 'IPSI' it can be seen that both numerical attributes have been pre-stratified (shown in figure 3). Figure 4 shows the type of correlation both the numerical attributes have, it can be recognized that there is little to no correlation between the two attributes. After analysing figure 3, box plot diagrams were added in to compare the two attributes as figure 3 shows there may be some outliers which aren't as visible in histograms and would be better seen in other diagrams, such as the boxplot in figure 5 in the appendix , it was spotted that from the box plot that 'IPSI' contains outliers whereas contra did not. The decision was made to remove these outliers less than or equal to 65. As the 5th percentiles value was 65. If the outliers were kept at their current value it would reduce the chosen model's prediction accuracy, and the overall classification accuracy. This does reduce our overall data set from 1520 to 1431 after removing outliers and null value. Nevertheless, it improves the quality of the model predictions, And decreases the chance of making a 'False negative' prediction.

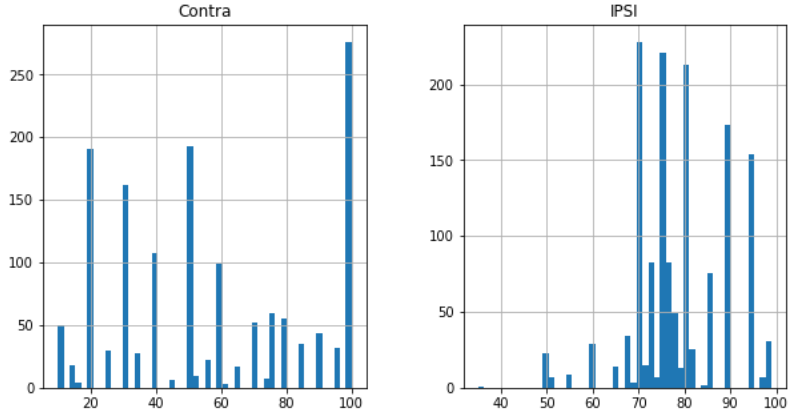


Figure 3: Histogram of numerical attributes

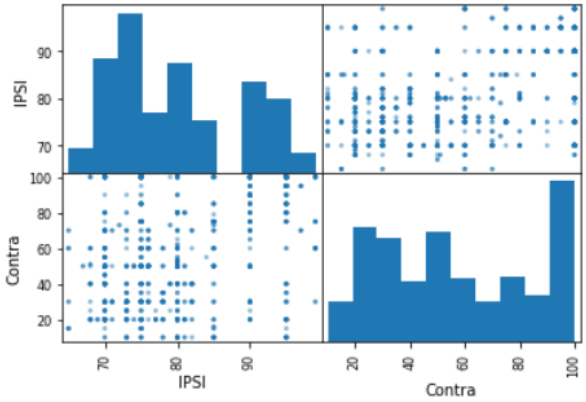


Figure 4: Correlation Diagram

1.4 Categorical Attributes

The tables in Figure 6 show the categorical data which are not are not highly skewed, they are evenly spread out, indicating they would show a good representation in performance

Indication	
A-D	489
CVA	407
TIA	388
ASx	216

IHD	
No	789
Yes	711

Hypertension	
No	777
Yes	723

Figure 6: Non skewed categorical data

The table below in Figure 7 show attributes which are highly skewed, this doesn't make them a reasonable choice for certain models such as regression-based models. This is because in a skewed data model the tail region could be considered outliers thus affecting the model's performance.

Diabetes	
No	1425
Yes	75

Arrhythmia	
No	1177
Yes	323

History	
No	1478
Yes	22

Figure 7: Skewed categorical data

1.8 Transforming data

All the categorical data was transformed into binary data, 0 and 1 representing No and Yes, so they can be used in the models as the models take numerical inputs for their computations. For the attribute 'Indication' One-Hot encoding was required to make each value have a 0 or 1, this would make sure at any one time only one of the values will be true and the rest false. In our case it looks like Figure 8 in the appendix. Once the One-Hot encoding was done each unique value in 'Indication' had been made its own attribute, so the attribute 'Indication' was no longer needed thus dropped.

1.9 Models

Below models were chosen because of the popularity and the scalability of the dataset. There is no protocol to selecting a model some may perform well and others not. These methods are chosen for testing purposes and fine-tuning which model if most efficient and provides the best results. grid search may give more accurate method even if the initial method does not perform well. It can be applied to calculate the best parameter to tune for optimal results.

5 models were chosen

1. SGD Classifier [1] – SGD classifier is a linear classifier that has internal algorithm to get the most suitable gradient decent. This performed the worst for the given data set as shown in the results below fir the given data set. It will be a good entry point to the classification problem. So, this can be used for the initial performance. We can compare other models with this model.

2. Random Forest Classifier [2] - Random Forest classifier is an ensemble classifier. Random forest classifier has multiple trees, these trees vote for each class then the most voted class will be the predictive class. You can clearly see the results below.

3. Support Vector Classifier - Kernel Linear (SVC2)[3] – Support vectors are used to divide the dataset into several classes in multiple dimensions. This was used to test the data with a linear kernel and compare the result with a nonlinear kernel.

4. Support Vector Classifier - Kernel RBF (SVC1) [4] - RBF stands for Radial Basis Function. Since RBF kernel is a non-linear kernel, if there is more nonlinearity, this kernel could produce a higher accuracy than the others. Performance with the linear and non-linear kernels are compared below.

5. Gausian Naive Biased[5]- This model is a simple probabilistic classifier that internally runs Bayes' theorem with strong independence assumptions between the features.

Each model was tested with the numerical attributes to give results and then each categorical attribute was tested alongside the numerical attributes. The model with the most optimal numerical and categorical attribute combination was tested with the best model

2 Results

(TN, FP, FN, TP)

	SGD	RFC	SVC1	SVC2	NB
Accuracy	77.55 %	89.70 %	89.96 %	85.77 %	85.07 %
Precision	0.74	0.88	0.89	0.84	0.84
Recall	0.76	0.88	0.89	0.83	0.83
F1 Score	0.74	0.89	0.89	0.83	0.83
Confusion Matrix	541 203 81 320	672 72 55 346	683 61 55 346	670 74 96 305	664 80 92 309

Validation with Numerical attributes

	SGD	RFC	SVC1	SVC2	NB
Accuracy	72.93 %	88.43 %	88.89 %	85.19 %	84.38 %
Precision	0.59	0.88	0.88	0.84	0.83
Recall	0.57	0.87	0.87	0.83	0.83
F1 Score	0.57	0.87	0.87	0.84	0.83
Confusion Matrix	473 91 210 90	522 42 58 242	521 43 55 245	510 54 72 228	502 62 68 232

Validation with Numerical attributes vs diabetes

	SGD	RFC	SVC1	SVC2	NB
Accuracy	72.57 %	85.18 %	88.43 %	84.61 %	84.38 %
Precision	0.72	0.85	0.87	0.83	0.83
Recall	0.74	0.83	0.86	0.81	0.83
F1 Score	0.73	0.84	0.86	0.82	0.83
Confusion Matrix	409 155 71 229	520 44 80 220	520 44 61 39	509 55 82 218	503 61 67 233

Validation with Numerical attributes vs Indication

	SGD	RFC	SVC1	SVC2	NB
Accuracy	75.24 %	86.00 %	88.31 %	85.19 %	85.65 %
Precision	0.71	0.87	0.88	0.84	0.84
Recall	0.73	0.87	0.87	0.82	0.84
F1 Score	0.69	0.87	0.87	0.83	0.84
Confusion Matrix	337 227 40 260	511 53 51 249	523 41 59 241	512 52 81 219	505 59 66 234

Validation with Numerical attributes vs IHD

	SGD	RFC	SVC1	SVC2	NB
Accuracy	72.21 %	87.62 %	88.89 %	85.30 %	84.96 %
Precision	0.75	0.84	0.88	0.84	0.84
Recall	0.61	0.83	0.86	0.82	0.84
F1 Score	0.60	0.83	0.87	0.83	0.84
Confusion Matrix	543 21 223 77	508 56 73 277	524 40 80 239	512 52 80 220	503 61 65 235

Validation with Numerical attributes vs Hypertension

	SGD	RFC	SVC1	SVC2	NB
Accuracy	75.68 %	87.85 %	88.31 %	85.30 %	85.30 %
Precision	0.65	0.84	0.88	0.83	0.84
Recall	0.65	0.84	0.87	0.81	0.83
F1 Score	0.65	0.84	0.87	0.82	0.84
Confusion Matrix	431 133 138 162	499 65 63 237	523 41 58 242	510 54 83 217	503 61 67 233

Validation with Numerical attributes vs Arrythmia

	SGD	RFC	SVC1	SVC2	NB
Accuracy	72.81 %	88.54 %	89.24 %	85.19 %	84.61 %
Precision	0.68	0.87	0.88	0.84	0.84
Recall	0.59	0.87	0.87	0.83	0.83
F1 Score	0.58	0.87	0.87	0.84	0.83
Confusion Matrix	528 36 228 72	516 48 55 245	521 43 55 245	510 54 73 227	501 63 66 234

Validation with Numerical attributes vs History

3 Evaluation and Discussion

Since this is a Classification problem, Accuracy cannot give the best understanding about the results. So, a check on both Precision and Recall values was done for the validation task.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Precision value gives us how many True Risk counts from to True Risk counts and No Risk counts that have been counted as Risks. Recall value gives us how many True Risk Counts with True Risk counts and Risk values that have been counted as No Risk. It can be a crucial situation if a Risk has been encountered as a No Risk, but it is not harmful is a No Risk patient encountered as a Risk person. So we need higher Precision values and low Recall values for a better model. From the results you can identify the model with the highest precision and lowest recall was SVC-RBF. This model consistently gave the best results in the confusion matrix for most TP TN and least FN. If the data set was larger this model would have produced even better precision for this classification problem.

We had highest Recall values in Random Forest and SVC -RBF classifier with 0.89 almost every feature set that has been tried. For further processing we can choose the threshold value to get close to 1 for better precision of values. From figure 9 to 12 in the Appendix are the graphs which show the precision, recall and accuracy and F1 score vs the model to determine which model is best. F1 score takes the combination of precision value and recall value, it could be said it conveys the balance among the two values, to work out F1 score it is '2*((precision*recall)/(precision + recall)'. The graphs show that numerical features always performed the best.

Appendix

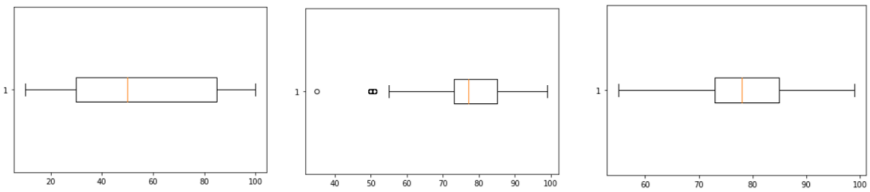
	Random	Id	Indication	Diabetes	IHD	Hypertensi	Arrhythmia	History	IPSI	Contra	label
2	0.602437	218242	A-F	no	no	yes	no	no	78	20	NoRisk
3	0.602437	159284	TIA	no	no	no	no	no	70	60	NoRisk
4	0.602437	106066	A-F	no	yes	yes	no	no	95	40	Risk
5	0.128157	229592	TIA	no	no	yes	no	no	90	85	Risk
6	0.676862	245829	CVA	no	no	no	no	no	70	20	NoRisk
7	0.916897	169990	A-F	no	no	no	yes	no	95	95	Risk
8	0.383408	196122	A-F	no	yes	yes	no	no	90	95	Risk
9	0.538333	261057	CVA	no	no	no	no	no	75	60	NoRisk
10	0.678157	256128	TIA	no	no	yes	no	no	81	20	NoRisk
11	0.689331	196936	A-F	no	no	yes	yes	no	95	100	Risk
12	0.678157	174588	CVA	no	yes	yes	yes	no	75	50	Risk
13	0.655217	271863	A-F	no	yes	yes	no	no	80	40	Risk
14	0.071533	274906	A-F	no	yes	no	no	no	76	50	NoRisk
15	0.025356	224025	CVA	no	yes	yes	yes	no	75	50	Risk
16	0.637037	167053	TIA	no	yes	no	no	no	78	30	NoRisk
17	0.025356	219417	CVA	no	yes	yes	yes	no	90	100	Risk
18	0.065821	275149	ASx	no	yes	yes	yes	no	90	100	Risk
19	0.046977	292898	CVA	no	no	yes	no	no	82	40	NoRisk
20	0.479682	284552	A-F	no	no	no	no	no	75	100	Risk
21	0.065821	101248	ASx	yes	no	yes	no	no	80	80	Risk
22	0.890427	250562	A-F	no	no	yes	no	no	75	25	NoRisk
23	0.981939	217006	ASx	no	yes	no	no	no	80	75	NoRisk
24	0.890427	184827	ASx	no	no	yes	no	no	85	10	NoRisk
25	0.885271	195912	CVA	no	no	no	no	no	68	60	NoRisk
26	0.616346	269505	TIA	no	no	yes	no	no	81	40	NoRisk
27	0.616346	190968	A-F	no	yes	no	no	no	76	70	NoRisk
28	0.616346	142470	CVA	no	no	yes	no	no	50	40	NoRisk

600092_data

Figure 1: Provided data

Attribute	Value Type	NumberOfValues	Values	Comment
Random	Real	Number of Records	Unique	Real number of help in randomly sorting the data records
Id	Integer	Max of Number of Records	Unique to patient	Anonymous patient record identifier: Should be unique values unless patient has multiple sessions
Indication	Nominal	Four	{a-f, asx, cva, tia}	What type of Cardiovascular event triggered the hospitalisation?
Diabetes	Nominal	Two	{no, yes}	Does the patient suffer from Diabetes?
IHD	Nominal	Two	{no, yes}	Does the patient suffer from Coronary artery disease (CAD), also known as ischemic heart disease (IHD)?
Hypertension	Nominal	Two	{no, yes}	Does the patient suffer from Hypertension?
Arrhythmia	Nominal	Two	{no, yes}	Does the patient suffer from Arrhythmia (i.e. erratic heart beat)?
History	Nominal	Two	{no, yes}	Has the patient a history of Cardiovascular interventions?
IPSI	Integer	Potentially 101	[0, 100]	Percentage figure for cerebral ischemic lesions defined as ipsilateral
Contra	Integer	Potentially 101	[0, 100]	Percentage figure for contralateral cerebral ischemic lesions
Label	Nominal	Two	{risk, norisk}	Is the patient at risk (Mortality)?

Figure 2: Data



Contra IPSI IPSI– outliers removed

Figure 5 Boxplot

	A-F	ASx	CVA	TIA
0	1	0	0	0
1	0	0	0	1
2	1	0	0	0
3	0	0	0	1
4	0	0	1	0
...	...	...	...	...
1495	1	0	0	0
1496	1	0	0	0
1497	0	0	0	1
1498	1	0	0	0
1499	0	0	1	0

Figure 8: Indication after One Hot

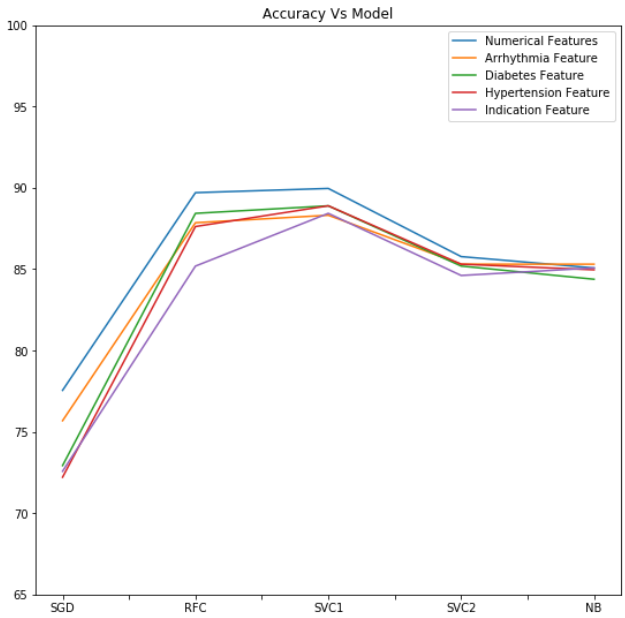


Figure 9: Accuracy vs Model data

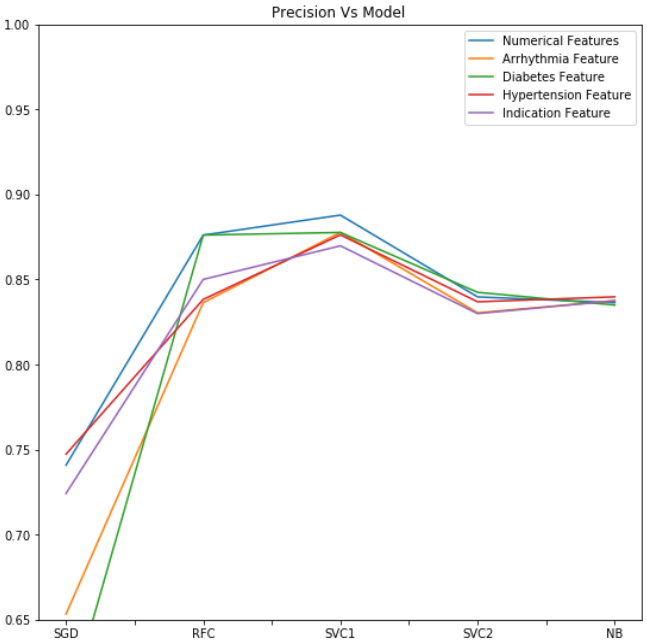


Figure 10: Precision vs Model

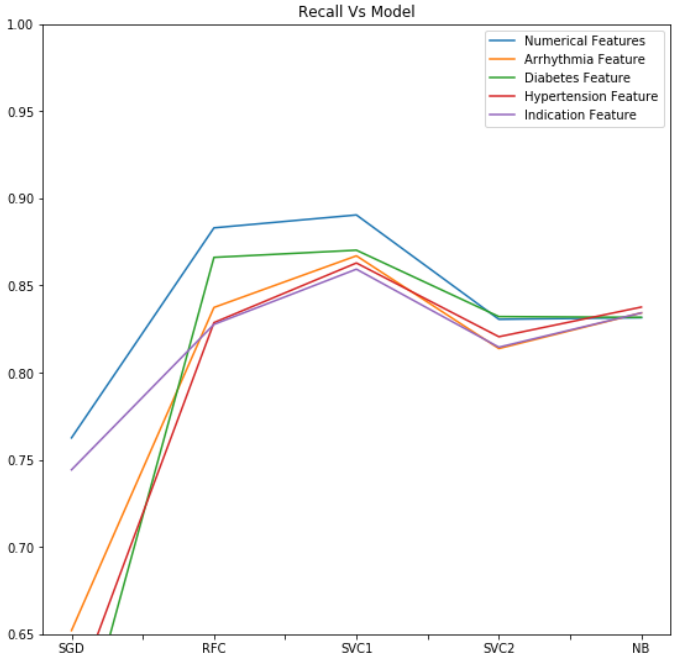


Figure 11: Recall vs Model

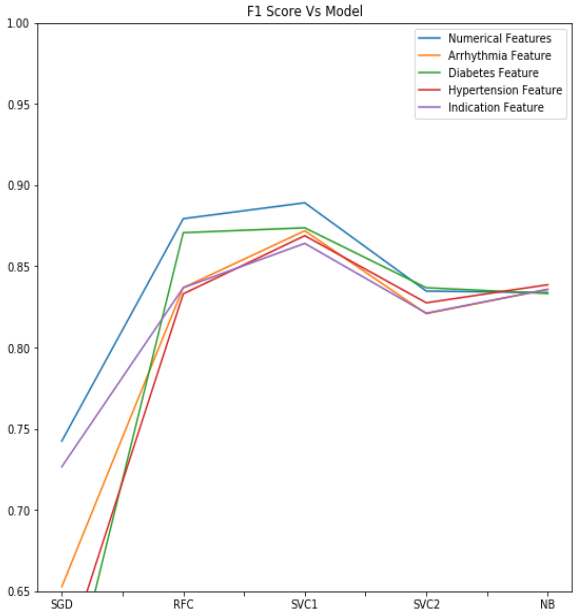


Figure 12: F1 Score vs Model

References

[1] Kabir, F., Siddique, S., Kotwal, M.R.A. and Huda, M.N., 2015, March. Bangla text document categorization using stochastic gradient descent (sgd) classifier. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1-4). IEEE.

[2] Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), pp.217-222.

[3] Han, S., Qubo, C. and Meng, H., 2012, June. Parameter selection in SVM with RBF kernel function. In *World Automation Congress 2012* (pp. 1-4). IEEE.

[4] Liu, B., Yang, Y., Webb, G.I. and Boughton, J., 2009, April. A comparative study of bandwidth choice in kernel density estimation for naive Bayesian classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 302-313). Springer, Berlin, Heidelberg.

[5] Joyce, J., 2003. Bayes' theorem.