University of the Balearic Islands

Master in complex systems

Applied data analysis and machine learning, Classical machine learning

Author

Pablo Gallardo Calleja

Analysis of a clinical database

Contents

1 Introduction

During the digestion, the body break down carbohydrates into molecules of different types of sugar. One of these is the glucose, which is one of the main sources of energy for our body. After the breakdown, glucose enters the bloodstream, where it is distributed through body. However, glucose cannot enter most tissue on its own; it needs the help of a hormone called insulin, which is produced in the pancreas. Therefore, the presence of this hormone is fundamental for overall well-being.

However, if a person has diabetes, the production of insulin on the body decrease, or the person becomes resistant to this hormone. This leads to an increase in glucose levels in the bloodstream, a condition known as hyperglycemia.

Poorly management or recognition of hyperglycemia can significantly increase the risk of both mortality and morbidity. Therefore, many tools are employed to monitor and control glucose levels in patients in the intensive care unit (ICU). However, in the case of inpatients outside of ICU, glucose control tends to be more arbitrary and may result in either a lack of treatment or large fluctuations in glucose levels [2],[1],[3]. This poor glycemic control can increase the likelihood of future patient readmission.

Consequently, in [3] the measurement of the HbA1c is analyzed its influence on readmission rate. The authors propose to use this marker to determine the presence of diabetes, with the aim of predicting the likelihood of future patient readmission.

2 Methodology

2.1 Data cleaning

The dataset will be cleaned and preprocessed to prepare it for the application of machine learning algorithms for data classification. The goal of this analysis is to remove features that are less relevant to the classification task. This will be achieved through data inspection, as well as by employing embedded approaches and wrapper methods for feature selection.

2.2 Data analysis

Once the dataset has been cleaned, it is ready to be analyzed. The following algorithms will be applied to analyze the dataset:

Algorithm	**Parameters**
k-neighbors (KNN)	n_neighbors= 2
Decision tree (DT)	depth= 6
Random Forest (RF)	depth= 6 n_estimators= 100
Support vector machine polynomial (SVM)	degree= 4
Support vector machine sigmoid (SVMS)	–
Support vector machine rbf (SVMR)	–
Support vector machine linear (SVML)	–
Logistic regression (LR)	$C = 0.9$
Naive Bayes (NB)	–

Table 1: The left column on the table list the algorithms used for the classification, while the right column displays the parameter values passed to the corresponding scikit-learn functions.

To evaluate the performance of the models, the followings metrics will be computed:

- **Accuracy:** Measures the proportion of correctly identified samples.
- **Cost matrix:** Mistakes on the classification of patient's medical condition can have serious consequences. Therefore, wrong classifications will be penalized using a cost matrix. Since it is especially important to correctly identify sick patients, the cost matrix will assign a higher penalty to false negatives (i.e., cases where a sick patient is wrongly classified as healthy):

$$C = \begin{pmatrix} -3 & 5 \\ 5 & 0 \end{pmatrix} \tag{1}$$

I consider that it's more more important to classify a patient as sick rather than healthy in case of doubt.

- **Precision:** Measure the ability of a classifier to avoid labeling negative samples as positive.
- **Recall:** Reflects the ability of the classifier to correctly identify all positive samples.
- **F1-score:** It can be seen as the harmonic mean of the precision and recall.
- **AUC:** Represents the area under the ROC curve, which plots the true positive rate against the false positive rate.

These scores will be computed for each model and across different parameter values to determine which configurations yield the best performance. Cross-validation will be used for each of parameter setting to obtain the average value of the metrics. These results will be visualized to identify the parameter values that lead to optimal overall model performance.

Finally, each model will be trained using the set of parameters that yielded the best performance. Cross-validation will be performed again to compute the evaluation metrics and asses the final performance of each model.

3 Data cleaning

The data analyzed is a subset of a larger dataset [4], which contains clinical records of patients. The selected subset includes 37 features related to patients with diabetes treated in hospitals across the United States:

Group	Column Name	Data Type
Demographics	Race	nominal
	Gender	nominal
	Age	nominal
Admission Details	Admission type Id	int64
	Discharge disposition Id	int64
	Admission source Id	int64
	Time in hospital	int64
Procedures and Utilization	Num lab procedures	int64
	Num procedures	int64
	Num medications	int64
	Number outpatient	int64
	Number emergency	int64
	Number inpatient	int64
Diagnoses	Diagnoses 1	float64
	Diagnoses 2	float64
	Diagnoses 3	float64
	Diagnoses 4	float64
	Number of diagnoses	int64
Diabetes Medications	Chlorpropamide, glimepiride, acetohexamide...	nominal
Others	Change	nominal
	DiabetesMed	nominal
	Readmitted	nominal

Table 2: Features that are going to be analyzed

A preliminary analysis of the dataset reveals some issues with the features:

- **Race and gender:** These features contain missing values. Conseqently, rows with missin values will be removed.

- **Metformin-rosiglitazone, examide, troglitazone, acetohexamide, tolbutamide:** These features contain one value. Since we are interested in classifying the data, columns with only one unique value, don't provide any useful information for the classification task. Therefore, these columns will be removed.

- **Diagnoses 4:** This feature present a unique value for each entry, with more than two decimals places. However, each entry should correspond to a disease code from [5], where each of codes have less than two decimals. Therefore, the information of this feature doesn't have interpretation, and can be safely removed.

This analysis has been performed to identify those features that are irrelevant to the classification process. These featured were removed to simplify the dataset for the classification task.

On the other hand, the patient number feature will be used to check for duplicate entries corresponding to the same patient. The results, shown in 5, shows that 154 patients appears more than once time in the dataset. These duplicate entries will be removed, and only the first occurrence of patient will be retained. This is to avoid trends on the classification. Furthermore, the patient number is just an ID assigned to identify patients within the system, and it will have no impact on the classification. Therefore, it can be safely removed from the dataset.

On the other hand, the dataset contain nominal data, which cannot be directly used on machine learning algorithms. Consequently, these categorical data will be mapped to numerical values to perform the classification process:

Column	Data	Map
readmitted / diabetesMed	Yes	1
	No	0
gender	Female	0
	Male	1
change	No	0
	Ch	1
medicine	No	0
	Steady	3
	Up	1
	Down	2
race	Caucasian	0
	Other	1
	Hispanic	4
	African American	3
	Asian	5

Table 3: Categorical encoding for selected features

Moreover, there is another nominal feature in the dataset which correspond to the patient's age. This feature is given in the following format:

$$[10x, 10(x+1)[\text{ with } x \in \{0, 1, 2, 3, 4...\} \tag{2}$$

For this feature, the map will assign the value x for each interval.

Finally, while the diagnoses are already in numerical format, each entry represent a disease code from [5]. These codes differs even when they refers to conditions that affect the same part of the body. Therefore, codes will be grouped into seven different categories, following the classification outlined in the table 2 from [1]. A unique numerical value will be assigned to category as follow:

Category	Map
Circulatory	0

Respiratory	1
Digestive	2
Diabetes	3
Injury	4
Musculoskeletal	5
Genitourinary	6
Neoplasms	7
Other	8

Table 4: ICD-9 Diagnosis Group Encoding

Finally, a Min-max scale is applied to ensure that all the feature has the same scale, helping to avoid convergence problems in the training process.

Before analyze the result, it is important to note that up to this point, the data has been cleaned using intuitive arguments based on the logic of the data or their values. However, it is possible that others features, which are not immediately obvious, may still be irrelevant for the classification task. Therefore, before analyzing the importance of each feature, a feature selection process will be carried out. This process will be done using two methods:

- **Embedded approach based on decision trees**: The number of times each feature is used in decision-making will be computed to assess its importance.

- **Permutation feature importance:** This method is particularly useful for non-lineal estimators. It consists on shuffle all the features and observing the degradation on model's performance score. While decision trees are effective for feature selection, they can assign more importance to the features that may no be predictive on unseen data. This method avoid to address this issue since it can be computed on unseen data.

The results:

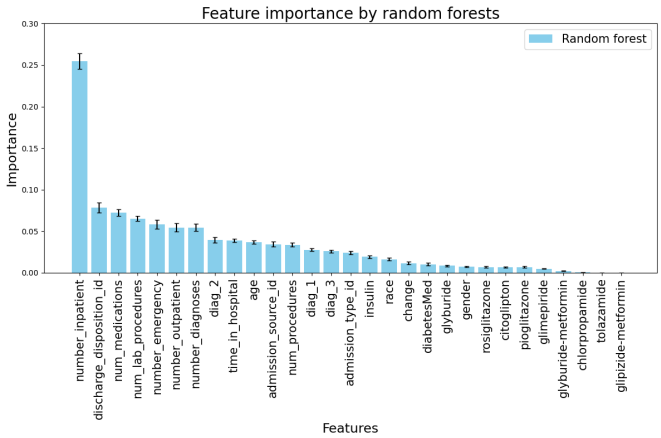

Figure 1: This picture compute the feature importance obtained by the generation of hundredths of tress in a Random Forest

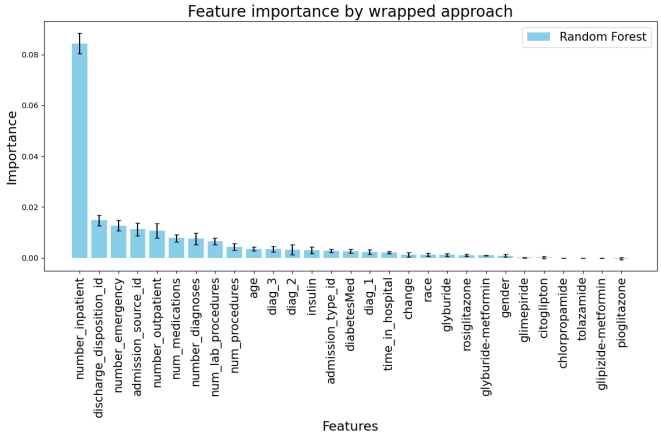

Figure 2: This picture shows the result of the permutation importance applied to a Random Forest

From the feature selection results, it can be observed:

- **Chlorpropamide, tolazamide, glyburide-metformin, glipizide-metformin, pioglitazone:** These are irrelevance for the classification task. This happens because there is an unbalance on the entries of this feature.

- **number_inpatient:** This is the most important feature, followed closely by the discharge disposition.
- The results obtained by the two methods differs for some features. This discrepancy occurs because the the embedded approach classifies features based on their intrinsic importance. However, the permutation approach is based on a maximization of model's accuracy.

4 Results

The results of the performance of each model:

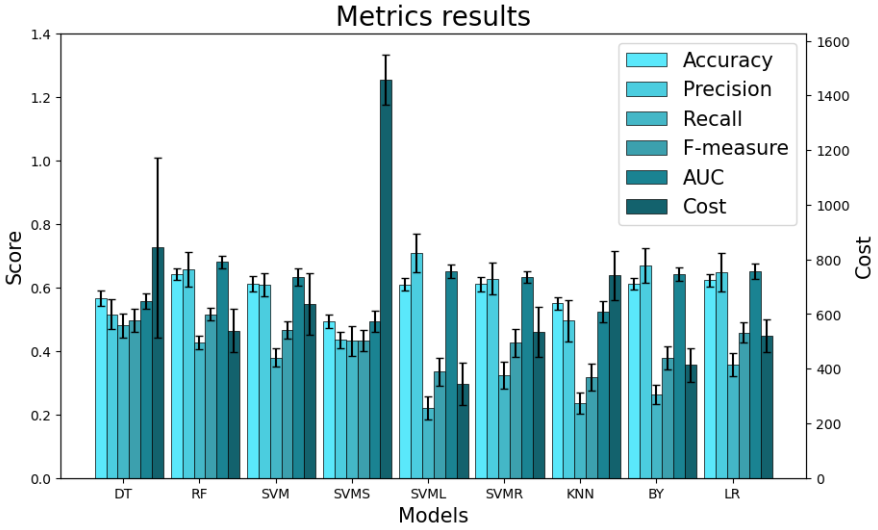

Figure 3: Results of the performance of each model. Decision tree (DT), Random Forest (RF), support vector machine with polynomial kernel (SVM), support vector machine with sigmoid kernel (SVMS), support vector machine with linear kernel (SVML), support vector machine with rbf kernel (SVMR), K-neighbors (KNN), logistic regression (LR) and naive bayes (NB)

The numerical result of the metrics can be found at 6. The picture shows that the accuracy of each model is between 0.55 and 0.65, meaning that, they perform better than a random classifier except for SVMS. SVMS present an accuracy of 0.490 ± 0.017, which means that it results are similar to choose randomly a class. However, there are significant differences in the cost of classification between models. As seen in the picture, the support vector

machine with a sigmoid kernel shows has a notably high cost, contrasting with the other models. Therefore, the choice of the model will have a substantial impact on the number of misclassifications. The following table displays the models with the best and worst performance for each metric:

Metric	Best model	Worst model
Accuracy	RF	SVMS
Precision	SVML	SVMS
Recall	DT	SVML
F1-measure	RF	KNN
AUC	RF	SVMS
Cost	SVMR	SVMS

Table 5: Models with the best and worst result for each metric

These results shows that the worst performing model is clearly the support vector machine with the sigmoid kernel, as it is outperformed in nearly all the metrics. On the other hand, the best model is the Random Forest, which performs the best overall. This model ranks first in three metrics and is ranked second or third in the cases where it doesn't win. This can also be seen with the ROC curve:

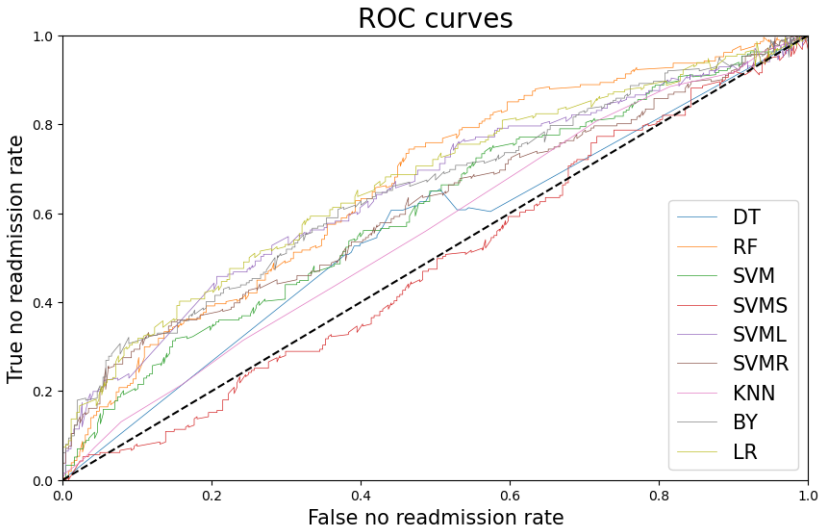

Figure 4: ROC curve

The picture shows that the best model is the Random Forest, as its furthest from the random classifier's curve. It is also shows graphically that the SVMS and KNN are the worst models, which barely perform better than a random classifier.

5 Conclusions

In this report, a subset of the dataset analyzed in [3] has been successfully examined. To achieve this, the data was cleaned an preprocessed, providing insights into the relevance of various features. This process allows us to determine that the features of chlorpropamide, tolazamide, glyburide-metformin, glipizide-metformin, pioglitazone, metformin-rosiglitazone, examide, troglitazone, acetohexamide, tolbutamide and diagnose 4 are irrelevant for classification purposes.

Among the various algorithms applied to the classification task, the Random Forest demonstrated the best overall performance and consistency across all evaluation metrics. In contrast, the results shows that other methods based on neighbors like KNN, single decision tree or support vector machine with a sigmoid kernel yielded accuracies barely distinguishable from that of a random classifier. Finally, it is also interesting to see that the Naive Bayes method also shows a performance that can compete with the Random Forest in some aspect. This could be an indicator that data is independent. However, it significantly under-performs compared to Random Forest in key metrics such as recall and F1-score.

A Appendices

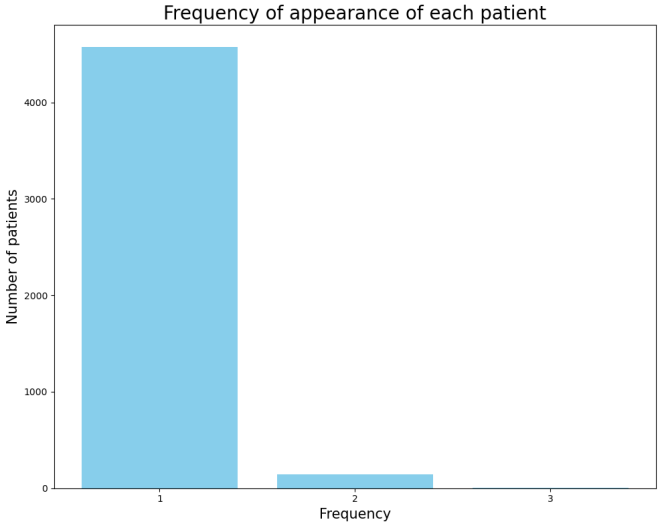

Figure 5: Frequency of appearance of each patient

Model	Accuracy	Precision	Recall	F1-measure	AUC	Cost
DT	0.564 ± 0.034	0.511 ± 0.058	0.485 ± 0.048	0.496 ± 0.048	0.554 ± 0.039	842 ± 30058
RF	0.645 ± 0.016	0.662 ± 0.045	0.426 ± 0.029	0.518 ± 0.031	0.678 ± 0.027	516 ± 88
SVM	0.606 ± 0.018	0.597 ± 0.056	0.373 ± 0.0416	0.458 ± 0.041	0.627 ± 0.025	658 ± 82
SVMS	0.490 ± 0.017	0.433 ± 0.030	0.433 ± 0.018	0.432 ± 0.017	0.494 ± 0.032	1477 ± 103
SVML	0.608 ± 0.024	0.706 ± 0.058	0.221 ± 0.018	0.336 ± 0.025	0.649 ± 0.017	347 ± 103
SVMR	0.613 ± 0.031	0.632 ± 0.044	0.336 ± 0.041	0.437 ± 0.035	0.635 ± 0.030	538 ± 140
KNN	0.557 ± 0.020	0.515 ± 0.047	0.244 ± 0.012	0.331 ± 0.018	0.537 ± 0.025	709 ± 99
BY	0.611 ± 0.019	0.673 ± 0.074	0.265 ± 0.031	0.378 ± 0.033	0.641 ± 0.025	414 ± 87
LR	0.622 ± 0.018	0.643 ± 0.040	0.359 ± 0.020	0.461 ± 0.022	0.647 ± 0.024	526 ± 87

Table 6: Numerical result of the performance of each model

References

[1] Claresa S Levetan et al. "Unrecognized Diabetes Among Hospitalized Patients". In: *Diabetes Care* 21.2 (Feb. 1998), pp. 246–249. ISSN: 0149-5992. DOI: 10.2337/diacare.21.2.246. eprint: https://diabetesjournals.org/care/article-pdf21/2/246/585515/21-2-246.pdf. URL: https://doi.org/10.2337/diacare.21.2.246.

[2] Guillermo E. Umpierrez et al. "Hyperglycemia: An Independent Marker of In Hospital Mortality in Patients with Undiagnosed Diabetes". In: *The Journal of Clinical Endocrinology & Metabolism* 87.3 (Mar. 2002), pp. 978–982. ISSN: 0021-972X. DOI: 10.1210/jcem.87.3.8341. eprint: https://academic.oup.com/jcem/article-pdf87/3/978/9155454/jcem0978.pdf. URL: https://doi.org/10.1210/jcem.87.3.8341.

[3] Beata Strack et al. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records". In: *BioMed Research International* 2014.1 (2014), p. 781670. DOI: https://doi.org/10.1155/2014/781670. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2014/781670. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/781670.

[4] *Cerner health facts*. en. https://www.uthsc.edu/cbmi/datacerner.php. Accessed: 2025-4-5.

[5] *ICD-9-CM volume 1 diagnosis codes*. en. http://www.icd9data.com/2015Volume1/default.htm. Accessed: 2025-4-5.