Matched-Pair Analysis Using Machine Learning to Predict 1-year Mortality in Newborn Twins

Everton Mendonça de Jesus[1], **Lucas Calais-Ferreira**[2], **Marcos Ennes Barreto**[1]

[1]AtyImoLab, Computer Science Department, Federal University of Bahia
Av. Adhemar de Barros s/n, Campus Ondina, Postcode: 40.170-110, Salvador, Brazil

[2]Centre for Epidemiology and Biostatistics
Melbourne School of Population and Global Health
The University of Melbourne, Melbourne, Australia

Abstract. *Twin pair analysis is a valuable tool for assessing familial risk factors related to several outcomes, including diseases. Machine learning models are standard, powerful tools for prediction, although their use for twin pair analysis is not fully suitable as most models do not account for the existing correlation between twin pairs. In this paper, we have focused on assessing the suitability of machine learning models to predict 1-year mortality using twin data extracted from Brazilian healthcare databases. We have evaluated five models and also used a proposed strategy for matched pair analysis to build an alternative dataset supposed to provide improvements for classification tasks. Our results showed that i) Gradient Boosting was the best classification model, and ii) the matched-pair strategy used did not improve our results as expected.*

Resumo. *A análise de pares de gêmeos é uma ferramenta importante para avaliar fatores de risco familiares relacionados a diversos problemas, inclusve doenças. Modelos de aprendizado de máquina são ferramentas consolidadas utilizadas em diversas tarefas de predição nas mais diversas áreas. Porém, elas não são totalmente adequadas para a análise de conjuntos de dados onde os pares de registros possuem forte correlação entre si, como é o caso de gêmeos. Este artigo avalia a adequabilidade de modelos de aprendizado de máquina para predizer a mortalidade até um ano de gêmeos nascidos no Brasil, utilizando dados extraídos de bases de dados públicas. Além disso, também é avaliado um método para análise de dados pareados, criando uma base de dados alternativa para verificar se a aplicação dos modelos de aprendizado de máquina promove ganhos no processo de classificação. OS resultados demonstraram que i) o modelo de aprendizado de máquina Gradient Boosting obteve os melhores resultados na tarefa de classificação e ii) a estratégia para análise de dados pareados não melhorou os resultados conforme esperado.*

1. Introduction

The epidemiological study of twin pairs is a valuable tool in public health, as it allows to investigate familial (both genetic and environmental, or lifestyle) risk factors associated with human disease [van Dongen et al. 2012]. Due to the similarity of twin pairs – identical or monozygotic (MZ) pairs share 100% of their genes – it is possible to study twin pairs who are discordant for an outcome (such as mortality) and discover which risk factors lead to such outcome, while holding genetic factors constant.

Fraternal or dizygotic (DZ) twin pairs share 50% of their genes as any other pair of siblings, but they also share many early environmental factors (including the mother's womb). Male-female twin pairs are a special case of DZ twins as they are also discordant for sex. This allows for investigating the role of sex differences in human traits and conditions while controlling for age, shared environmental factors and 50% of genetic differences. This has been done before for studying neonatal mortality in a dataset of all twins born in the USA from 2000-2005, finding a substantial higher risk of neonatal mortality for the male sex (of about 50%), even after adjusting for familial factors [Bogl et al. 2017].

Previous studies have used Brazilian public healthcare data to investigate outcomes related to twin pairs, specifically on morbidity/mortality of the second twin based on the time interval between twin delivery [Fava et al. 2001], perinatal variables [Costa et al. 1998] and obstetric complications [Sá et al. 2008]. However, these studies have used small (not sufficiently representative) datasets or did not properly accounted for familial factors [Silva et al. 2017], as expected from a twin study design. Our work aimed at to fill this gap through the deployment and analysis of different classification models targeting the prediction of 1-year infant mortality over paired data.

The use of machine learning models to efficiently classify paired records (such as twins) is very challenging. This issue is mentioned in the literature as the "matched pair problem", which comprises scenarios where paired data must be analyzed through dependent tests. We have developed bespoke models and assessed their performance running over paired and independent/disjoint records. We also used an approach proposed by [Adler et al. 2011] to check for performance improvements in our models.

The objectives of our study are twofold. Firstly, we aimed at to check which machine learning model is more suitable to predict the probability of a newborn twin decease in the first year of his life. The second goal was to assess if the matched-pair methodology proposed by [Adler et al. 2011] provides improvement in the accuracy, precision, recall and F1-score metrics when applied to the machine learning models used in this study. Developing a good approach to analyze datasets where records are organized in pairs is important to understand the relations between them and how these relations can be used to extract better insights from data.

This paper is structured as follows: Section 2 presents a literature review on related publications using similar classification methods. Section 3 describes the machine learning models used in this study, followed by the methodology used to obtain and perform linkage in the databases, deploy and evaluate these models (Section 4). We present our results and discussion in Section 5 and some concluding remarks and future research prospects on infant mortality supported by machine learning techniques in Section 6.

2. Related Work

For this study, we searched public repositories (PubMed, ACM, Scopus, MedRxiv and IEEE) for publications focusing on "paired data analysis", "twin pairs", and the use/design of prediction models specially related to mortality without restrictions on the publication date. As "paired data" is generally associated with biological data, such as DNA/RNA sequencing and analysis, to retrieve existing works properly related to ours, we restricted our search to methodological aspects of twin data analysis or the use of twin

(population-based) data. We ended up with seven papers discussing methods and case studies, besides other four conceptual works related to twin pair analysis.

Regarding methods, [Carlin et al. 2005] discuss the use of specialized regression models for twin data, with emphasis on coefficients more appropriate to within-twin-pair and between-pair analysis, as well as on the interpretation of continuous and binary outcomes. They claim that, despite the popularity of regression models for risk factor analysis, their use over twin data poses some complexity related to the modelling of existing correlations between twin records which is not straightforward. The use of multiple regression models and their comparison with maximum-likelihood models for twin pair analysis was earlier discussed by [Rao et al. 1987] and [Cherny et al. 1992].

[Theiler 2013] argues that machine learning models using paired data sets for training present improved performance, but his evaluation was based on image analysis within a plume detection application. A performance comparison of feature selection models is presented by [Liang et al. 2018]. Although using genomic data sets, we have included this work as it presents a comprehensive analysis of such models categorized as statistical tests, logistic regression or boosting, and observed complexities when used for breast cancer analysis.

In [Silva et al. 2017], the authors have proposed a prediction model with subsequent notification whether a newborn is at high risk of death in 1-year time. They have tested different classification models over live births and mortality data from Brazilian public databases (SINASC and SIM databases), but they do not account for twin pairs.

The work developed in [Rittenhouse et al. 2019] aims to improve preterm newborn identification by using machine learning algorithms. As cited by the authors, preterm is one of the main causes of neonatal deaths [Liu et al. 2016] and creating ways to previously identify these cases leads to more accurate clinical interventions.

As cited before, one of the main problems when dealing with twin pairs is the matched pair problem and how to consider the intrinsic relation between the twins. As far as we know, there are scarce studies considering existing relations present in twin data when analyzed by machine learning models in the healthcare domain.

The work proposed in [Adler et al. 2011] presents a methodology to deal with matched records originated from paired organs or repeated measurements from the same organs and subjects. The authors proposed techniques to glaucoma diagnosis using both eyes of glaucoma patients. They compared the results of different machine learning algorithms in terms of misclassification rates.

In [Liang et al. 2018], a review of matched-pair feature selection techniques applied to gene expression data analysis is presented. The authors compared two real datasets against ten feature-selection methods by applying them to three classification methods. After the tests, they concluded that one of the methods (WL_2Boost – Boosting Weighted L2 Loss) had best performance for a feature list of moderate size and when execution time is not a constraint. For high dimensional features, logistic regression methods such as RP-CLR and PCU-CLR have shown better performance.

As mentioned earlier, we found few papers discussing the matched-pair problem applied to twin data and how to use this kind of data in machine learning models to

extract information. The matched-pair approach can also be used in problems where data points are organized in pairs, so finding a methodology to deal with this question is quite necessary and valuable. Based on this, we considered important to investigate previous methods and techniques used to analyze data records with an intrinsic relation, such as our newborn twins cohort.

3. Machine Learning Models for Prediction

In the last years, machine learning has evolved and was widely adopted in healthcare for different tasks, including electronic health records (EHR) linkage and analysis [Shickel et al. 2018], genomic analysis [Libbrecht and Noble 2015], and development of risk prediction models [Christodoulou et al. 2019], [Denaxas et al. 2018].

In this section, we aimed at to provide an overview of the machine learning models used in this study: Logistic Regression, Decision Tree, Support Vector Machines, Random Forest, and Gradient Boosting. All these algorithms are defined as "supervised" learning methods, meaning that they depend on labeled datasets (a class attribute to allow for categorizing each record into one specific class) to build their knowledge.

Logistic Regression is a statistical method used to predict the value taken by a categorical variable, usually binary (such as pass/fail, win/lose), from a set of continuous or binary variables. It is used in various topics in healthcare, for example in [Liang et al. 2018], [Adler et al. 2011] and [Gomes et al. 2010].

Decision Tree is a machine learning algorithm that analyzes the data and creates a "tree" (model) by using attributes as decision nodes and taking different branches to reach leaf nodes, representing the classes. It is an efficient algorithm with good accuracy to most applications. The use of decision trees in medicine is described in [Podgorelec et al. 2002]. A more recent studies using decision trees in healthcare is shown in [Game et al. 2019] and [Kaur et al. 2019].

Support Vector Machines (SVM) is a widely used machine learning model for binary classification problems. The main goal of the algorithm is to split the data points into two different sides (classes) separated by a well-defined line (known as hyperplane). It can be applied to trivially (linearly) separable data points, as well as to more complex (non-linearly separable) data points through the use of different classification functions (kernels). Some studies describing the use of SVM in healthcare are [Son et al. 2010], [Razzaghi et al. 2016], [Game et al. 2019] and [Naraei et al. 2016].

Random Forest works by combining different decision trees into a group to create a powerful classification model. The user can define how many decision trees will be created in his model, and the final classification is achieved by combining individual results. Some previous works that use Random Forests are described in [Kaur et al. 2019] and [Imani et al. 2019].

Gradient Boosting is another derivative from decision trees. It is similar to the AdaBoost algorithm, using an ensemble of decision trees to perform classification. The main difference between these models is that Gradient Boosting can be deeper (more decision levels) than AdaBoosting. The use of Gradient Boosting in healthcare problems is described in [Hatton et al. 2019] and [Zafar et al. 2019].

Given the existence of several classification models with performance (accuracy)

highly dependent on the input dataset, parameters setting and domain knowledge, the standard approach to design classification and prediction models is testing a set of models and then choosing the one presenting better results and, desirably, generalization.

4. Methodology

We have applied a standard data science approach in this work, comprising data linkage, preprocessing and analysis, as described in this section.

4.1. Data sources and linkage

In this study, we have used public databases (extracted through the Ministry of Health's TABNET website[1]) capturing live births (SINASC) and death records (SIM). From SINASC, twins born between 2012 and 2016 were extracted through a bespoke algorithm that checks relevant attributes to identify whether a pair of records represents a twin pair. The attributes from SINASC used in the mentioned algorithm were: 'NUMERODN', 'LOCNASC', 'CODESTAB', 'CODINST', 'CODMUNNASC', 'IDADEMAE', 'ESTCIVMAE', 'ESCMAE', 'CODMUNRES', 'GRAVIDEZ', 'PARTO', 'DTNASC', 'NATURALMAE', 'CODMUNNATU', 'CODUFNATU', 'DTNASCMAE', 'RACACORMAE' and 'SEMAGESTAC'.

After that, a deterministic linkage between SINASC and corresponding entries in SIM was carried out through a common attribute (NUMERODN) to both databases. This search was restricted to the year of birth and the subsequent year, to account for 1-year mortality. We have used a class attribute (PRESENTENOSIM) to indicate whether a newborn twin has deceased or not, respectively with values 1 and 0.

After the data linkage process, a total of 303,379 twin records were extracted ('Dataset 1'). From this total, 11,868 were matched in the SIM database, meaning that they deceased in a 1-year time from birth.

As cited in [Adler et al. 2011], a suitable technique for dealing with the matched pair problem when building training datasets is to arbitrarily select one record out of the pair during the bootstrap phase. So, after the extraction and linkage process, the full dataset with 303,379 records was processed and, for each pair of twins, one record was randomly selected, generating a new dataset of 104,023 records ('Dataset 2').

4.2. Data pre-processing

Before deploying and evaluating some classification models, we have addressed data preparation issues. Originally, 'Dataset 1' had 171 attributes. We ran a feature selection technique to reduce the feature space to 45 relevant attributes (shown in Table 1).

Tabela 1. List of attributes used for classification.

CODINST	ORIGEM	PREFIXODN	CODESTAB	CODMUNNASC	LOCNASC	IDADEMAE	ESTCIVMAE	ESCMAE
QTDFILVIVO	QTDFILMORT	CODMUNRES	GESTACAO	GRAVIDEZ	PARTO	CONSULTAS	DTNASC	HORANASC
SEXO	APGAR1	APGAR5	RACACOR	PESO	IDANOMAL	DTCADASTRO	CODANOMAL	NUMEROLOTE
DTRECEBIM	DIFDATA	NATURALMAE	CODMUNNATU	ESCMAE2010	DTNASCMAE	RACACORMAE	QTDGESTANT	QTDPARTNOR
QTDPARTCES	SEMAGESTAC	CONSPRENAT	STDNEPIDEM	STDNNOVA	CODPAISRES	PAREADO	FAMILY_ID	PRESENTENOSIM

Some attributes needed manual processing due to missing values. The attribute 'CODANOMAL', which indicates congenital malformation or chromosomal anomaly in

[1] https://datasus.saude.gov.br/informacoes-de-saude-tabnet/

newborns, was filled with 'None' for all missing values, meaning no abnormalities. The attribute 'IDANOMAL', which indicates if the newborn has a congenital anomaly, was filled with 9, meaning it was ignored during the newborn registration (according to the SINASC manual). The attribute 'SEXO' (gender) was set to 1 (male), 2 (female) or 0 (ignored). Missing values from the following attributes were filled with 0: 'QTDFILMORT', 'QTDGESTANT', 'QTDPARTNOR', 'QTDPARTCES', 'CODPAISRES' and 'CONSPRENAT'. After pre-processing, 'Dataset 1' reduced from 303,379 to 256,499 records and 'Dataset 2' reduced from 104,023 to 98,459 records.

The attributes 'CODINST', 'CODANOMAL' and 'FAMILY_ID', originally of type 'string', were encoded as numerical values as required by machine learning models. Additionally, all attributes were re-scaled.

Lastly, the datasets were divided into training and test subsets at the 70%-30% proportion, respectively. 'Dataset 1' ended with 179,549 records for training and 76,950 for test and validation, whereas 'Dataset 2' ended with 68,921 records for training and 29,538 for test and validation. Also, the attribute 'PRESENTENOSIM', that indicates if an individual has deceased or not, was selected as the class to be used during classification. All data linkage and pre-processing routines, as well as the classification models, were implemented in Python using the *scikit-learn* library.

4.3. Model selection

After preparing the datasets, we chose five models parameterised as follows:

- **Logistic Regression:** we set the parameter `max_iter` to 10,000.
- **SVM:** we used the algorithm variation C-Support Vector Classification (SVC).
- **Decision Tree, SVM, Random Forest** and **Gradient Boosting**: we used all parameters with default values.

These models were chosen due to their wide application in different healthcare studies, as cited in Section 3. Also, these methods are easy to implement and evaluate in several languages and frameworks, allowing researchers to easily reproduce the experiments presented in this work.

5. Results and Discussion

After running the chosen models, we performed a comparative analysis of the results obtained with both datasets, based on standard metrics. We also checked for any improvements when using the random choice approach proposed by [Adler et al. 2011]. Table 2 and Table 3 show the results for 'Dataset 1' and 'Dataset 2', respectively.

Tabela 2. Results obtained over 'Dataset 1'.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.972	0.711	0.395	0.508
Decision Tree	0.958	0.430	0.445	0.437
SVM (SVC)	0.972	0.749	0.345	0.472
Random Forest	0.973	0.734	0.436	0.547
Gradient Boosting	**0.974**	**0.717**	**0.477**	**0.573**

Tabela 3. Results obtained over 'Dataset 2'.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.974	0.687	0.364	0.476
Decision Tree	0.957	0.357	0.380	0.368
SVM (SVC)	0.972	0.700	0.262	0.381
Random Forest	0.975	0.697	0.406	0.514
Gradient Boosting	0.973	0.642	0.412	0.502

As noted, despite the sizes of the datasets used in this work, no substantial improvement or decrease in accuracy was observed when changing between datasets – the results are statistically close. Additionally, the other metrics presented a slightly greater variation: for all models, it is possible to see that precision, recall, and F1-score have presented better results over 'Dataset 1' when compared to 'Dataset 2'.

When comparing individual models, Gradient Boosting has presented the best results. Notably, its F1-score stands out when compared to the other models. It is also important to highlight that all Gradient Boosting results were better over 'Dataset 1' (marked in bold in Table 2) than over 'Dataset 2'.

Analyzing the ROC curves presented in Figures 1 and 2, it is possible to see the differences between the two approaches. It i clear that all the models tested over 'Dataset 1' have better results when compared to the results obtained over 'Dataset 2'. Again, Gradient Boosting has presented a better ROC curve, following the results presented in Tables 2 and 3.

Based on our experiments, it is possible to state that Gradient Boosting is a suited machine learning model to predict 1-year mortality. We can also state that the matched-pair strategy – that consists of randomly selecting one of the twin pairs – did not provide significant improvements to our classification models.

Despite the accuracy metrics be similar over both datasets, the other metrics (precision, recall, F1-score and ROC curves) have presented significant differences. Furthermore, all results indicate that 'Dataset 1' provides better results when compared to 'Dataset 2'.

6. Conclusions

In this study, we have aggregated public data of live births and mortality to build a twin cohort and evaluate machine learning models targeted to predict 1-year mortality between twin pairs.

We have investigated if a method to deal with matched pair twins provides any improvement during the prediction if a newborn will die in a specific time window. To check for that, two datasets were prepared and tested. The first one ('Dataset 1') containing all twins extracted from the Brazilian newborn database (SINASC), grouped in pairs and considered as independent records. The second dataset ('Dataset 2') was built based on a strategy that randomly selects one twin from the pair into a new dataset.

After that, the two datasets were used in Logistic Regression, Decision Tree,

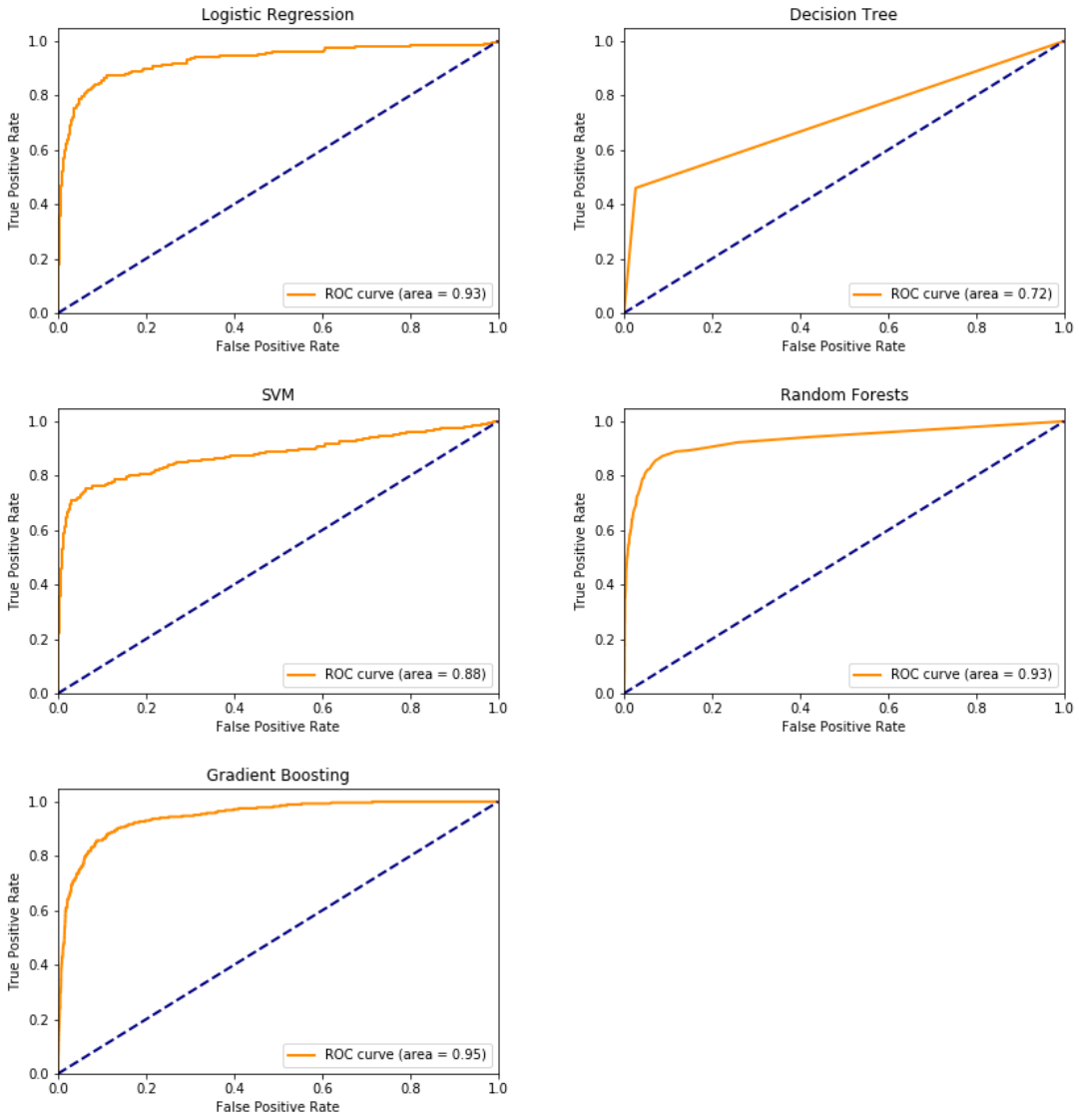

Figura 1. ROC Curves for 'Dataset 1'.

SVM, Random Forest, and Gradient Boosting classification, having their results analyzed and compared through standard statistical metrics.

The first results pointed out some differences between the two strategies depending on the analyzed metric. When looking only to accuracy, all models presented similar results for both datasets; but when looking at the other metrics (precision, recall, F1-score, and ROC curves), we can observe more significant differences. In general, 'Dataset 1' has allowed better results for all models, meaning that the matched-pair strategy used to create the 'Dataset 2' did not provide improvements to our models.

This also indicates that more studies are necessary to find an effective way to use matched pair records in machine learning models. Our future work includes testing new ways of dealing with matched pair records, by using paired T-tests and dimensionality reduction techniques, like Principal Component Analysis (PCA), as well as assessing other models (including deep learning techniques) and hyperparameter optimization.

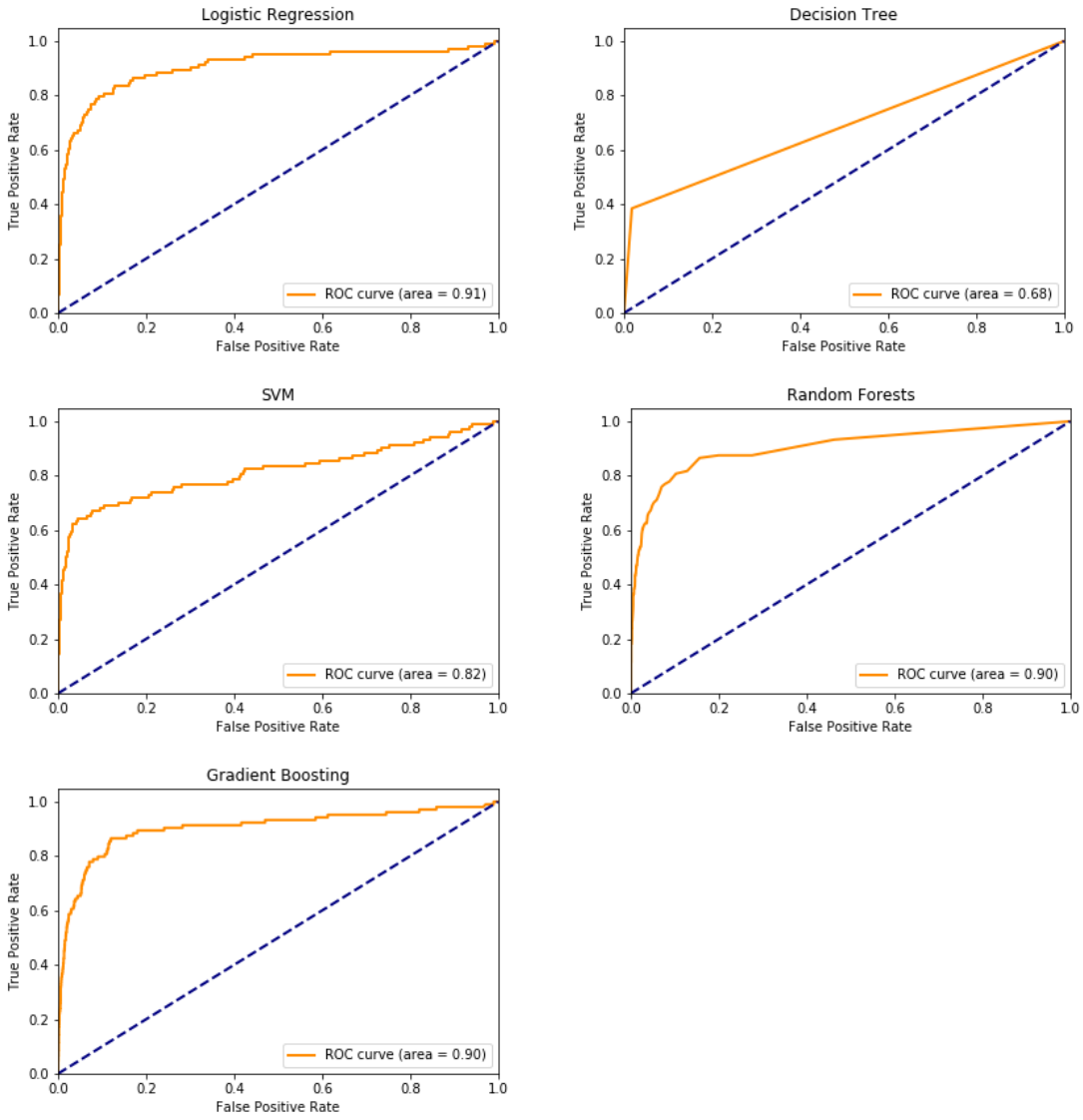

Figura 2. ROC Curves for 'Dataset 2'.

Acknowledgments

This work is supported by The Royal Society (UK), grant AL\191004. Marcos E. Barreto is a Newton International Fellow Alumnus (The Royal Society, UK) and is also with the Department of Statistics of the London School of Economics and Political Science (LSE).

Referências

Adler, W., Brenning, A., Potapov, S., Schmid, M., and Lausen, B. (2011). Ensemble classification of paired data. *Comput. Stat. Data Anal.*, 55(5):1933–1941.

Bogl, L. H., Jelenkovic, A., Vuoksimaa, E., et al. (2017). Does the sex of one's co-twin affect height and BMI in adulthood?: A study of dizygotic adult twins from 31 cohorts. *Biology of Sex Differences*, 8(1):14.

Carlin, J. B., Gurrin, L. C., Sterne, J. A. C., Morley, R., and Dwyer, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology*, 34(5).

Cherny, S. S., DeFries, J. C., and Fulker, D. W. (1992). Multiple regression analysis of twin data: a model-fitting approach. *Behavior Genetics*, 22(4):489–497.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Calster, B. V. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12 – 22.

Costa, H. d. L. F. F., Rocha, A. C. O., Galvão, A. F., Souza, J. d. A., Rigaard, A. C. d. O., and Costa, L. O. B. F. (1998). É pior o prognóstico do segundo gemelar? *Revista Brasileira de Ginecologia e Obstetrícia*, 20(5).

Denaxas, S., Stenetorp, P., Riedel, S., Pikoula, M., Dobson, R., and Hemingway, H. (2018). Application of clinical concept embeddings for heart failure prediction in uk ehr data.

Fava, J. L., Souza, E., and Camano, L. (2001). Intervalo entre o nascimento de gêmeos: Morbidade e mortalidade do segundo gemelar. *Revista Brasileira de Ginecologia e Obstetrícia*, 23(7).

Game, P. S., Vaze, V., and Emmanuel, M. (2019). Optimized Decision tree rules using divergence based grey wolf optimization for big data classification in health care. *Evolutionary Intelligence*.

Gomes, A. S., Klück, M. M., Riboldi, J., and Fachel, J. M. G. (2010). Modelo preditivo de óbito a partir de dados do Sistema de Informações Hospitalares. *Revista de Saúde Pública*, 44:934 – 941.

Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., and Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders*, 246:857–860.

Imani, F., Chen, R., Tucker, C., and Yang, H. (2019). Random forest modeling for survival analysis of cancer recurrences. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 399–404.

Kaur, P., Kumar, R., and Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, 78(14):19905–19916.

Liang, S., Ma, A., Yang, S., Wang, Y., and Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. *Computational and Structural Biotechnology Journal*, 16:88 – 97.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J., Cousens, S., and Black, R. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an

updated systematic analysis with implications for the sustainable development goals. *The Lancet*, 388.

Naraei, P., Abhari, A., and Sadeghian, A. (2016). Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In *2016 Future Technologies Conference (FTC)*, pages 848–852.

Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I. (2002). Decision trees: An overview and their use in medicine. *Journal of medical systems*, 26:445–63.

Rao, D. C., Vogler, G. P., M., M., and Russell, J. M. (1987). Maximum-likelihood estimation of familial correlations from multivariate quantitative data on pedigrees: A general method and examples. *American Journal of Human Genetics*, 41:1104–1116.

Razzaghi, T., Roderick, O., Safro, I., and Marko, N. (2016). Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values. *PLOS ONE*, 11(5):e0155119.

Rittenhouse, K. J., Vwalika, B., Keil, A., Winston, J., Stoner, M., Price, J. T., Kapasa, M., Mubambe, M., Banda, V., Muunga, W., and Stringer, J. S. A. (2019). Improving preterm newborn identification in low-resource settings with machine learning. *PLOS ONE*, 14(2):1–12.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2018). Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

Silva, C., Alves, J., Braga, O., Júnior, J., Andrade, L., and Oliveira, A. (2017). Usando o classificador naive bayes para geraÇÃo de alertas de risco de Óbito infantil. *Revista Eletrônica de Sistemas de Informação*, 16.

Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., and Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, 16:253–9.

Sá, R. A. M., Silva, N. R., and Rezende, K. R. F. (2008). Gestação gemelar: problemas em dobro? *Femina*, 36(12).

Theiler, J. (2013). Matched-pair machine learning. *Technometrics*, 55.

van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G., and Boomsma, D. I. (2012). The continuing value of twin studies in the omics era. *Nature Reviews Genetics*, 13(9):640–653.

Zafar, F., Raza, S., Khalid, M. U., and Tahir, M. A. (2019). Predictive analytics in healthcare for diabetes prediction. In *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, ICBET' 19, page 253–259, New York, NY, USA. Association for Computing Machinery.