Data Mining and Risk Analysis Supporting Decision in Brazilian Public Health Systems

Raimundo Valter, Silas Santiago,
Ronaldo Ramos, Mauro Oliveira
Computer Science Dept., IFCE
Fortaleza, Brazil
Email: {valter.costa, silas.santiago, ronaldo, mauro}@ifce.edu.br

Luis Odorico M. Andrade,
Ivana Cristina de H. C. Barreto
Public Health Dept., FIOCRUZ
Fortaleza, Brazil
Email: {odorico,ivana_barreto}@ufc.br

***Abstract*—Health data monitoring is a key activity to reduce maternal, neonatal and infant mortality rates. Data available in Brazilian health databases points that It is possible to predict death risk in early stages of gestation and newborn development. In this research, we consider the information availability still in gestational period to propose different death risk prediction models for this public of interest. We also detail the data mining process to apply machine learning-based techniques in death risk classification for maternal, neonatal and infant patients. We present an experiment pipeline to estimate average performance and evaluated machine learning models with different features combinations. Additionally, is shown a web service which provides multiple predictive models by information availability. Results shows Random Forest obtaining better performance when compared to the other machine learning methods.**

***Index Terms*—Brazilian health data, data mining, information availability.**

I. Introduction

Historically, data analysis was always a guide in the decision-making process. Modern computing techniques and the vast amount of information available in Brazil by public transparency points to new opportunities. When it comes to linked data, though, there are many challenges. Despite the difficulties, researches in this area continue to demonstrate that it is possible to immediately associate data and extract solutions.

The World Health Organization (WHO) [1], [2] reports many of the maternal and infant occurred deaths, are due to gestation or parturition complications and can be avoided by performing simple actions. For this purpose, population health parameters monitoring is a key activity to reduce maternal (gestation and puerperium), neonatal and infant mortality rates.

When considering this context, collecting some data during each gestation stage period allows generating relevant information to identify death risk for mothers and babes. Answering this requisite, the web tool Intelligent Governance in Health Systems (GISSA) was proposed. This system supports governance in public health care. GISSA consists of a set of components which allow data collection, integration and visualization to support the decision-making process.

Researches [3]–[5] have already demonstrated the correlation in these data identifying the feasibility of inferring the death risk in this specific public. Nevertheless, there are distinct moments in which the data is available, suggesting the death risk identification may precede even the birth, in order to guarantee a certain confidence degree. Figure 1 outlines three periods of interest in case of mothers (pregnancy up to puerperium) and babes (neonatal and infant).

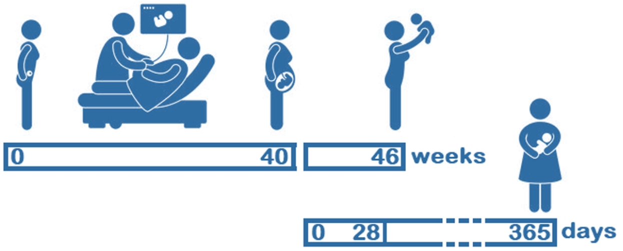

Fig. 1. Periods: gestational + puerperium (0 to 46 weeks), neonatal (0 to 28 days), and infant (0 to 365 days).

Using this data collected at every single moment allows the proposal of different pattern recognition models to the death risk identification for maternal, neonatal and infant patients. These predictive models could be applied to different stages of gestation and baby development. This individualized or joined information may be used to support public health strategies.

This paper details the data mining process in GISSA datasets to evaluate the application of supervised machine learning methods for neonatal, infant and maternal mortality for death risk prediction. The main contribution of this work includes a proposal of a web service which provides multiple predictive models ordered by information availability. A Proof of Concept (PoC) is also presented to demonstrate its use.

The remainder of this paper is organized as follow. Section II overviews related works about infant and pregnancy mortality risk prediction with machine learning. Section III details data mining process used to build and evaluate a couple of machine learning-based death risk prediction models. Section IV presents and discusses performance results. Section V ends with some conclusions and future works.

II. Related Work

In [3], it was experimented to apply Fuzzy logic to death prediction of children group in the neonatal period. In this study, it was identified a set of characteristics of interest: newborn birth weight, gestational age at parturition, Apgar score, and previous report of stillbirth. These features showed to be enough for the fuzzy inference. From the 24 rules identified by specialists, with it is possible to predict neonatal death with an accuracy of 90.0%. This study points for the model applicability from the child's birth, since 3/4 of these attributes are measured only in birth, been not possible using them to predict mortality conditions in early stages.

Another research, performed in [4], presents a simple and practical approach to identify whether the infant mortality coefficient of a given city will be above or below the Brazilian national mean rate. By the use of regression trees model, it is sufficient to observe the total of prenatal medical appointments and mother's educational level hitting 65.0% of cases.

In [6], authors presents and evaluate the Intelligent Health Analysis System (LAIS), to support decision-making in preventive actions involving pregnant mothers and newborns. This system uses data mining techniques to generate death risk alerts using probability-based methods for training and evaluation of predictive models. The authors applied data from the Mortality Information System (SIM) and Live Birth Information System (SINASC) databases available on the DATA-SUS portal. Results showed that probabilistic algorithm Naive Bayes performed better when compared to others machine learning techniques. The obtained accuracy and Area Under the Receiver Operating Characteristic Curve (AUROCC) were 98.2% and 92.1%, respectively.

III. Methodology

Since the project workflow focuses on the data mining process from data collection to the deployment, it is applied Cross Industry Standard Process for Data Mining (CRISP-DM) [7] as the predominant methodology.

A. Data Mining Steps in Gissa

In order to better represent the methodology applied, CRISP-DM was simplified in four main steps. Figure 2 summarizes macro activities that guide data mining based approach in this paper.

B. Business Understanding

This step was about the business domain understanding of GISSA and includes the comprehension of the entities, relationships, and fields in the databases. As an artifact, data dictionary is produced which describes tables of SINASC and SIM databases.

C. Data Preparation

Raw data in GISSA's scenario is stored in relational databases. This enables applications to efficiently store and query with Structure Query Language (SQL) large amount of data (about 1.5M samples) [8]. Considering the problem definition, the dataset and its description was built based on SIM and SINASC tables through data preparation phase summarized in a couple of steps:

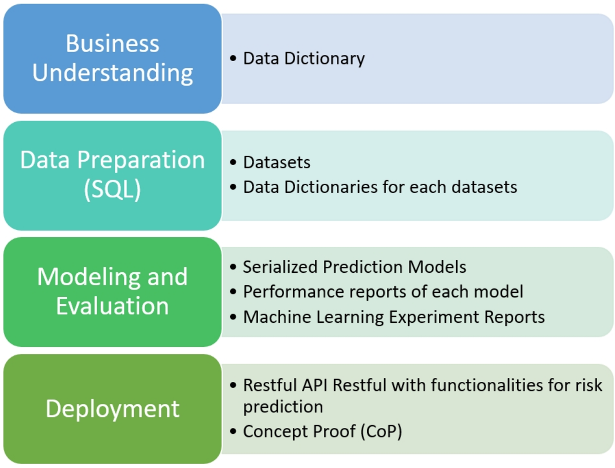
Business Understanding
• Data Dictionary
Data Preparation (SQL)
• Datasets
• Data Dictionaries for each datasets
Modeling and Evaluation
• Serialized Prediction Models
• Performance reports of each model
• Machine Learning Experiment Reports
Deployment
• Restful API Restful with functionalities for risk prediction
• Concept Proof (CoP)

Fig. 2. Data mining steps in GISSA.

1) **Select Data**: selection of columns and rows of interest in SIM and SINASC tables;
2) **Integrate Data**: tables union is done defining data classification as well (samples that incur in death or not). Some fields appears with missed values due to lack of information coming from different tables;
3) **Clean Data**: filling default missed values and replacement of inconsistent values in the table resulted by integration step;
4) **Construct Data**: some features are extracted based on each problem definition;
5) **Format Data**: completely filled registers (without ignored informations) are selected randomly and recorded in Comma-Separated Values (CSV) format.

The Figure 3 represents the data preparation process, which involves apply SQL scripts resulting in each dataset.

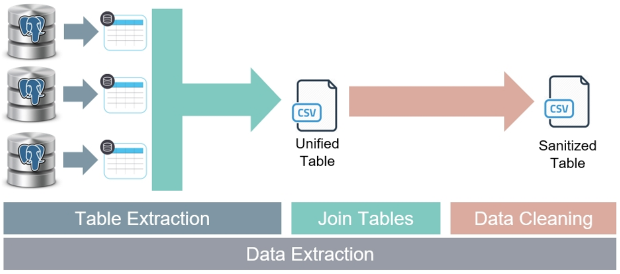

Fig. 3. Overview for data preparation process.

After data preparation process, datasets classification remains in two groups (dead or alive) for binary classification as in Table I. Maternal dataset is composed by women in pregnancy or puerperium stages. Neonatal dataset is composed by newborns children, from 0 up to 28 days of life. Infant dataset is composed by children from 0 up to 365 days of life.

After this phase each dataset is standardized with StandardScaler class, available at Scikit-Learn library [9]. This

TABLE I
DATASETS COMPOSITION

Dataset	dead	alive	total
maternal dataset	508	2531	3039
neonatal dataset	657	682	1339
infant dataset	911	952	1863

operation results in zero mean and scaling data to unit variance considering each feature separately in all samples selected. Each value is scaled by the expression: $x_{scaled} = (x - \mu)/\sigma$, where μ and σ represent, respectively, the mean and standard deviation for a given feature in dataset.

Exploratory data analysis was performed to verify the generated datasets. This allows verify predictor variables quality by graphics visualization and statistical measurements. All these actions aims to prevent biased and overfitting models.

D. Modeling

Some tasks are performed in order to modeling and assess applicability: (1) data loading and preprocessing; (2) exploratory data analysis; (3) hyperparameters optimization; (4) cross-validation executions. After these steps, models are ranking by AUROCC and accuracy. Figure 4 shows the sequence of steps to guide modeling and deployment in API restful used by by GISSA portal. This pipeline is applied to the three datasets considering each group of features selected.

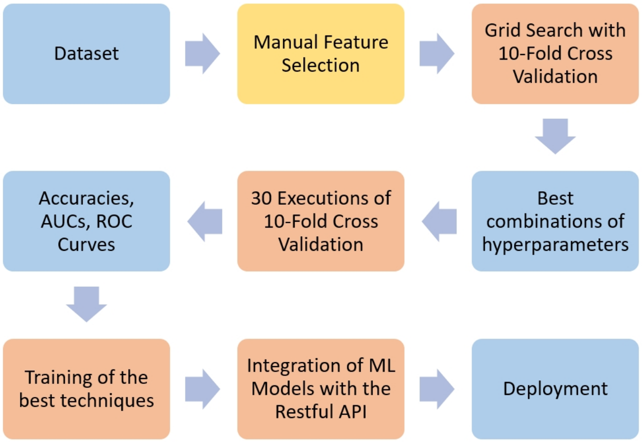
Dataset
Manual Feature Selection
Grid Search with 10-Fold Cross Validation
Best combinations of hyperparameters
30 Executions of 10-Fold Cross Validation
Accuracies, AUCs, ROC Curves
Training of the best techniques
Integration of ML Models with the Restful API
Deployment

Fig. 4. Experiment overview for model selection and evaluation.

For each dataset, the attributes are listed in order of information availability. This feature arrangement allows the definition and evaluation of multiple predictive models, depending on the available information. Besides that, this strategy also enables the prediction in different periods of interest, once multiple models with progressive feature numbers are proposed.

The features in maternal dataset are (F_1) Birthplace, (F_2) Education level (mother), (F_3) Child's race, (F_4) Child's gender, (F_5) Number of healthy parturition, (F_6) Gestational age (in weeks), (F_7) month starting prenatal, (F_8) Child positioning for parturition, (F_9) Parturition type, (F_{10}) Assisted parturition, (F_{11}) Induced parturition, (F_{12}) Cesarean occurrence before parturition, (F_{13}) Birth indicative, (F_{14}) Robson classification [10] for pregnancy, (F_{15}) Apgar 5 minutes for child at birth, (F_{16}) Age of child at death, (F_{17}) Death occurred in relation to parturition and (F_{18}) Death indicative of child.

The features in neonatal and infant datasets are (F_1) Age of father at birth, (F_2) Age of mother at birth, (F_3) Level of education (mother), (F_4) Marital status of mother, (F_5) Number of prenatal consultations, (F_6) Start month of prenatal consultations, (F_7) Start week of prenatal consultations, (F_8) mother's Brazilian Code of Occupation (BCO), (F_9) Number of previous pregnancies, (F_{10}) Number of stillbirths, (F_{11}) Number of live births, (F_{12}) Number of cesarean parturition, (F_{13}) Number of healthy births, (F_{14}) race of mother , (F_{15}) gender of child, (F_{16}) Pregnancy type, (F_{17}) Birth occurrence place, (F_{18}) Robson classification, (F_{19}) Assisted parturition code, (F_{20}) Cesarean occurrence before parturition begins, (F_{21}) Cesarean status before parturition begins, (F_{22}) Birthplace, (F_{23}) Apgar 1 minute index for child at birth, (F_{24}) Apgar 5 minutes index for child at birth, (F_{25}) Weight of child at birth, (F_{26}) Race of child and (F_{27}) Malformed occurrence status.

E. Evaluation

The performance of supervised classifiers Naive Bayes [11] (NB), Decision Tree (DT) [12] and Random Forest (RF) [13] was measured and evaluated for the binary classification task.

The application of the Grid Search optimization strategy combined with the K-Fold Cross Validation (CV) technique makes it possible to obtain different model performance estimates for each hyperparameters combination. From these results, we can choose the most appropriated model (with the lowest CV error). In addition, the use of CV maximizes the confidence of the selected hyperparameters values ensuring a better generalization (reducing overfitting).

The hyperparameters adjustment in supervised algorithms Decision Tree and Random Forest were performed by the following. For Random Forest, the parameters n_estimators, criterion and max_depth of the RandomForest Classifier class available in the Scikit-learn library were considered. Table II shows the parameters and tested values for this classifier.

TABLE II
EVALUATED PARAMETERS FOR RANDOM FOREST

Parameters	Description	Tested Values
n_estimators	Number of trees in the forest	10, 50, 100
max_depth	Maximum depth of the tree	5, 10, 15, 20
criterion	Function to measure the quality of a split	”gini”, ”entropy”

For the Decision Tree, criterion and splitter parameters were considered. The Table III presents the tested values.

Since Gaussian Naive Bayes computes *a priori* and *a posteriori* probabilities from datasets, there are no parameters to be tunned.

TABLE III
EVALUATED PARAMETERS FOR DECISION TREE

Parameters	Description	Tested Values
criterion	Function to measure the quality of a split	"gini", "entropy"
splitter	strategy used to choose the split at each node	"best","random"

Aim to obtain the best combination of parameters, the Grid Search technique combined to K-Fold Cross Validation was performed with $k = 10$. The optimal values for Random Forest were $criterion = "gini"$, $max\_depth = 10$ and $n\_estimators = 100$. For the Decision Tree technique were found $criterion = "gini"$, and $splitter = "best"$.

1) Cross-Validation Experiment: For each supervised estimator, the experiment pipeline was run 30 times, for the estimation of a confidence interval and average performance benchmarks. Algorithm 1 details the experiment process. For a given dataset D, a group of supervised machine learning techniques T and features set A, the experiment firstly generate a set of features combinations C (TOP 01, TOP 02, $\dots$, TOP M) and evaluate each technique for a different features subset by cross-validation.

Algorithm 1 Experiment pseudocode

```
$D \leftarrow GetDataset(\ )$
$A \leftarrow \{a_1, a_2, \dots, a_m\}$
$T \leftarrow \{t_1, t_2, \dots, t_k\}$
$C \leftarrow FeaturesCombination(A)$
foreach $t \in T$ do
    $t_{hyperparameters} = GridSearch(t, D, A)$
end foreach
foreach $round \in 30\ rounds$ do
    foreach $c \in C$ do
        foreach $t \in T$ do
            $S \leftarrow Subset(D, c)$
            $S \leftarrow FeatureStandardization(S)$
            $ACC_{cv}, AUROCC_{cv} \leftarrow CrossVal(S, folds = 10)$
        end foreach
    end foreach
end foreach
foreach $c \in C$ do
    foreach $t \in T$ do
        $results[c][t] = ComputeMetrics()$
    end foreach
end foreach
$S \leftarrow BestCombinations(results)$
```

F. Deployment

From the evaluation of experiments described previously, it is possible to obtain the predictive models with better accuracy and AUROCC results for each attributes combination. The selected predictive models for each addressed classification scenario (maternal, neonatal and infant) were serialized and made available in a restful API. This software modularization allows simple integration with any system, either web or mobile. The intelligence module uses supervised machine learning classifiers to compute predictions and probabilities. A total of 27 predictive models is generated for neonatal mortality, 27 models for infant mortality and 18 models for maternal mortality.

For each scenario, the model represents a classifier which receives a features vector of a given size. For example, considering neonatal mortality risk, model TOP04 is trained with the Gaussian Naive Bayes algorithm and receives as input a vector with four attributes. In this context, GISSA web system works as the PoC, consuming services provided by an restful API, in order to demonstrate the proposed models. Figure 5 illustrates the architecture.

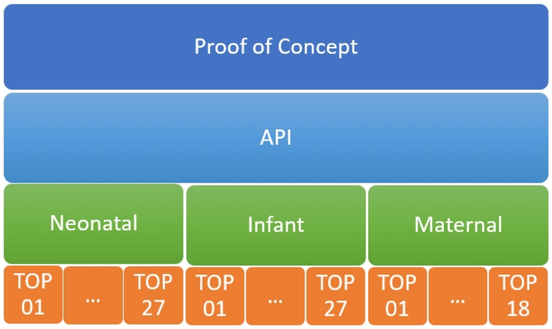

Fig. 5. PoC and API architecture.

Still considering TOP04 model, GISSA application must informs the scenario and a feature vector as an POST request according to the format.

```
POST http://<server>:5001/predict
{
    "data": "[21.0, 19.0, 4.0, 2.0]",
    "model": "MMInfantil",
}
```

API executes prediction for the selected model, returning the class value (0 for alive or 1 for death prediction) and the death probability as a POST response.

```
Response: [{'class': 1, 'prob': 0.79}]
```

IV. RESULTS

A serie of experiments was performed and the best predictive model algorithms considering performance results were chose. The average values for accuracy and AUROCC are presented, according to the experiment procedures detailed in section III. All models and its results are presented in Tables IV, V and VI.

Receiver Operating Characteristic (ROC) curve is represented in the cartesian plane, where the Y axis represents the sensitivity and the X axis represents 1 - specificity. Sensitivity refers to the likelihood of a test being positive given that the

TABLE IV
NEONATAL MORTALITY EXPERIMENTS

Features Set	Classifier	Mean AUROCC	Mean ACC
TOP01	GaussianNB	0.5180	54.07%
TOP02	GaussianNB	0.5407	55.21%
TOP03	GaussianNB	0.5554	57.83%
⋮	⋮	⋮	⋮
TOP15	RandomForest	0.7391	81.68%
TOP16	RandomForest	0.7379	81.68%
TOP17	RandomForest	0.7454	82.02%
TOP18	RandomForest	0.7459	81.86%
TOP19	RandomForest	0.7448	81.87%
TOP20	RandomForest	0.7504	82.73%
TOP21	RandomForest	0.7517	82.73%
TOP22	RandomForest	0.7511	82.79%
TOP23	RandomForest	0.8155	88.27%
TOP24	RandomForest	0.8261	89.75%
TOP25	RandomForest	0.8394	90.82%
TOP26	**RandomForest**	**0.8876**	**93.90%**
TOP27	RandomForest	0.8872	93.94%

individual has died. Specificity refers to the likelihood of the test being negative, once the individual has not died [14]. For the Figures 6, 7 and 8 blue dots describes the mean ROC graph that represents all 30 experiments randomly initialized of the considered model, the gray shadow (when observable) in background is a composition of all results separately.

Overall accuracy is a measure that provides the percentage of correctly categorized examples. It concerns the ratio between the accounting of correctly classified examples and the total of evaluated examples. This metric is accepted for evaluation and describes the accuracy of classification entirely [15].

AUROCC represents the overall performance of an estimator, once this metric considers all computed values of sensitivity and specificity. The more the estimator ability to discriminate against individuals with and without risk of death, more the curve will approximate to the upper left corner and AUROCC will approximate to 1 [14].

The predominance of the Random Forest algorithm as the best technique for the evaluated datasets is notorious. For the neonatal and infant mortality datasets, the Random Forest algorithm obtained better performance for 24 of the 27 predictive models evaluated. For the maternal mortality dataset, this algorithm presented better accuracy and AUROCC for all models.

For the neonatal dataset, the combination that scored highest AUROCC and accuracy was TOP26, with 0.8876 and 93.90%, respectively. The area under the ROC curve for this combination is shown in Figure 6. This combination concerns with a Random Forest model with 26 predictive attributes.

For the infant dataset, the best combination was also the TOP26, with 0.9909 and 99.73%, respectively. This similarity is associated with the composition of the neonatal and infant datasets, given that both differ in the way how class attribute is determined for each instance. The ROC curve for this combination is shown in Figure 7.

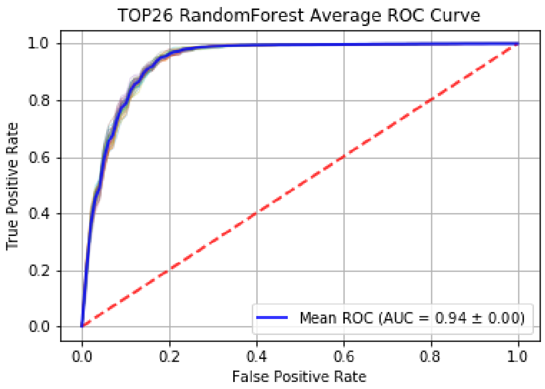

Fig. 6. ROC curve for neonatal death risk.

TABLE V
INFANT MORTALITY EXPERIMENTS

Features Set	Classifier	Mean AUROCC	Mean ACC
TOP01	GaussianNB	0.5277	54.34%
TOP02	GaussianNB	0.5495	56.15%
TOP03	GaussianNB	0.5760	60.70%
⋮	⋮	⋮	⋮
TOP15	RandomForest	0.7477	82.33%
TOP16	RandomForest	0.7480	82.37%
TOP17	RandomForest	0.7512	82.69%
TOP18	RandomForest	0.7533	82.81%
TOP19	RandomForest	0.7518	82.70%
TOP20	RandomForest	0.7903	86.32%
TOP21	RandomForest	0.7893	86.25%
TOP22	RandomForest	0.7893	86.20%
TOP23	RandomForest	0.8452	91.28%
TOP24	RandomForest	0.8521	92.00%
TOP25	RandomForest	0.8744	93.15%
TOP26	**RandomForest**	**0.9909**	**99.73%**
TOP27	RandomForest	0.9906	99.82%

Lastly, for the maternal dataset, TOP15 the combination was the one that obtained the highest value for accuracy and AUROCC, with 0.9163 and 97.50%, respectively. The ROC curve for this combination is shown in Figure 8.

V. CONCLUSION

This paper evaluated three problem scenarios for death prediction to support decision-making in health management. From the data mining process in GISSA portal, it was possible to build and evaluate a set of machine learning models trained with neonatal, infant and maternal data with different feature combinations.

In this approach, we consider the information availability as a guideline to generate, evaluate and select the predictive models. Each model represents the best algorithm for a feature

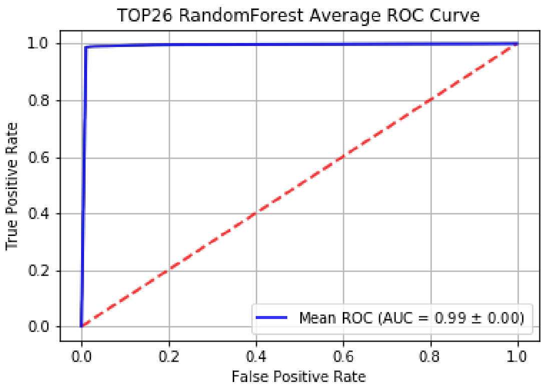

Fig. 7. ROC curve for infant death risk.

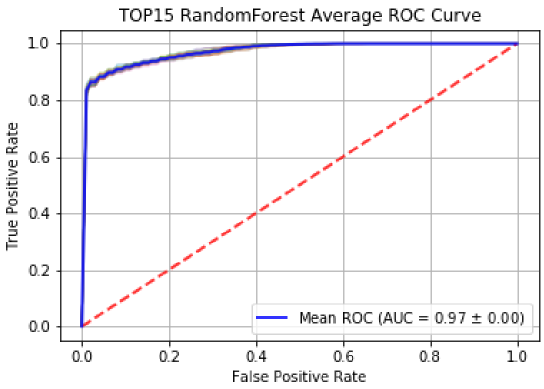

Fig. 8. ROC curve for maternal death risk.

TABLE VI
MATERNAL MORTALITY EXPERIMENTS

Features Set	Classifier	Mean AUROCC	Mean ACC
TOP01	RandomForest	0.6120	61.20%
TOP02	RandomForest	0.6913	71.64%
TOP03	RandomForest	0.8277	92.81%
⋮	⋮	⋮	⋮
TOP10	RandomForest	0.8840	95.49%
TOP11	RandomForest	0.8892	95.77%
TOP12	RandomForest	0.8946	96.14%
TOP13	RandomForest	0.8957	96.26%
TOP14	RandomForest	0.9060	97.11%
TOP15	**RandomForest**	**0.9163**	**97.50%**
TOP16	RandomForest	0.9147	97.39%
TOP17	RandomForest	0.9133	97.38%
TOP18	RandomForest	0.9143	97.41%

combination. The Random forest estimator was the predominant method in feature combinations for the three scenarios, which indicates its generalization capability for health data. To demonstrate the utility of our approach, a microservice to serve all these models for each scenario was built. A PoC was also implemented to demonstrate the use of the restful API.

Future works includes expand restful API system adding new predict models (services). Intends evaluate other supervised methods for described scenarios or even apply semi-supervised approaches to deal with labeled and unlabeled datasets.

ACKNOWLEDGEMENT

Our sincere acknowledgements to the Cearense Foundation for Scientific and Technological Development (FUNCAP) for funding this research. We also thank the AVICENA ™ team for provide access to data and support from IT technicians and health care specialists.

REFERENCES

[1] W. H. Organization. (2018) Newborns: reducing mortality. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality

[2] ——. (2018) Maternal mortality. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[3] L. F. C. Nascimento, P. M. S. R. Rizol, and L. B. Abiuzi, “Establishing the risk of neonatal mortality using a fuzzy predictive model,” *Cadernos de Saúde Pública*, vol. 25, pp. 2043 – 2052, 09 2009.

[4] A. D. P. Chiavegatto Filho, “Uso de big data em saúde no brasil: perspectivas para um futuro próximo,” *Epidemiologia e Serviços de Saúde*, vol. 24, pp. 325 – 332, 06 2015.

[5] C. L. da Silva, “Lais, uma solução baseada em classificadores para geração de alertas em sistema de saúde,” Ph.D. dissertation, Universidade Estadual do Ceará, 2017.

[6] R. Ramos, C. Silva, M. W. L. Moreira, J. J. P. C. Rodrigues, M. Oliveira, and O. Monteiro, “Using predictive classifiers to prevent infant mortality in the brazilian northeast,” in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Oct 2017, pp. 1–6.

[7] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Citeseer, 2000, pp. 29–39.

[8] J. Grus, *Data science from scratch: first principles with python*. ” O’Reilly Media, Inc.”, 2015.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] J. P. Vogel, A. P. Betrán, N. Vindevoghel, J. P. Souza, M. R. Torloni, J. Zhang, Ö. Tunçalp, R. Mori, N. Morisaki, E. Ortiz-Panozo *et al.*, “Use of the robson classification to assess caesarean section trends in 21 countries: a secondary analysis of two who multicountry surveys,” *The Lancet Global health*, vol. 3, no. 5, pp. e260–e270, 2015.

[11] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[13] Y. Liu and H. Wu, “Water bloom warning model based on random forest,” in *Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2017 International Conference on*. IEEE, 2017, pp. 45–48.

[14] B. Lopes, I. C. d. O. Ramos, G. Ribeiro, R. Correa, B. d. F. Valbon, A. C. d. Luz, M. Salomão, J. M. Lyra, and R. Ambrósio Junior, “Bioestatísticas: conceitos fundamentais e aplicações práticas,” *Rev Bras Oftalmol*, vol. 73, no. 1, pp. 16–22, 2014.

[15] A. Dainotti, A. Pescape, and K. C. Claffy, “Issues and future directions in traffic classification,” *IEEE network*, vol. 26, no. 1, pp. 35–40, 2012.