Treating Colon Cancer Survivability Prediction as a Classification Problem

Ana Silva, Tiago Oliveira, José Neves, and Paulo Novais
Algoritmi Centre/Department of Informatics, University of Minho, Braga, Portugal
a55865@alunos.uminho.pt, {toliveira,jneves,pjon}@di.uminho.pt

KEYWORD

colon cancer; prediction; machine learning

ABSTRACT

This work presents a survivability prediction model for colon cancer developed with machine learning techniques. Survivability was viewed as a classification task where it was necessary to determine if a patient would survive each of the five years following treatment. The model was based on the SEER dataset which, after preprocessing, consisted of 38,592 records of colon cancer patients. Six features were extracted from a feature selection process in order to construct the model. This model was compared with another one with 18 features indicated by a physician. The results show that the performance of the six-feature model is close to that of the model using 18 features, which indicates that the first may be a good compromise between usability and performance.

1. Introduction

Colorectal cancer is one of the most common types of cancer of the digestive system. It is the third most common cancer overall with an incidence rate of 9.7% and the fourth most deadly with a mortality rate of 6.41% (Ferlay et al., 2012). This is a pathology that affects the walls of the colon and rectum and consists in the abnormal growth of the cells lining these portions of the digestive tract (Vachani and Prechtel-Dunphy, 2015). The term colorectal is used to represent two different pathologies, colon cancer and rectal cancer. Although these two pathologies have aspects in common, they are different diseases and are characterized by different dynamics and molecular pathways (Yamauchi et al., 2012). Colorectal cancer affects mostly the elderly and among its risk factors are smoking, inherited gene mutations, and personal family history of colorectal cancer. Most colorectal cancers develop in the colon, which consists of the cecum, ascending colon, tansverse colon, descending colon, and sigmoid. This work will focus on this colorectal cancer variant.

In terms of treatment, surgical resection is the preferred choice when it comes to colon cancer. As it is an aggressive treatment, most of the times followed by chemotherapy, it is often unclear whether patients will be able to endure it or not. Therefore, estimating the survivability of colon cancer patients is an important clinical decision making element for health care professionals, one that may help them to decide if a patient will need palliative care or to inform patients more accurately. However, it is not an easy task and even seasoned oncologists have trouble in making such predictions. As such, the objectives of this work are:

- To develop an individualized survivability prediction model for colon cancer patients in years 1, 2, 3, 4, and 5 after treatment;
- To determine the ideal number of features to make a prediction and to operationalize the prediction model in an application;
- To determine which features are important for colon cancer survivability prediction;

©Ediciones Universidad de Salamanca / cc by-nc-nd


The starting point was the the Surveillance, Epidemiology, and End Results (SEER) program (National Cancer Institute, 2015), a large cancer registry from the United States (US) with data from 1973 to 2012, featuring a total of 8,689,771 cancer cases. After the extraction of data of colon cancer patients in several pre-processing steps, different machine learning strategies were applied in order to produce a survival prediction model in the form of several classifiers.

The structure of this paper is as follows. Section 2 introduces previous works in colon cancer survivability prediction. Section 3 explains the prediction system under development with the specification of the type of inputs it should receive and the outputs it should produce. It also describes the steps and machine learning methods used to develop the prediction model. The corresponding experimental results are disclosed and discussed in Section 4. Finally, Section 5 provides concluding remarks about the work done so far and future work considerations.

2. Related Work

Most of the existing approaches for colon cancer survivability prediction are based on the SEER data. An example is the web-based calculator[1] developed in (Bush and Michaelson, 2009) whose underlying prediction model is the Nodes + Prognostic Factors (NAP), based on the number of positive lymphatic nodes combined with other prognostic features. The model has an underlying biological motivation, reflected in the use of the probability of a cancerous cell invading healthy tissues to formulate equations for cancer lethality, combined with other prognostic features estimated by means of simulation of several statistical tests. The model requires inputs for 9 features and provides a prediction of the mortality risk over the period of 15 years.

Another SEER-based approach is the one followed in (Chang et al., 2009), also made available in the form of a web application [2]. The prediction model has 5 input features, derived through a Cox regression analysis to evaluate simultaneous effects of multiple variables on survivability. This resulted in adjusted survival functions stratified by 5 features. The conditional survival probabilities for a period of 10 years produced by the model are calculated on the basis of the adjusted survival functions for the features, controlled for the influence of other covariates in the final model.

A similar approach was followed in (Weiser et al., 2011), in which a survival prediction model for a period of 5 years was developed based on multi-variable regression, with Cox proportional hazards modelling, using 7 prognostic features [3]. All the features were chosen *a priori*, on the basis of their well established independent association with overall survival and their availability in the SEER data.

In (Snow et al., 2001) an artificial neural network model and a regression-based model were developed to predict patient survival status 5 years after treatment. The models have 12 input features and were based on data from the National Cancer Data Base (NCDB), a cancer registry in the United Kingdom. This work had a strong machine learning component and is among the first to apply methods from this field of computer science to colon cancer survival prediction. Another example is the work in (Al-Bahrani et al., 2013), in which a 5-year survival prediction model was developed using ensemble machine learning with supervised classification. The number of selected features for prediction in this work was 13 and the resulting model achieved an overall high performance in terms of precision, accuracy, and receiver operating characteristic (ROC).

The work developed herein distances itself from the works in (Bush and Michaelson, 2009; Chang et al., 2009; Weiser et al., 2011) by treating survival prediction as a classification problem and applying varied machine learning methods to obtain a model capable of individualized survival prediction. In this regard, it is influenced by the methodology followed in (Al-Bahrani et al., 2013), whose work will serve as a reference for direct

[1] Application available at `http://www.lifemath.net/cancer/coloncancer/outcome/index.php`.
[2] Application available at `http://www3.mdanderson.org/coloncalculator`.
[3] Application available at `http://nomograms.mskcc.org/Colorectal/OverallSurvivalProbability.aspx`

©Ediciones Universidad de Salamanca / cc by-nc-nd


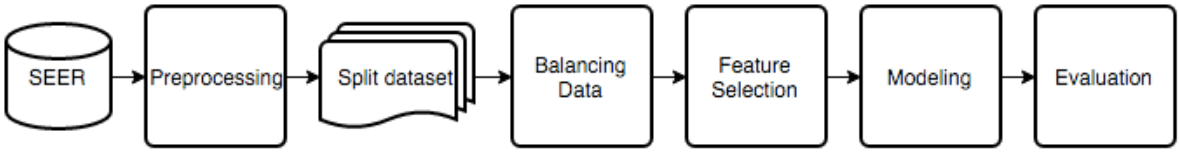

Figure 1: Workflow for the development of the prediction model.

comparison. At the same time, this work aims to produce 5-year survival predictions using fewer features than the existing approaches, which may be the deciding factor for the adoption of a clinical decision support application.

3. Development of the Prediction Model

The colon cancer survival prediction system should have the ability to accept a determined number of inputs for selected prediction features and produce an output stating whether the patient will survive each of the five years following treatment, along with a confidence value for the prediction. The survival prediction was handled as a classification problem, so that five classification models for each year were developed. In order to provide a prediction for each year with a single interaction, the models were posteriorly combined, in a programmatic manner.

The development of these prediction models involved several phases, from the preprocessing of SEER data to the selection of the best model. All of them are depicted in the workflow of Figure 1 and each one is described in the ensuing sections.

The RapidMiner software[4] was chosen to develop the prediction model. It has a workflow-based interface that offers an intuitive application programming interface (API).

3.1 Preprocessing, Split Dataset, and Balancing Data

In order to load the data provided by SEER to RapidMiner, the data in raw format had to be converted into *csv* format, through a developed script. The data of colorectal cancer from SEER contained 515,791 records and 146 attributes, with only some of them being applicable to a limited period within the time of data collection. After the preprocessing phase and selecting the colon cancer patients, the data was reduced to 38,592 records.

In order to minimize the occurrence of missing data due to the applicability of the attributes, during the Preprocessing phase was defined a period of interest, from 2004 onwards. Additionally, attributes that are not applicable to this type of cancer (e.g., the human epidermal growth factor receptor 2 result is an indicator used in breast cancer only(Wolff et al., 2007)), empty attributes, and attributes that are not directly related with the vital status of the patient were removed (e.g. the number identifying the registry of the patient). Only patients with age greater than or equal to 18 years old were selected for further processing – because of colon cancer is not common in young people –, and the missing values were replaced with the *unknown* code. Patients who were alive at the end of the data collection whose survival time had not yet reached 60 months (five years), the maximum period for which the model under development is supposed to predict survivability, and those who passed away of causes unrelated to the cancer were sampled out from the training set as their inclusion was considered to be unsuited to the problem at hand. The numeric attributes were converted to nominal (e.g. sex) and the binary classes (*survived* and *not survived*) were derived for the target labels 1-, 2-, 3-, 4- and 5-year

[4]Software available at https://rapidminer.com/.

©Ediciones Universidad de Salamanca / cc by-nc-nd

Table 1: Class distribution for each target label in the sub-datasets.

	Target Labels				
	1 Year	2 Year	3 Year	4 Year	5 Year
Not Survived	24.51%	32.60%	36.96%	39.35%	41.07%
Survived	75.49%	67.40%	63.04%	60.65%	58.93%

Table 2: Attributes selected in the Feature Selection process.

Attribute	Description
Age recode with < 1 year old	Age groupings based on age at diagnosis (single-year ages) of patients (< 1 year, 1-4 years, 5-9 years, ..., 85+ years)
CS Site-Specific Factor 1	The interpretation of the highest Carcinoembryonic Antigen (CEA)[5] test results
CS Site-Specific Factor 2	The clinical assessment of regional lymph nodes
Derived AJCC Stage Group	The grouping of the TNM information combined
Primary Site	Identification of the site in which the primary tumor originated
Regional Nodes Examined	The total number of regional lymph nodes that were removed and examined by the pathologist

survival. Finally, based on existing attributes new ones were calculated, such as the ratio of positive nodes over the total examined nodes, the number of regional lymph negative nodes and the relapse of the patients for colon cancer. The attributes were diminished to 61 after the Preprocessing, including the added attributes and the target labels.

Data was split into five sub-datasets during the Split Dataset phase by target label, according to the survival year. Table 1 shows the class distribution of each sub-dataset.

Observing Table 1 is seen that the classes are not equally represented. Several studies (Chawla, 2005; Leon and Jalao, 2014) show how important the problem of using imbalanced datasets is, from both the algorithmic and performance perspectives. An overview of classification algorithms for the resolution of this kind of problem (Ganganwar, 2012) concluded that hybrid sampling techniques, i.e., combining over-sampling of the minority class with under-sampling of the majority class, can perform better than just oversampling or undersampling. As such, in the Balancing Data phase, hybrid sampling was applied in order to generate balanced sub-datasets, as described in (Ganganwar, 2012). It resulted in five sub-datsets with 38,592 records each.

3.2 Feature Selection

The Feature Selection phase was an essential phase where the most influential features on the survival of colon cancer patients were determined using the Optimize Selection operator (RapidMiner, 2016c) of RapidMiner. It implements a deterministic and optimized selection process with decision trees and *forward selection*. The process was applied to each sub-dataset for the target label. Only the selected features in common to all the sub-datasets were used to construct the prediction models. Table 2 shows the selected features and their meaning.

The selected features, a total of 6, were compared with a set of 18 features (shown in Table 3) indicated by a specialist physician on colon cancer. Several prediction models were constructed with these two sets of features,

©Ediciones Universidad de Salamanca / cc by-nc-nd


Table 3: Attributes selected by a specialist physician on colon cancer.

Attribute	Description
Age at Diagnosis	The age of the patient at diagnosis
CS Extension	Extension of the tumor
CS Site-Specific Factor 8	The perineural Invasion
CS Tumor Size	The size of the tumor
Derived AJCC T, N and M	The AJCC T, N and M stage (6th ed.)
Grade	Grading and differentiation codes
Histologic Type	The microscopic composition of cells and/or tissue for a specific primary
Laterality	The side of a paired organ or side of the body on which the reportable tumor originated
Primary Site	*
Race Recode (White, Black, Other)	Race recode based on the race variables
Regional Nodes Examined	*
Regional Nodes Positive	The exact number of regional lymph nodes examined by the pathologist that were found to contain metastases
Regional Nodes Negative	(Regional nodes examined - Regional nodes positive)
Regional Nodes Ratio	(Regional nodes negative over Regional nodes examined)
Relapse	The relapse of the patients for colon cancer
Sex	The sex of the patient at diagnosis

* Described in Table 2.

mapping the attributes in the sub-datasets and later used to generate and evaluate the prediction models.

3.3 Modeling and Evaluation

The classification strategies used in the Modeling phase consisted of ensemble methods. The classification schemes applied were meta-classifiers. This type of classifier is used to boost basic classifiers and improve their performance. All the possible combinations of the classifiers were explored, according to the algorithms and type of attributes allowed. The tested meta-classifiers were:

- **Bagging** (Breiman, 1996): Also called bootstrap aggregating. It splits the data into *m* different training sets on which *m* classifiers are trained. The final prediction results from the equal voting of each generated model on the correct result. Bagging is used to improve stability and classification accuracy, reduce variance and avoid overfitting.
- **AdaBoost** (Freund and Schapire, 1997): This meta-classifier calls a new weak classifier at each iteration. A weight distribution which indicates the weight of examples in the classification is updated. It focuses on the examples that have been misclassified so far in order to adjust subsequent classifiers and reduce relative error.

©Ediciones Universidad de Salamanca / cc by-nc-nd


- **Bayesian Boosting** (RapidMiner, 2016a): A new classification model is produced at each iteration and the training set is reweighed so that previously discovered patterns are sampled out. The inner classifier is sequentially applied and the resulting models are later combined into a single model. The boosting operation is conducted based on probability estimates. It is particularly useful for discovering hidden groups in the data.
- **Stacking** (Džeroski and Ženko, 2004): This meta-classifier is used to combine base classifiers of different types. Each base classifier generates a model using the training set, then a meta-learner integrates the independently learned base classifier models into a high level classifier by re-learning a meta-level training set. This meta-level training set is obtained by using the predictions of base classifiers in the validation dataset as attribute values and the true class as the target.
- **Voting** (Kittler, 1998): Each inner classifier of the meta-classifier receives the training set and generates a classification model. The prediction of an unknown example results from the majority voting of the derived classification models.

Since survivavility prediction is being handled as a classification problem, a group of basic classifiers were selected to be used in ensembles with the above-described meta-classifiers. The group includes some of the most widely used learners (RapidMiner, 2016b) available in RapidMiner, namely:

- **k-NN (Lazy Modeling)** (Han et al., 2006): this algorithm is based on learning by analogy. The training examples are described by *n* attributes and each of them represents a point in a *n*-dimensional space. The test example is compared with them by searching the pattern space and it is classified according the *k* training examples closest to it. The similarity is determined in terms of a distance metric, such as the Euclidean distance.
- **Naive Bayes (Bayesian Modeling)** (Unnikrishnan et al., 2011): it is a simple probabilistic classifier, based on the application of the Bayes theorem with the strong (naive) assumption of independence between every pair of features.
- **Decision Tree (Tree Induction)** (Radhakrishnan and Priyaa, 2015): the data is classified using a hierarchical splitting mechanism (repeatedly splitting on the values of attributes), looking like an inverted tree with the root at the top and growing downwards. Each node of the tree corresponds to one of the input attributes. Normally, the recursion stops when all or most of the examples or instances have the same label value.
- **Random Forest (Tree Induction)** (Kotu and Deshpande, 2014): set of a specified number of random trees is generated, working like the Decision Tree. However, it uses only a random subset of attributes for each split. The resulting model is a voting model of all the random trees.

A total of fourteen classification schemes were explored for each set of attributes (6 and 18 attributes) for 1, 2, 3, 4, and 5 survival years. The learning combinations of meta-classifiers with basic classifiers are as follows. The Stacking model used k-NN, Decision Tree, and Random Forest classifiers as base learners, and a Naive Bayes classifier as a Stacking model learner. The Voting model used k-NN, Decision Tree and Random Forest as base learners. The other models were used in combination with each basic classifier. For evaluation purposes, 10-fold cross-validation (Refaeilzadeh et al., 2009) was used to assess the prediction performance of the generated prediction models and avoid overfitting.

©Ediciones Universidad de Salamanca / cc by-nc-nd

Table 4: Survivability Percentage Accuracy.

	Accuracy											
	1 Year		2 Year		3 Year		4 Year		5 Year		Average	
Ensemble Model	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes
Stacking	98.28%	96.15%	97.63%	96.78%	98.02%	97.12%	98.02%	97.26%	97.83%	96.81%	97.96%	96.82%
Voting	97.96%	95.87%	97.41%	96.49%	98.11%	96.57%	98.15%	97.03%	98.09%	96.62%	97.94%	96.52%
Bayesian Boosting with Decision Tree	97.83%	96.33%	97.53%	96.76%	97.81%	96.95%	97.84%	96.98%	97.85%	96.72%	97.77%	96.75%
AdaBoost with Decision Tree	97.83%	96.35%	96.89%	96.78%	97.81%	96.95%	97.84%	97.02%	97.85%	96.74%	97.64%	96.77%
Bagging with Decision Tree	96.88%	95.17%	96.92%	95.97%	97.04%	96.05%	97.1%	96.08%	97.08%	95.76%	97.004%	95.806%
Bayesian Boosting with Random Forest	83.18%	86.79%	84.29%	88.13%	84.4%	88.46%	84.97%	89.16%	85.11%	88.32%	84.39%	88.172%
AdaBoost with Random Forest	82.12%	87.3%	83.64%	87.28%	84.78%	88.95%	83.04%	89.53%	84.17%	88.67%	83.55%	88.346%
Bagging with Random Forest	84.71%	88.81%	84.89%	90.22%	85.81%	90.97%	86.33%	91.15%	85.87%	90.53%	85.52%	90.34%
Bayesian Boosting with Naive Bayes	81.95%	82.19%	83.94%	83.94%	83.23%	84.55%	84.08%	85.02%	83.13%	84.99%	83.27%	84.14%
AdaBoost with Naive Bayes	82.38%	82.08%	83.04%	83.95%	83.41%	84.57%	83.6%	85.11%	83.72%	84.96%	83.23%	84.13%
Bagging with Naive Bayes	80.84%	82.14%	80.18%	83.97%	80.58%	84.5%	80.02%	84.95%	80.05%	84.96%	80.33%	84.10%
Bayesian Boosting with K-NN	97.69%	94.51%	97.58%	94.73%	97.26%	94.78%	97.28%	94.63%	97.19%	94.6%	97.4%	94.65%
AdaBoost with K-NN	97.69%	94.51%	97.58%	94.73%	97.26%	94.78%	97.28%	94.63%	97.19%	94.6%	97.4%	94.65%
Bagging with K-NN	97.69%	94.47%	97.5%	94.77%	97.17%	94.76%	97.3%	94.66%	97.13%	94.54%	97.36%	94.64%

4. Experimental Results and Discussion

Each classification scheme was evaluated using the prediction accuracy and the area under the ROC curve (AUC) for 1, 2, 3, 4, and 5 years. The accuracy is the percentage of correct responses among the examined cases (Bradley, 1997). The AUC can be interpreted as the percentage of randomly drawn data pairs of individuals that have been accurately classified in the two populations (Klepac et al., 2014), and it is commonly used as a measure of quality for classification models (Bradley, 1997). Tables 4 and 5 present all the results obtained for prediction accuracy and AUC respectively. The average performances in terms of accuracy and AUC of the learning schemes for the 5 years are shown in Figures 2 and Figure 3 respectively.

From the observation of the figures and the tables, it is obvious that almost all the classification methods demonstrated high performances, particularly the ones using decision trees. Out of those, the Stacking models showed a slightly better average performance both in terms of accuracy (Figure 2) and AUC (Figure 3).

Comparing the results of the 6-attribute stacking models with those of the 18-attribute models, it is possible to say that the differences are small. With an average of 96.82% for accuracy and 0.989 for AUC, the 6-attribute stacking models had prediction accuracies for years 1 to 5 of 96.15%, 96.78%, 97.12%, 97.26% and 96.81% (as seen in Table 4), and AUCs of 0.984, 0.987, 0.990, 0.991 and 0.991 (as seen in Table 5). The 18-attribute models had an average accuracy of 97.96%, with values for years 1 to 5 of 98.28%, 97.63%, 98.02%, 98.02% and 97.83%. The average AUC was 0.993, and the remaining values were 0.991, 0.993, 0.994, 0.994 and 0.994, for years 1 to 5. The results show that it is possible to build a model with less than half of the features indicated by the expert physician. Regarding the attributes obtained in the feature selection process, with the exception of the site-specific factors, they were all connected with the features indicated by the specialist physician. It should be noted that, in addition to the close performances, the difference between the number of attributes used is important. To apply the attributes in a practical way (for instance, in a tool), the health care professional will lose much time if he must introduce 18 attributes. This is a critical point, as it may lead to the rejection of the tool.

Comparing this approach with others mentioned in Section 2, fewer features were necessary to develop the prediction model. Moreover, in the approach followed in (Al-Bahrani et al., 2013), the closest to the one

©Ediciones Universidad de Salamanca / cc by-nc-nd

Table 5: Survivability AUC.

	AUC											
	1 Year		2 Year		3 Year		4 Year		5 Year		Average	
Ensemble Model	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes
Stacking	0.991	0.984	0.993	0.987	0.994	0.99	0.994	0.991	0.994	0.991	0.993	0.989
Voting	0.988	0.979	0.988	0.982	0.989	0.983	0.99	0.985	0.988	0.984	0.989	0.983
Bayesian Boosting with Decision Tree	0.977	0.963	0.984	0.97	0.979	0.969	0.984	0.973	0.986	0.967	0.982	0.9684
AdaBoost with Decision Tree	0.978	0.967	0.972	0.972	0.981	0.973	0.982	0.974	0.987	0.971	0.98	0.971
Bagging with Decision Tree	0.981	0.977	0.971	0.97	0.974	0.969	0.976	0.972	0.978	0.965	0.976	0.971
Bayesian Boosting with Random Forest	0.894	0.927	0.911	0.932	0.91	0.938	0.91	0.941	0.914	0.934	0.908	0.934
AdaBoost with Random Forest	0.888	0.924	0.908	0.932	0.909	0.936	0.896	0.94	0.9	0.937	0.9	0.934
Bagging with Random Forest	0.925	0.952	0.933	0.959	0.939	0.963	0.94	0.966	0.938	0.963	0.935	0.961
Bayesian Boosting with Naive Bayes	0.896	0.888	0.9	0.9	0.916	0.912	0.916	0.917	0.912	0.913	0.908	0.906
AdaBoost with Naive Bayes	0.901	0.89	0.907	0.902	0.917	0.912	0.914	0.918	0.915	0.914	0.911	0.907
Bagging with Naive Bayes	0.872	0.887	0.885	0.906	0.896	0.92	0.9	0.926	0.898	0.923	0.89	0.912
Bayesian Boosting with K-NN	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
AdaBoost with K-NN	0.977	0.945	0.976	0.947	0.973	0.948	0.973	0.946	0.972	0.946	0.974	0.946
Bagging with K-NN	0.98	0.948	0.979	0.954	0.977	0.953	0.977	0.954	0.977	0.952	0.978	0.952

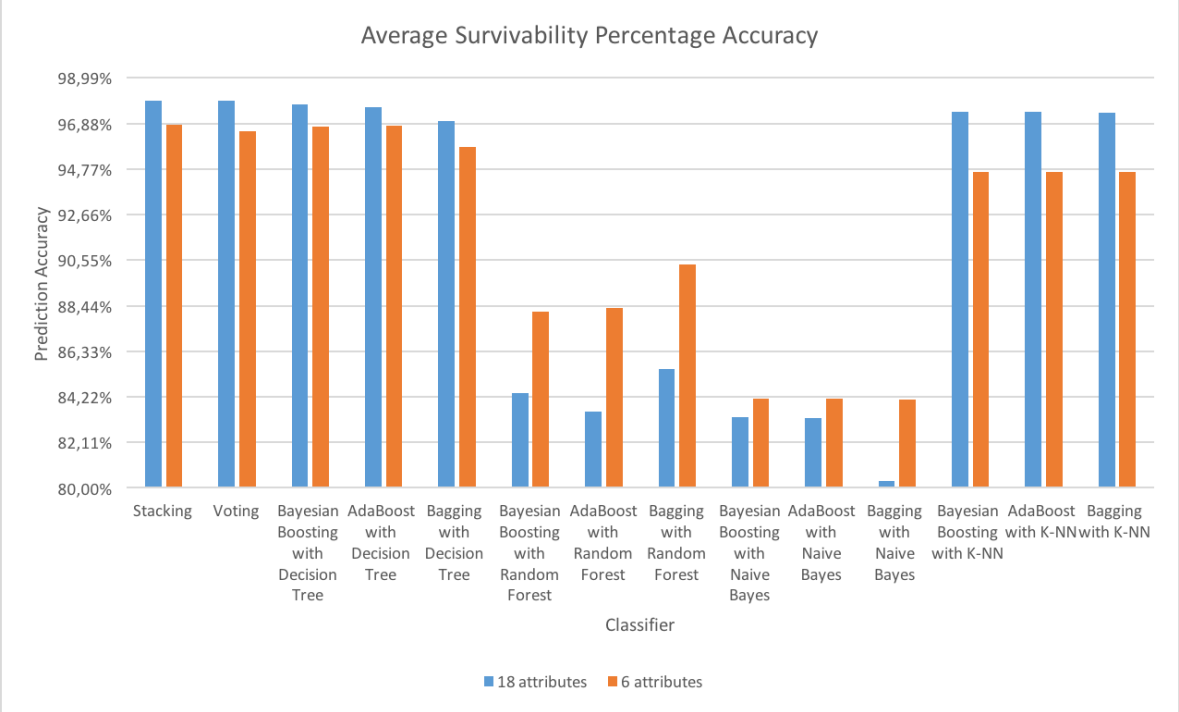

Figure 2: Average survivability percentage accuracy: comparison of the 18-attribute models with the 6-attribute models.

©Ediciones Universidad de Salamanca / cc by-nc-nd

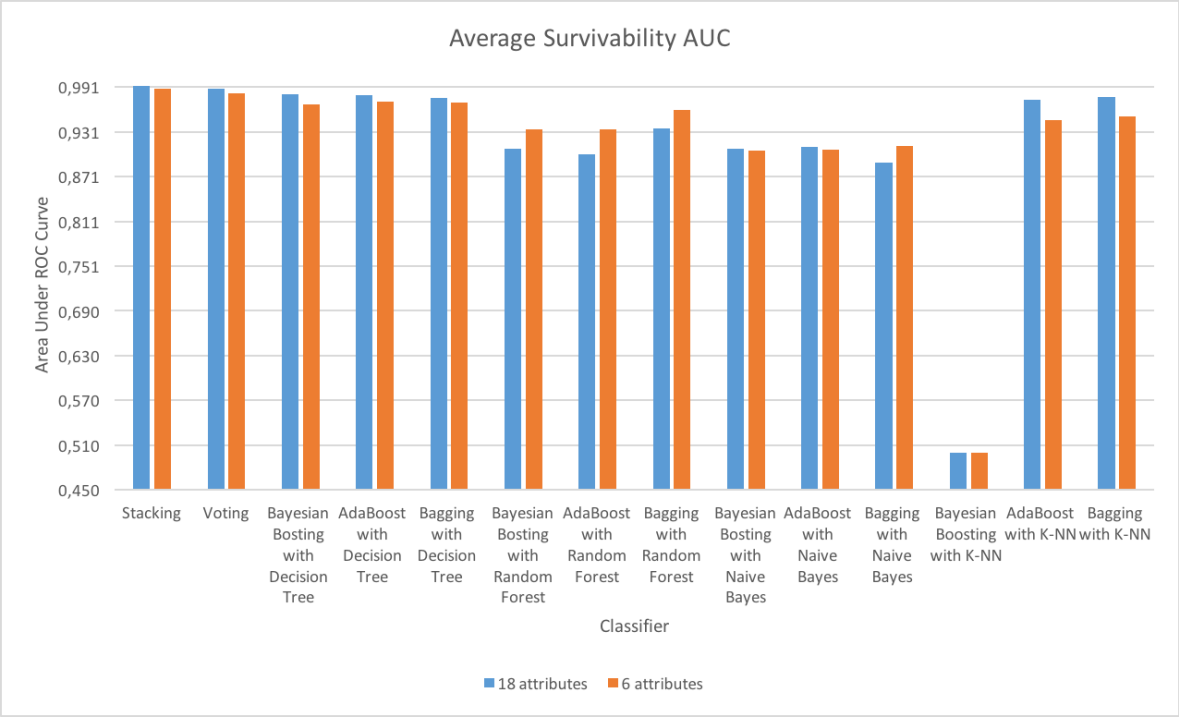

Figure 3: Average survivability AUC: comparison of the 18-attribute models with the 6-attribute models.

©Ediciones Universidad de Salamanca / cc by-nc-nd

followed herein, the best model of colon cancer survivability prediction was based on a Voting classification scheme, with prediction accuracies of 90.38%, 88.01%, and 85.13% and AUCs of 0.96, 0.95, and 0.92 for years 1, 2 and 5. As such, the present work represents an improvement and was able to achieve considerably better results.

5. Conclusions and Future Work

This work involved the use of different meta-classification schemes to construct survival prediction models for colon cancer patients. The best model found uses a Stacking classification scheme, combining k-NN, Decision Tree, and Random Forest classifiers as base learners and a Naive Bayes classifier as a stacking model learner.

The ideal number of features for colon cancer survivability prediction was found to be 6. The selected set includes: age, CS site-specific factor 1, CS site-specific factor 2, derived AJCC stage group, primary site, and regional nodes examined. Overall the developed model was able to present a good performance with fewer features than most of the existing approaches.

As future work we intend to conduct a similar analysis for rectal cancer, a pathology with similar characteristics to colon cancer. Additionally, a mobile application to make the model available to the health care community is under development for different mobile platforms, ready to assist health care professionals in carrying out their duties at any time. In order to ensure that the model is able to adapt and adjust, an on-line learning scheme is also being prepared. In this way, it will be possible for users to dynamically feed new cases to the prediction system and make it change in order to provide better survival predictions. This type of model could also prove to be very useful when integrated in computer-interpretable guideline systems, such as the one described in (Carneiro et al., 2008; Costa et al., 2011; Lima et al., 2011; Oliveira et al., 2013; Oliveira et al., 2014; Novais et al., 2016), as a way to provide dynamic knowledge to rule-based decision support. Future work also includes the development of conditional survivability models that allow the user to get a prediction knowing that the patient has already survived a number of years after diagnosis and treatment. Additionally, we intend to conduct experiments to assess how well the tool fulfils the needs of health care professionals and identify aspects to improve.

Acknowledgements

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013. The work of Tiago Oliveira is supported by a FCT grant with the reference SFRH/BD/85291/ 2012.

6. References

Al-Bahrani, R., Agrawal, A., and Choudhary, A., 2013. Colon cancer survival prediction using ensemble data mining on SEER data. In *2013 IEEE International Conference on Big Data*, pages 9–16. doi:10.1109/BigData.2013.6691752.

Bradley, A. P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159. ISSN 00313203. doi:10.1016/S0031-3203(96)00142-2.

Breiman, L., 1996. Bagging Predictors. *Machine Learning*, 24(2):123–140. ISSN 1573-0565. doi:10.1023/A:1018054314350.

Bush, D. M. and Michaelson, J. S., 2009. Derivation : Nodes + PrognosticFactors Equation for Colon Cancer accuracy of the Nodes + PrognosticFactors equation . Technical report.

©Ediciones Universidad de Salamanca / cc by-nc-nd

Carneiro, D., Costa, R., Novais, P., Neves, J., Machado, J., and Neves, J., 2008. Simulating and Monitoring Ambient Assisted Living. In *Proceedings of the ESM 2008 - The 22nd annual European Simulation and Modelling Conference*, pages 175–182. Le Havre.

Chang, G. J., Hu, C. Y., Eng, C., and et al., 2009. Practical application of a calculator for conditional survival in colon cancer. *Journal of Clinical Oncology*, 27(35):5938–5943. ISSN 0732183X. doi:10.1200/JCO.2009.23.1860.

Chawla, N. V., 2005. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*, pages 853–867. ISBN 9780387254654. doi:10.1007/0-387-25465-X{_}40.

Costa, A., Novais, P., Corchado, J. M., and Neves, J., 2011. Increased performance and better patient attendance in an hospital with the use of smart agendas. *Logic Journal of IGPL*. doi:10.1093/jigpal/jzr021.

Džeroski, S. and Ženko, B., 2004. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54(3):255–273. ISSN 1573-0565. doi:10.1023/B:MACH.0000015881.36452.6e.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F., 2012. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. Last visited on 27/12/2015.

Freund, Y. and Schapire, R. E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55(1):119–139. ISSN 0022-0000. doi:10.1006/jcss.1997.1504.

Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. *Int. J. Emerg. Technol. Adv. Eng*, 2(4):42–47. ISSN 2250-2459.

Han, J., Pei, J., and Kamber, M., 2006. *Data Mining, Southeast Asia Edition*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. ISBN 9780080475585.

Kittler, J., 1998. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, 1(1):18–27. ISSN 1433-755X. doi:10.1007/BF01238023.

Klepac, G., Klepac, G., Kopal, R., and Mri, L., 2014. *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. IGI Global, Hershey, PA, USA, 1st edition. ISBN 1466662883, 9781466662889.

Kotu, V. and Deshpande, B., 2014. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier Science. ISBN 9780128016503.

Leon, M. R. C. D. and Jalao, E. R. L., 2014. Prediction Model Framework for Imbalanced Datasets. (c):33–41.

Lima, L., Novais, P., Neves, J., Bulas, C. J., and Costa, R., 2011. Group Decision Making and Quality-of-Information in e-Health Systems. *Logic Journal of the IGPL*, 19(2):315–332.

National Cancer Institute, 2015. Surveillance, Epidemiology and End Results Program. Last visited on 10/01/2015.

Novais, P., Oliveira, T., and Neves, J., 2016. Moving towards a new paradigm of creation, dissemination, and application of computer-interpretable medical knowledge. *Progress in Artificial Intelligence*, pages 1–7. ISSN 2192-6360. doi:10.1007/s13748-016-0084-2.

Oliveira, T., Leão, P., Novais, P., and Neves, J., 2014. Webifying the Computerized Execution of Clinical Practice Guidelines. In Bajo Perez, J., Corchado Rodríguez, J. M., and et al., editors, *Trends in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection SE - 18*, volume 293 of *Advances in Intelligent Systems and Computing*, pages 149–156. Springer International Publishing.

Oliveira, T., Novais, P., and Neves, J., 2013. Development and implementation of clinical guidelines: An artificial intelligence perspective. *Artificial Intelligence Review*, pages 1–29. doi:10.1007/s10462-013-9402-2.

Radhakrishnan, S. and Priyaa, D. S., 2015. An Ensemble approach on Missing Value Handling in Hepatitis Disease Dataset. *International Journal of Computer Applications*, 130(17).

©Ediciones Universidad de Salamanca / cc by-nc-nd

RapidMiner, 2016a. RapidMiner Documentation: Bayesian Boosting. Last visited on 03/01/2016.

RapidMiner, 2016b. RapidMiner Documentation: Operator Reference Guide. Last visited on 03/01/2016.

RapidMiner, 2016c. RapidMiner Documentation: Optimize Selection. Last visited on 03/01/2016.

Refaeilzadeh, P., Tang, L., and Liu, H., 2009. Cross-Validation. In LIU, L. and ÖZSU, M., editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US. ISBN 978-0-387-35544-3. doi: 10.1007/978-0-387-39940-9_565.

Snow, P. B., Kerr, D. J., Brandt, J. M., and Rodvold, D. M., 2001. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. *Cancer*, 91(8 Suppl):1673–1678. ISSN 0008-543X.

Unnikrishnan, S., Surve, S., and Bhoir, D., 2011. *Advances in Computing, Communication and Control: International Conference, ICAC3 2011, Mumbai, India, January 28-29, 2011. Proceedings*. Communications in Computer and Information Science. Springer Berlin Heidelberg. ISBN 9783642184406.

Vachani, C. and Prechtel-Dunphy, E., 2015. All About Rectal Cancer. Last visited on 27/12/2015.

Weiser, M. R., Gönen, M., Chou, J. F., Kattan, M. W., and Schrag, D., 2011. Predicting survival after curative colectomy for cancer: Individualizing colon cancer staging. *Journal of Clinical Oncology*, 29(36):4796–4802. ISSN 0732183X. doi:10.1200/JCO.2011.36.5080.

Wolff, A. C., Hammond, M. E. H., Schwartz, J. N., and et al., 2007. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Journal of clinical oncology*, 25(1):18–43. ISSN 1527-7755. doi:10.1200/JCO.2006.09.2775.

Yamauchi, M., Lochhead, P., Morikawa, T., Huttenhower, C., Chan, A. T., Giovannucci, E., Fuchs, C. S., and Ogino, S., 2012. Colorectal cancer: a tale of two sides or a continuum? *Gut*, 61(6):794–797. doi:10.1136/gutjnl-2012-302014.

©Ediciones Universidad de Salamanca / cc by-nc-nd