Predictive Clustering Learning Algorithms for Stroke Patients Discharge Planning

Luigi Lella[1,*], Luana Gentile[2,†], Christian Pristipino[3,‡] and Danilo Toni[4,§]
[1]*ASUR Marche, via Oberdan n.2, Ancona, Italy*
[2]*Dept. of Human Neurosciences, Sapienza University, Rome, Italy*
[3]*San Filippo Neri Hospital, ASL1 Roma, Rome, Italy*
[4]*Dept. of Human Neurosciences, Sapienza University, Rome, Italy*

Keywords: Pattern Recognition and Machine Learning, Big Data in Healthcare, Data Mining and Data Analysis, Decision Support Systems.

Abstract: Stroke patients discharge planning is a complex task that could be carried out by the use of a suitable decision support system. Such a platform should be based on unsupervised machine learning algorithms to reach the best results. More specifically, in this kind of prediction task clustering learning algorithms seem to perform better than the other unsupervised models. These algorithms are able to independently subdivide the treated clinical cases into groups, and they can serve to discover interesting correlations among the clinical variables taken into account and to improve the prediction accuracy of the treatment outcome. This work aims to compare the prediction accuracy of a particular clustering learning algorithm, the Growing Neural Gas, with the prediction accuracy of other supervised and unsupervised algorithms used in stroke patients discharge planning. This machine learning model is also able to accurately identify the input space topology. In other words it is characterized by the ability to independently select a subset of attributes to be taken into consideration in order to correctly perform any predictive task.

1 INTRODUCTION

According to the Italian Ministry of Health website, approximately 196,000 strokes occur every year in Italy, of which 20% are relapses. As defined by the World Health Organization, stroke is a "neurological deficit of cerebrovascular cause that persists beyond 24 hours or is interrupted by death within 24 hours" (Italian Ministry of Health website, 2020).

Stroke is caused by an interruption of oxygenated blood supply due to an occlusion or a rupture of the arteries supplying the brain. As a result, brain functions controlled from that area (limb movement, language, vision, hearing or other) are partially or totally impaired or lost (Donnan et al., 2008). About 10-20% of people with stroke die within a month and another 10% within the first year after the event. Only 25% of stroke survivors recover completely, 75% survive with some form of disability, and half of these suffer from a deficit so severe that they lose self-sufficiency.

Major risk factors include age, high blood pressure, tobacco smoking, obesity, high blood cholesterol, diabetes mellitus, a previous TIA or stroke, and atrial fibrillation. Diagnosis is performed by a physical examination and it is supported by neuroimages (CT and/or MR).

During hospitalization, different pharmacological and/or interventional treatments are put in place to preserve vital functions and minimize brain damage. The National Institutes of Health Stroke Scale (NIHSS) is used to assess stroke severity (Putra Pratama et al., 2019; Lyden et al., 2009).

In the discharge phase, it is important to make a proper plan, not only to enhance individual recovery, but also to reduce the high social burden and the use of health system resources (Mess et al., 2016)

* www.linkedin.com/in/luigi-lella
† https://www.neuroscienze.uniroma1.it/
‡ https://www.aslroma1.it/presidi-ospedalieri/presidio-ospedaliero-san-filippo-neri
§ https://www.neuroscienze.uniroma1.it/

Lella, L., Gentile, L., Pristipino, C. and Toni, D.
Predictive Clustering Learning Algorithms for Stroke Patients Discharge Planning.
DOI: 10.5220/0010187502960303
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 296-303
ISBN: 978-989-758-490-9

(Pereira et al.,2014). To help in planning, functional outcome after stroke is evaluated using the modified Rankin Scale (mRS) which measures the degree of disability or dependence in activities of daily living. The lower the score the higher the likelihood of being able to live at home with a degree of independence after being discharged from the hospital or a long-term care ward (Saver et al.,2010)(Wilson et al.,2002).

Aim of this paper is to present an innovative and effective prediction model for discharge planning, based on a machine learning algorithm derived from data gathered during hospitalisation in the acute phase.

2 PREVIOUS WORK

A considerable body of literature exists on the use of machine learning algorithms based on the assimilation of the data of previously treated clinical cases. The accuracy of prediction generally tends to increase over time, as new data become available (Bishop, 2006). However, there is no algorithm capable of providing the best predictive accuracy for each category of problem (Alpaydin, 2020).

Machine learning algorithms can be subdivided into supervised and unsupervised. In the supervised learning human experts select the correct answers of the machine learning model, while unsupervised learning does not need the intervention of human experts (Alpaydin, 2020).

Within the unsupervised learning paradigm, it is possible to operate a further subdivision between symbolic models that seek to reach a formal representation of knowledge (using for example logical representations, inference rules or decision trees) and sub-symbolic models where the acquired knowledge is stored in complex representations such as artificial neural networks. Clustering learning algorithms (Van Hulle, 1989; Kohonen, 1988; Kohonen, 1989; Kohonen, 1990) are subsymbolic unsupervised models that allow to achieve the best results in an unsupervised manner. If a class attribute is chosen, such as the length of the hospital stay or the outcome expressed using a suitable evaluation scale, these algorithms are able to divide the clinical cases that can occur in clusters, corresponding to hospitalizations that end with the same length of hospital stay or with the same outcome. Even if these models are not classifiers, after being trained with a set of clinical cases used as test sets, they are able to assign a test set record to a correct class, achieving a considerable high prediction accuracy when new treated cases are processed. Among the learning cluster algorithms, Kohonen's Self Organizing Maps (Kohonen, 1989) have been widely used in the health sector, but better results can be achieved by using a more adaptive algorithm such as Fritzke's Growing Neural Gas (GNG) (Fritzke,1994). GNG is an incremental network model based on a simple Hebb-like learning rule. Unlike previous approaches like the “neural gas” method of Martinetz and Shulten (Martinetz and Shulten, 1991) (Martinetz and Shulten,1994) this model has fixed parameters and it is able to continue the learning phase, adding other neural units and connections, until a performance goal is achieved. By being able to accurately identify the input space topology, this model is able to identify the variables to be considered in order to effectively operate the prediction of a class attribute.

For example, if the mRS is chosen as class attribute (that is the study variable to be predicted), the GNG model could predict that a male individual aged between 40 and 65, with a particular form of diagnosed stroke and treated with certain pharmacological therapies, at the end of the hospital stay, is able to reach a mRS score of 1. This prediction could be made without taking into consideration other variables such as risk factors, results of instrumental and laboratory tests etc. In other cases the same model could need a greater number of variables in order to make an accurate prediction.

The GNG model has already been successfully used in similar areas, such as in the prediction of the length of hospital stay (Lella and Licata, 2017)

Functional outcome after stroke is related to many variables, from biometric data (gender, age, weight, pressure levels) to risk profiles, from the results of laboratory tests to the results of instrumental tests to therapies.

Given this complex interaction of many variables, using machine learning appears to be the best solution choosing an unsupervised model that can improve its predictive accuracy over time. The purpose of these algorithms is to identify over the course of time new possible and clinically meaningful correlations among the available data.

This approach has proved particularly effective in the clinical setting, especially to predict the therapeutic outcome of stroke patients (Zdrodowska, 2019; Zdrodowska et al., 2018; Chen et al., 2017) (Alaiz-Moreton, 2018). The best results were obtained through the use of the Random Forest algorithm (Breiman, 2001; Alaiz-Moreton, 2018) (Tin Kam Ho, 1998; Tin Kam Ho, 1995), an

ensemble learning method for classification obtained through the aggregation of decision trees, and the PART model, a partial decision tree algorithm that does not need to perform global optimization like other rule based learners (Zdrodowska et al., 2018; Ali and Smith, 2006). Excellent results were also obtained by using supervised models such as the Support Vector Machine (SVM) (Zdrodowska et al., 2018), a model based on a binary non-probabilistic linear classifier (Ben Hur et al.,2001) (Cortes and Vapnik, 1995).

3 METHODOLOGY

On the basis of the above literature review, it was decided to use GNG networks to predict the status of stroke patients one day and seven days after entering the hospital, as well as at hospital discharge and after three months, comparing the results with the ones achieved by the best models used in this kind of study, i.e. the Random Forest model, the PART model and the SVM model.

The ZeroR model (Witten et al., 2011), the OneR (Holte, 1993), the Naive Bayes (John and Langley, 1995) and the J48 (Witten et al., 2011) were also taken into consideration.

The ZeroR model is used as a benchmark to verify that all the other tested algorithms have been configured and used correctly. ZeroR always predicts the most frequent class variable in the presence of any combination of input variables. Given its simplicity, it has generally a much lower level of prediction accuracy than the other algorithms. If this does not occur, the found result may be due to a bad data selection and coding, or to a bad configuration of the models.

The OneR, which stands for "one Rule", is a one-level decision tree. In various areas and predictive tasks, this model has proved to be much more efficient than other more complex models, and it is always advisable to check whether the problem in question can be effectively treated using this model which requires a reduced amount of resources.

The Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

The J48 is a decision tree based on the "divide and conquer" strategy used recursively. At each training step, the node characterized by the highest amount of information is selected and split into a series of nodes corresponding to some possible values that the original node can assume. The process ends when all the considered instances refer to the same class attribute value.

The study involved a subset of 20,000 samples taken from the Italian subset of the SITS registry (Safe Implementation of Treatments in Stroke website, 2020), a non-profit, research- driven, independent and international monitoring initiative for stroke patients.

Of these, just the data that hold the fields of the outcome 24 hours after the patient's access to the hospital (Global Outcome 24), the mRS at 7 days (Rankin at 7 days), the mRS at hospital discharge (Rankin at hospital discharge) and the mRS at 3 months (Rankin at 3 months) were taken into consideration. These three variables were chosen as class attributes for the tests.

The first class attribute is a qualitative clinical variable that can take 7 possible values ("muchBetter", "better", "unchanged", "worse", "muchWorse", "dead"). The other three class variables can only take the scores 0,1,2,3,4,5 and 6.

The first subgroup of records, with the global outcome at 24 hours specified, consisted of 13008 records characterized by 69 non class attributes; the second subgroup, with the mRS at 7 days specified, was made by 10460 records characterized by 99 non class attributes; the third subgroup, with the mRS at hospital discharge specified, was constituted by 4989 records characterized by 129 non class attributes; the fourth subgroup, with the mRS at 3 months specified, was made by 10777 records characterized by 152 non class attributes. The reason why the number of non-class attributes is different in the four subsets is that an higher length of hospitalization also increases the number of available data deriving from further tests performed. Therefore, the number of non-class variables that can be considered increases.

The characteristics of the four subsets of records are presented in table 1.

Table 1: Features of the considered data sets.

Data set	Total no.(male,female); avg age(min,max); hemorrhagic stroke; ischemic stroke
1–Global Outcome 24	13008(6974,6034); 71(14,102); 1902; 11106
2–Rankin at 7 days	10460(5491,4969); 71(14,101); 1376; 9084
3–Rankin at Hospital Discharge	4989(2780,2209); 68(14,102); 557; 4432
4- Rankin at 3 months	10777(5685,5092); 71(14,104); 755; 10022

Part of the patient data was incomplete, but it was not necessary to perform any type of data cleaning due to the data entry performed by a codified online form.

Data were discretized and normalized before being processed. All predictive machine learning models taken into consideration were trained with 60% of the samples and tested with the remaining 40% of the samples. The Weka 3.8.4 platform was used to test ZeroR, OneR, J48, Naive Bayes, SVM and Random Forest models, while a Java implementation was used to test the GNG model.

The sequential minimal optimization algorithm (Platt, 1998) was used to train the SVM model.

The GNG model was tested with the following parameters:

λ=100, ε_b=0.2, ε_n=0.006, α=0.5, α_{max}=50, δ=0.995.

The training was stopped when the mean square error, i.e. the main of the local square error related to each unit (expected distortion error), dropped below the threshold of E=0.7.

4 RESULTS

The results in terms of prediction accuracy (i.e. the number of correct predictions over the total number of predictions) are shown in table 2.

All the tested models reached an higher prediction accuracy than the ZeroR model, and the OneR model resulted to be the second worst algorithm, confirming the complexity of the prediction task. Naive Bayes, J48, SVM and PART performed better with a low number of training records characterized by more non class attributes.

The best results were obtained by the GNG model and by the Random Forest model, followed by PART and SVM. The result confirmed the correctness of the choice of the unsupervised models over the supervised ones and the chosen clustering learning model proved to be more performing than the Random Forest ensemble model.

Once trained, it is also possible to use the GNG model to identify which non-class attributes are linked to particular values of the selected class attribute. For example, it is possible to identify which clinical variables are linked to the worsening of patients during the first 24 hours of hospitalization.

Using the Girvan-Newman algorithm (Girvan and Newman, 2002) it is possible to identify communities of nodes starting from the portion of the trained model of self-organizing neural network associated with a deterioration of state. In the case of the first GlobalOutcome24 subset the selected nodes are those in which the values of the class attribute code corresponding to "worse", "muchWorse" and "dead" exceed a threshold value that has been set equal to 0.7. Girvan-Newman's algorithm identifies the communities of nodes by eliminating those connections characterized by the greatest number of shorter paths that link each pair of nodes. After training the first model with the records of the first subset of input records (GlobalOutcome24) and selecting the network portion corresponding to the values of the class attribute "muchWorse" and "dead", the Girvan-Newman algorithm was used to remove the first 100 connections characterized by the greatest number of shorter paths between pairs of nodes.

Table 2: Prediction accuracy of the tested models.

Tested Models	Prediction Accuracy on data set 1; data set 2; data set 3; data set 4
ZeroR	30.67; 19.55; 24.05; 27.73
OneR	38.72; 29.96; 37.36; 43.73
Naive Bayes	29.44; 46.21; 80.36; 56.76
J48	68.9; 71.48; 82.08; 44.88
SVM	41.48; 78.46; 99.19; 82.57
PART	72.26; 76.54; 83.04; 46.89
Random Forest	97.99; 97.54; 97.25; 84.89
GNG	99.17; 99.64; 99.05; 89.88

Table 3: Clinical variables related to death in the first 24 hours of hospitalization.

Cluster no.	Label (weight)
1	**Hypertension(0.99)**; HighTemperature(0.97)...
2	**NIHSS1A=3(1.00)**; GenderMale(1.00)...
3	**NIHSS5A=4(0.99)**; NIHSS6A=4(0.99)...
4	**LowAPTTvalues(1.00)**; GenderMale(1.00)...
5	**NIHSS11=9(1.00)**; GenderMale(1.00)...
6	**GenderMale(1.00)**;Hypertension(0.96)
7	Hypertension(0.99); **NIHSS5B=4(0.98)**...
8	Hypertension(0.99); **Age>=80(0.96)**...
9	Hypertension(0.99); **Diabetes(0.95)**...
10	Hypertension(0.99); **HighTemperature(0.96)**...
11	**CurrentInfarct(1.00)**; GenderMale(1.00)...
12	NIHSS5A=4(0.98); **Age65-80(0.83)**...
13	NIHSS1C=2(0.95); **NIHSS6A=4(0.90)**...
14	**Hyperlipidaemia(1.00)**;GenderMale(1.00)...
15	**NIHSS4=2(1.00)**; GenderMale(1.00)...
16	**PreviousStroke<3months(1.00)**...
17	**NIHSS8=2(1.00)**; GenderMale(1.00)...
18	**CerebralOedema(1.00)**; GenderMale(1.00)...
19	NIHSSB=2(0.93); **NIHSS1C=2(0.88)**...
20	**NIHSS7=4(1.00)**; GenderMale(1.00)...

Subsequently, the non-class attributes associated with the nodes of the individual communities were extracted, that is, those characterized by corresponding code values higher than the threshold value of 0.7.

For each node, using a weight function tf-idf (Baeza, 1999), the attribute most frequent in its own community and less frequent in the overall set of extracted communities was chosen. In this way, all the non-class attributes associated with the worsening of the clinical picture were identified. The results of this processing are shown in table 3 and table 4.

Table 4: Clinical variables related to the worsening in the first 24 hours of hospitalization.

Cluster no.	Label (weight)
1	**NIHSS1B=2(1.00)**; Hypertension(1,00)...
2	Hypertension(1.00); **NIHSS5A=4(0.88)**...
3	Hypertension(0.97); **NIHSS5B=4(0.87)**...
4	Hypertension(0.92); **NIHSS4=2(0.88)**...
5	NIHSS1B=2(1.00); **Hypertension(0.94)**...
6	Hypertension(0.97); **GenderFemale(0.96)**...
7	**Age65-80(0.98)**; NIHSS3=2(0.98)...
8	**NIHSS5B=4(0.99)**; NIHSS4=2(0.99)...
9	Hypertension(1,00); **LowAPTTvalues(0.93)**...
10	NIHSS4=2(0.99); **NIHSS11=2(0.88)**...
11	Hypertension(0.98); **CurrentInfarct(0.92)**...
12	**NIHSS10=1(0.99)**; Age18-65(0.87)
13	Hypertension(0.86); **GenderMale(0.81)**; ...

For each cluster of nodes extracted by the use of Girvan-Newman's algorithm, the non-class attributes associated to the corresponding nodes are listed. Due to the GNG model training method, it is reasonable to assume that each identified cluster is related to a subset of analysed clinical cases.

The clinical variables associated with the various clusters which are related to cases of serious worsening and death patient are displayed in the results tables 3 and 4. In the first lines the most important clusters are represented, i.e. those associated with a greater number of neural units and therefore with a greater number of clinical cases.

For each cluster the non-class attributes, i.e. the clinical variables responsible for the worsening or the death of the patient, are sorted in descending order by weight associated with the attribute. The weight is a coefficient between 0 and 1 which indicates how much the neuronal units associated with the selected cluster are activated when the attribute is present in the considered clinical case. Values close to 0 mean that the attribute is not very relevant for the worsening or death of the patient, those close to 1 are instead considered the main responsible factors.

The obtained results show that the presence of hypertension, assessed on the basis of the systolic and diastolic blood pressure values, is considered an important potential factor for the worsening the patient's state which can also lead to death. The presence of hypertension alone, however, is not sufficient to infer the risk of worsening or death. Considering for example the first cluster of table 3 related to the death cases, hypertension is an important factor having a weight equal to 0.99, but it must also be accompanied by other factors such as hightemperature, hyperlipidaemia, genderfemale, NIHSS4 = 2 and age65-80.

Using the tf-idf algorithm, the most representative clinical variables of the considered cluster were selected. These are represented in bold and they allow to highlight which are the attributes most related to the worsening or death of the patient.

For example, the attributes of table 3 most related to death cases with the relative weights are PreviousStroke<3months (1.00), Hyperlipidaemia (1.00), LowAPTTvalues(1.00), CerebralOedema (1.00), GenderMale (1.00), NIHSS8 = 2 (1.00), NIHSS4 = 2 (1.00), NIHSS1A = 3 (1.00), NIHSS7 = 4 (1.00), Hypertension (0.99), NIHSS5A = 4 (0.99), NIHSS5B = 4 (0.98), HighTemperature (0.96), Age> = 80 (0.96), Diabetes (0.95), NIHSS6A = 4 (0.90), NIHSS1C = 2 (0.88), Age65-80 (0.83). This means that the presence of a previous stroke that occurred no more than three months before, accompanied by hyperlipidaemia or a cerebral oedema or a low level of APTT, especially if the patient is male can be considered a quite worrying clinical picture. An age greater than 80 is to be considered a more important risk factor than an age between 65 and 80.

The clinical variables of table 4 related to the severe worsening of the patient with the relative weights are NIHSS1B = 2 (1.00), NIHSS5B = 4 (0.99), NIHSS10 = 1 (0.99), Age65-80 (0.98), GenderFemale (0.96), Hypertension (0.94), LowAPTTvalues(0.93), NIHSS11 = 2 (0.88), NIHSS4 = 2 (0.88), NIHSS5A = 4 (0.88), NIHSS5B = 4 (0.87), GenderMale (0.81). This means that the presence of a low sense of orientation, of plegia or dysarthria especially if the patient is aged between 65 and 80 and hypertensive is to be considered a potential worsening factor. Female patients are considered more at risk of serious worsening than male patients.

A similar study was carried out to identify the clinical variables most correlated with a serious level of disability after 3 months (corresponding at a mRS of 3, 4 or 5). The results are shown in table 5.

As it can be seen, high blood pressure levels are always related to the appearance of a disabling stroke. The mRS at three months is also strongly affected by the patient's state 7 days after hospitalization. Generally there are only slight differences in outcomes between male and female patients, although clusters 1, 4 and 6 show a greater correlation with the male sex while clusters 3 and 5 are more correlated with the female sex.

The first cluster is represented by patients aged between 65 and 80 characterized by the presence of occlusions that lead to the appearance of an ischaemic penumbra (perfusion infarct mismatch). The second cluster is represented by underweight patients over the age of 65 characterized by diabetes and hyperlipidaemia and occlusions and a hyperdensity of the arteries. The third cluster relates to underweight patients characterized by a mRS of 4 both after the first 7 days and upon discharge from the hospital. Such patients are characterized by low APTT levels. The fourth cluster is made up of patients aged between 18 and 65, with high blood cholesterol levels, diabetes and occlusions. The fifth cluster is represented by patients over eighty characterized by hyperlipidaemia and low levels of APTT. The sixth cluster is associated with pre-diabetic patients aged between 65 and 80 years.

Table 5: Clinical variables related to the worsening in the first 24 hours of hospitalization.

Cluster no.	Label (weight)
1	**mRS_4_dis(0,90)**; mRS_3_7d(0,99); ...
2	**LowApttValues(0,96)**; Hypertension(1,00); ...
3	**Age65-80(0,85);** Hypertension(0,97); ...
4	**GenderMale(0,91);** Hypertension(0,99); ...
5	**Age>=80(0,92)**; GenderFemale(0,97); ...
6	**Prediabetes(0,83)**; Hypertension(0,97); ...

The age range most at risk of incurring permanent disabilities at 3 months after treatment is consistent with the fact that in this cohort 78% of patients with a 3-month mRS equal to 3, 4 or 5 are over 65-year-olds.

Analysing the subset of patients with a 3-month mRS of 3, 4 or 5 aged between 65 and 80, in 72% of these cases a significant ischaemic penumbra is detected (perfusion infarct mismatch) and in 44% of cases also an hyperdense artery sign. This subset of patients is clearly identifiable with cluster 1.

Considering instead the subset of patients with a 3-month mRS of 3, 4 or 5 aged between 65 and 80 characterized by low APTT levels, in 85% of cases there are also high serum glucose levels. This subset of cases can be associated with clusters 2 and 3.

Analysing the subset of patients with a 3-month mRS equal to 3, 4 or 5 aged between 18 and 65 years it is found that in this case 67% of the patients are male, 81% have low APTT levels and 80% had high serum glucose levels. This subset of patients is clearly identifiable with cluster 4.

The 87% of the subset of patients with a 3-month mRS equal to 3, 4 or 5 over the age of 80 with hyperlipidaemia are patients with low APTT levels. This group can be associated with cluster 5.

5 CONCLUSIONS

The findings of this study suggest that the use of clustering learning algorithms allows to identify in an unsupervised way a set of clinical variables which can be taken into consideration in order to carry out a good prediction of the clinical outcome.

The Growing Neural Gas model has proved particularly effective in predicting the patient outcome compared to other algorithms used in the same application area. The best result in terms of predictive accuracy achieved by this model is due to its ability to exactly identify the input space topology, which also makes it particularly robust to noise and lack of data. By analyzing the final configuration of the trained GNG network, it is also possible to obtain useful information on the attributes which are most correlated with certain outcomes. Statistical analyses carried out on the data used as training set and test set seem to confirm the consistency of the extracted knowledge. The developed model is ready to be tested in prospective studies in the real world.

REFERENCES

Alaiz-Moreton, H., Fernandez-Roblez L., Alfonso Cendon, J., Castejon-Limas, M., Sanchez-Gonzalez L., Perez H., 2018. *Data mining techniques for the estimation of variables in health-related noisy data.* Advances in intelligent systems and computing, 649, 482-491.

Ali S., Smith K.A., 2006. *On learning algorithm selection for classification*. Applied Soft Computing, 6, 119-138.

Alpaydin E., 2020. *Introduction to Machine Learning*. 4th

Edition. MIT Press.

Baeza Y., 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., ISBN 0-201-39829-X.

Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.N., 2001. *Support vector clustering*. Journal of Machine Learning Research. 2, 125–137.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*, Springer.

Breiman L., 2001. *Random Forests*. Machine Learning. 45(1), 5–32. doi:10.1023/A:1010933404324.

Chen Y.C., Suzuki T., Suzuki M., Takao H., Murayama Y., Ohwada H., 2017. *Building a classifier of onset stroke prediction using random tree algorith*m. International Journal of Machine Learning and Computing, 7(4), 61-66.

Cortes, C., Vapnik, V.N., 1995. *Support-vector networks*. Machine Learning. 20(3), 273–297. CiteSeerX10.1.1.15.9362. doi:10.1007/BF00994018.

Donnan, G.A., Fisher M., Macleod M., Davis S.M., 2008. *Stroke*. The Lancet. 371(9624), 1612–23.

Fritzke B., 1994. *A Growing Neural Gas Network Learns Topologies*. Part of: Advances in Neural Information Processing Systems 7, NIPS.

Girvan M., Newman M.E.J., 2002. *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA 99, 7821–7826.

Holte,R.C., 1993. *Very simple classification rules perform well on most commonly used datasets*. Machine Learning.

Italian Ministry of Health website, 2020. http://www. salute.gov.it/portale/salute/p1_5.jsp? lingua=italiano& id=28&area=Malattie_cardiovascolari, last accessed 2020/04/24.

John, G.H.; Langley, P., 1995. *Estimating Continuous Distributions in Bayesian Classifiers*. Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence. Morgan Kaufmann. 338–345. arXiv:1302.4964

Kohonen T., 1988. *An introduction to neural computing*. Neural Networks, 1, 3-16.

Kohonen T., 1989. *Self-Organization and Associative Memory*, Berlin: Springer-Verlag.

Kohonen T., 1990. *The Self Organizing Map*. Proc of the IEEE, 78(9).

Lella L., Licata I., 2017. *Prediction of Length of Hospital Stay using a Growing Neural Gas Model*. In Proceedings of the 8th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2017), 175-178.

Lyden P., Raman R., Liu L., Emr M., Warren M., Marler J., 2009. *National Institutes of Health Stroke scale certification is reliable across multiple venues*. Stroke, 40(7), 2507-2511. doi:10.1161/STROKEAHA.116. 015434.

Martinetz, T. M., Schulten, K J., 1991. *A "neural-gas" network learns topologies*. In Kohonen, T., Makisara, K, Simula, 0., and Kangas, J., editors, Artificial Neural Networks, North-Holland, Amsterdam, 397-402.

Martinetz, T. M., Schulten, K J., 1994. *Topology representing networks*. Neural Networks, 7(3), 507-522.

Mess M., Klein J., Yperzeele L., Vanacker P., Cras P., 2016. *Predicting discharge destination after stroke: A systematic review*. Clin Neurol Neurosurg. 142(15-21). doi:10.1016/j.clineuro.2016.01.004.

Pereira S., Foley N., Salter K., McClure J.A., Meyer M., Brown J., Speechley M., Teasell R., 2014. *Discharge destination of individuals with severe stroke undergoing rehabilitation: a predictive model*. Disabil Rehabil. 36(9), 727-731. doi:10.3109/09638288. 2014.902510.

Platt,J., 1998. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Technical Report MSR-TR-98-14.

Putra Pratama A., Tresno T., Wahyu Purwanza S., 2019. *Development the national institutes of health stroke scale (NIHSS) for predicting disability and functional outcome to support discharge planning after ischemic stroke*. Journal Ners, 14(3).

Saver J.L., Filip B., Hamilton S., Yanes A., Craig S., Cho M., Conwit R., Starkman S., FAST-MAG Investigators and Coordinators, 2010. *Improving the reliability of stroke disability grading in clinical trials and clinical practice: the Rankin Focused Assessment (RFA)*. Stroke. 41 (5): 992– doi:10.1161/STROKEAHA.109.571364. PMC 2930146. PMID 20360551

Safe Implementation of Treatments in Stroke website, 2020. https://sitsinternational.org, last accessed 2020/04/24.

Tin Kam Ho, 1998. *The Random Subspace Method for Constructing Decision Forests*. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844, DOI:10.1109/34.709601.

Tin Kam Ho, 1995. *Random Decision Forests*. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 278–282.

Van Hulle M. M., 1989. *Self Organizing Maps*. Handbook of Natural Computing, 585-622.

Wilson J. L., Hareendran A., Grant M., Baird T., Schulz U.G., Muir K.W., Bone I., 2002. *Improving the Assessment of Outcomes in Stroke: Use of a Structured Interview to Assign Grades on the Modified Rankin Scale*. Stroke. 33 (9): 2243–2246. doi:10.1161/01.STR.0000027437.22450.BD. PMID 12215594

Witten,I. H., Frank,E., Hall,M.A., 2011. *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.

Zdrodowska M., 2019. *Attribute selection for stroke prediction*. Sciendo. Doi 10.2478/ama-2019-0026.

Zdrodowska M., Dardzinska M, Chorazy M., Kulakowska A., 2018. *Data Mining Techniques as a tool in neurological disorders diagnosis*. Acta Mechanica et Automatica, 12(3), 217-220.

APPENDIX

For the input data preprocessing the following coding rules were used:

The age was codified in four main classes: age>=80; 65<=age<80; 18<=age<65 and 0<=age<18.

The clinical variables “Hypertension” and “Hyperlipidemia” refer to the risk factors which are specified in the data entry form. They are not automatically computed by the use of a codifying formulas.

If the glucose level is lower than 69 the value is codified as “Hypoglycemia”, if the glucose level is between 70 and 99 the value is codified as “Normoglycemia”, if the glucose level is between 100 and 124 the value is codified as “Prediabetes”, if the glucose level is higher than 125 the value is codified as “Diabetes”.

If the cholesterol level is lower than 199 the value is codified as “NormalCholesterol”, if the cholesterol level is higher than 200 the value is codified as “HighCholesterol”.

If the temperature level is lower than 98.5 Fahrenheit the value is codified as “NormalTemperature”, if the temperature level is higher than 98.6 the value is codified as “HighTemperature”.

If the APTT is lower than 29 the value is codified as “LowApttValues”, if the APTT level is between 30 and 39 the value is codified as “NormalApttValues”, if the APTT level is higher than 40 the value is codified as “HighApttValues”.

If the BMI is lower than 18.4 the value is codified as “Underweight”, if the BMI is between 18.5 and 24,9 the value is codified as “Normweight”, if the BMI is between 25 and 29,9 the value is codified as “IncreasedOverweight”, if the BMI is between 30 and 34,9 the value is codified as “ModerateOverweight”, if the BMI is between 35 and 39,9 the value is codified as “SevereOverweight”, if the BMI is higher than 40 the value is codified as “VerySevereOverweight”.