Assessing Stroke Risk: Study on the Predictive Accuracy of Bayesian Networks

Alessio Drigatti
Zurich University of Applied Sciences
School of Engineering
drigaale@students.zhaw.ch

Abstract

This study explores the application of Bayesian Networks to predict stroke risk using a dataset comprising 28,685 subjects after adjustments. Utilizing a Tabu search algorithm, the study models the relationships among variables such as age, gender, hypertension, and smoking status, identified as stroke risk factors by previous research. The resulting Bayesian Network demonstrated an Area Under the Curve (AUC) of 0.79 on the validation dataset, indicating a good predictive accuracy. This research underscores the potential of Bayesian Networks in predictive health analytics, suggesting avenues for future work in enhancing model performance through finer variable categorization and balanced modeling approaches.

1 Introduction

Recent advancements in predictive health analytics have introduced innovative methods for forecasting medical outcomes and identifying at-risk populations. Among these, Bayesian networks have emerged as a powerful tool for modeling complex relationships between various risk factors and health outcomes.

Park et al. and Song et al. highlighted the utility of Bayesian networks in enhancing our understanding of stroke, with Park focusing on post-stroke outcomes and Song on identifying key risk factors (Park et al., 2018) (Song et al., 2022).

This study aims to model a simple Bayesian Network using a replica dataset from Harvard University with a comparatively small number of variables to investigate and compare its performance. The paper is structured as follows: First, we explore the dataset. Then, we explain the relationships for the Bayesian Network in the context of current research. Next, we discuss the resulting network and its performance. Finally, we compare its performance with that of other studies.

2 Data Description

This study utilizes a replication dataset for stroke prediction (M, 2021), comprising 43,000 subjects and the following variables: *gender, age, hypertension, heart disease, average glucose level, bmi, smoking status, stroke*. Approximately 59% of the subjects are female, and 41% are male. The average age is just over 42 years (Median: 44, Min: 0.08, Max: 82). For this analysis, subjects under 18 years old, comprising about 17.4% of the dataset, are excluded. Additionally, around 31% of the subjects are removed for lacking smoking status information. After these adjustments, the variables distribution still exhibits a relatively similar balance, except for the observation with stroke. The resulting dataset consists of 28,685 subjects.

For this study, average glucose levels are categorized as follows: *Normal* (< 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* (> 126 mg/dl). BMI is categorized as *Normal* (BMI < 30) and *Overweight* (BMI $\geq$ 30). Finally, age is categorized as *Young* (18 to 30), *Adult* (30 to 60), and *Senior* (> 60). The dataset and variable distribution are described as follows:

- **Gender**: Female 17,594 | Male 11,084
- **Age**: Adult 14,513 | Senior 8,988 | Young 5,177
- **Hypertension**: No 25,152 | Yes 3,526
- **Heart Disease**: No 27,000 | Yes 1,678
- **Average Glucose Level**: Diabetic 5,858 | Normal 17,334 | Prediabetic 5,486
- **BMI**: Normal 15,595 | Overweight 13,083
- **Smoking Status**: Formerly 7,299 | Never 14,981 | Smokes 6,488
- **Stroke**: No 28,040 | Yes 638

Finally, the dataset is split into two sets: 80% for model training and 20% for model evaluation, ensuring the proportion of subjects with stroke is maintained.

2.1 Bayesian Network

Bayesian networks are graphical models that allow representing probability relationships among a set of variables. This study employs a Tabu search algorithm to learn the structure of a Bayesian Network from the dataset. This heuristic search algorithm seeks optimal or near-optimal solutions in large search spaces by performing local searches while avoiding cyclic paths that would lead back to previously examined solutions (Heckerman et al., 1995).

Prior knowledge about stroke risk factors, as identified by previous research, can be incorporated into the model. Elkind et al., in 1998, identified age, gender, diabetes, smoking, hypertension, weight, among others, as risk factors for stroke (Elkind and Sacco, 1998). Many of these risk factors have been consistently confirmed in numerous studies (Zhang et al., 2019). Based on the current state of knowledge, the following relationships (whitelist) are defined for this study:

$$Hypertension \rightarrow Stroke \tag{1}$$
$$AverageGlucoseLevel \rightarrow Stroke \tag{2}$$
$$SmokingStatus \rightarrow Stroke \tag{3}$$
$$BMI \rightarrow Stroke \tag{4}$$
$$Age \rightarrow Stroke \tag{5}$$

Furthermore, the following relationships (blacklist) are defined, as various variables cannot influence each other:

$$Any* \rightarrow Age \tag{6}$$
$$Any* \rightarrow Gender \tag{7}$$
$$Gender \rightarrow Age \tag{8}$$
$$Age \rightarrow Gender \tag{9}$$

3 Results

Based on the defined relationships, the resulting network is shown in Figure 1. Notably, age influences all variables, aligning with general knowledge. Furthermore, gender impacts heart diseases and smoking status, assuming an asymmetrical distribution of smokers, hence the Tabu algorithm recognizes a relationship. The paper does not delve further into these relationships as they are generally understandable and exhibit no anomalies.

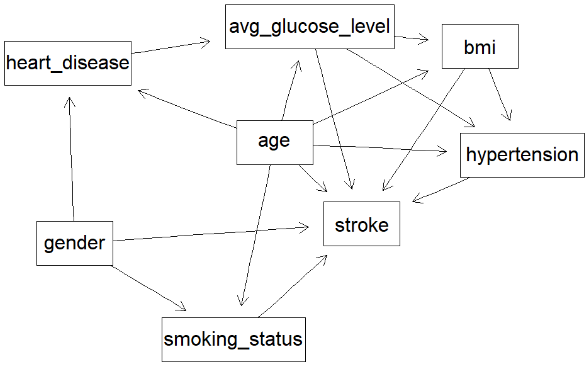

Figure 1: Bayes Network based on TABU-Algorithm and described relationships

The Bayesian Network adapted to our data achieved an Area Under the Curve (AUC) of 0.79 on the validation dataset. Figure 2 illustrates the ROC-Curve based on the validation dataset. The curve is significantly above the line of randomness, indicating that the classifier provides useful discrimination between the two classes. The line is not perfectly symmetrical, suggesting slightly better specificity than sensitivity, potentially due to the unbalanced dataset (only about 2.3% of the subjects experienced a stroke).

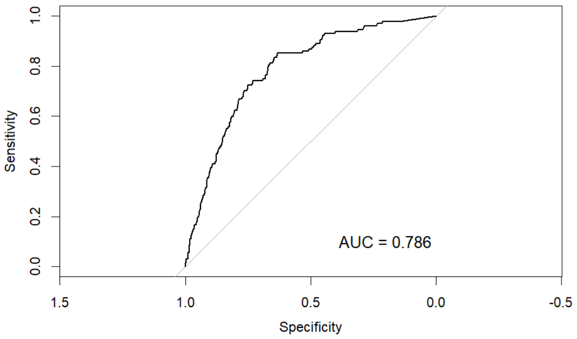

Figure 2: ROC-Curve of model performance on validation dataset

4 Conclusion and Discussion

This study examined the performance of Bayesian Networks in predicting strokes based on various risk factors. The achieved AUC value of 0.786 indicates a good predictive accuracy of the classifier. Overall, the ROC curve shows that the classifier performs significantly better than a random classifier, yet there is room for improvement. For instance, Yang et al. in their systematic review reported a pooled total AUC of 0.872 with a 95% CI of (0.862-0.88) (Yang et al., 2023).

Possible approaches to enhance performance include finer categorization of discrete and continuous variables and a more balanced dataset. Future work could compare the performance of Bayesian Networks against simpler methods such as logistic regression.

References

Mitchell S Elkind and Ralph L Sacco. 1998. Stroke risk factors and stroke prevention. In *Seminars in neurology*, volume 18, pages 429–440. © 1998 by Thieme Medical Publishers, Inc.

David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243.

Mark M. 2021. Replication Data for: Prediction of Cerebral Stroke.

Eunjeong Park, Hyuk-jae Chang, and Hyo Suk Nam. 2018. A bayesian network model for predicting post-stroke outcomes with available risk factors. *Frontiers in neurology*, 9:401721.

Wenzhu Song, Lixia Qiu, Jianbo Qing, Wenqiang Zhi, Zhijian Zha, Xueli Hu, Zhiqi Qin, Hao Gong, and Yafeng Li. 2022. Using bayesian network model with mmhc algorithm to detect risk factors for stroke. *Math. Biosci. Eng*, 19(12):13660–13674.

Yujia Yang, Li Tang, Yiting Deng, Xuzi Li, Anling Luo, Zhao Zhang, Li He, Cairong Zhu, and Muke Zhou. 2023. The predictive performance of artificial intelligence on the outcome of stroke: A systematic review and meta-analysis. *Frontiers in neuroscience*, 17:1256592.

Shuai Zhang, Wei Zhang, and Guangqian Zhou. 2019. Extended risk factors for stroke prevention. *Journal of the National Medical Association*, 111(4):447–456.