Data Science in Health: Graded Lab 1

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

Alessio Drigatti
March 31, 2024

Table of Content

Research Topic

How good is the **predictive performance** *of a Bayesian network in* **assessing stroke risk** *based on various risk factors?*

Dataset

- Replica dataset from Harvard Dataverse [1]
- 43'000 Subjects
- Target variales: age, gender, hypertension, heart disease, average glucose level, BMI, smoking status and stroke (dataset contains many more)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)
- Average glucose levels are categorized as: *Normal* ($<$ 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* ($<$ 126 mg/dl)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)
- Average glucose levels are categorized as: *Normal* ($<$ 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* ($<$ 126 mg/dl)
- BMI is categorized as: *Normal* ($<$ 30) and *Overweight* ($\geq$ 30)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)
- Average glucose levels are categorized as: *Normal* ($<$ 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* ($<$ 126 mg/dl)
- BMI is categorized as: *Normal* ($<$ 30) and *Overweight* ($\geq$ 30)
- Age is categorized as: *Young* (18 to 30), *Adult* (30 to 60), and *Senior* ($>$ 60)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)
- Average glucose levels are categorized as: *Normal* ($<$ 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* ($<$ 126 mg/dl)
- BMI is categorized as: *Normal* ($<$ 30) and *Overweight* ($\geq$ 30)
- Age is categorized as: *Young* (18 to 30), *Adult* (30 to 60), and *Senior* ($>$ 60)
- Train/Val-Split (80/20%)

Dataset

Preprocessing

- Removed persons $<$ 18 years old (17.4%)
- Removed observations with missing smoking status (29%)
- Average glucose levels are categorized as: *Normal* ($<$ 100 mg/dl); *Prediabetic* (100 to 126 mg/dl); *Diabetic* ($<$ 126 mg/dl)
- BMI is categorized as: *Normal* ($<$ 30) and *Overweight* ($\geq$ 30)
- Age is categorized as: *Young* (18 to 30), *Adult* (30 to 60), and *Senior* ($>$ 60)
- Train/Val-Split (80/20%)
- Resulting dataset has **28'685** observations

Bayesian Network

We know...

- Age, gender, diabetes, smoking, hypertension, weight, among others, are risk factors for stroke [2], [3]
- Age and gender is not influenced by other variables

Bayesian Network

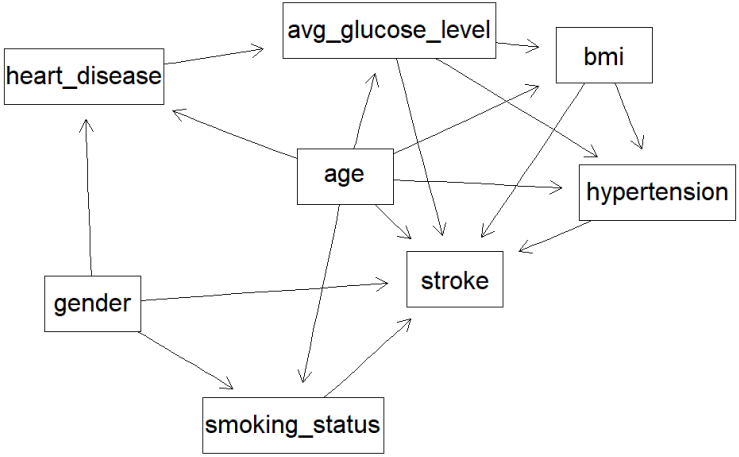
Whitelist

$$Hypertension \rightarrow Stroke \quad (1)$$
$$AverageGlucoseLevel \rightarrow Stroke \quad (2)$$
$$SmokingStatus \rightarrow Stroke \quad (3)$$
$$BMI \rightarrow Stroke \quad (4)$$
$$Age \rightarrow Stroke \quad (5)$$

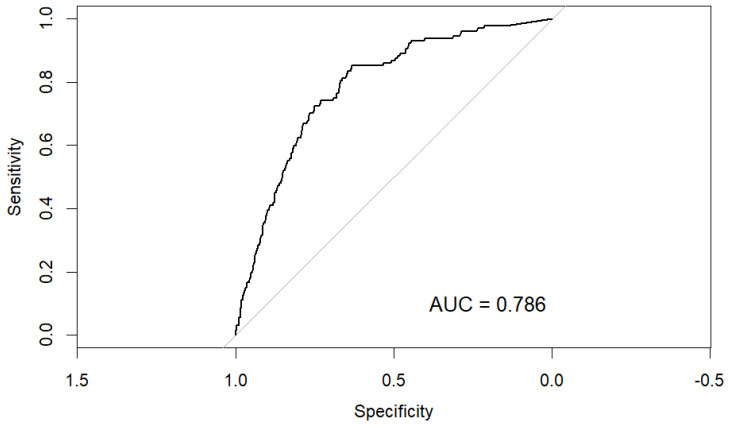
Blacklist

$$Any* \rightarrow Age \quad (6)$$
$$Any* \rightarrow Gender \quad (7)$$
$$Gender \rightarrow Age \quad (8)$$
$$Age \rightarrow Gender \quad (9)$$

Results


avg_glucose_level
bmi
heart_disease
age
hypertension
stroke
gender
smoking_status

Results


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.786
1.5
1.0
0.5
0.0
-0.5
Specificity

Conclusion and Outlook

1. AUC-Value of 0.786 is *okayish*

Conclusion and Outlook

1. AUC-Value of 0.786 is *okayish*
2. Room for improvement

Conclusion and Outlook

1. AUC-Value of 0.786 is *okayish*
2. Room for improvement
3. Yang et al. reported in their systematic review a pooled total AUC of 0.872 with a 95% CI of (0.862-0.88) [4].

Conclusion and Outlook

1. AUC-Value of 0.786 is *okayish*
2. Room for improvement
3. Yang et al. reported in their systematic review a pooled total AUC of 0.872 with a 95% CI of (0.862-0.88) [4].
4. Perhaps enhance performance with more balanced datase or a finer categorization numeric risk factors.

Conclusion and Outlook

1. AUC-Value of 0.786 is *okayish*
2. Room for improvement
3. Yang et al. reported in their systematic review a pooled total AUC of 0.872 with a 95% CI of (0.862-0.88) [4].
4. Perhaps enhance performance with more balanced datase or a finer categorization numeric risk factors.
5. Compare performance with other simple models e.g. Logistic regression

Resources I

[1] M. M, "Replication Data for: Prediction of Cerebral Stroke," version DRAFT VERSION, 2021. DOI: 10.7910/DVN/44RCPZ. [Online]. Available: https://doi.org/10.7910/DVN/44RCPZ.

[2] M. S. Elkind and R. L. Sacco, "Stroke risk factors and stroke prevention," in *Seminars in neurology*, © 1998 by Thieme Medical Publishers, Inc., vol. 18, 1998, pp. 429–440.

[3] S. Zhang, W. Zhang, and G. Zhou, "Extended risk factors for stroke prevention," *Journal of the National Medical Association*, vol. 111, no. 4, pp. 447–456, 2019.

[4] Y. Yang, L. Tang, Y. Deng, *et al.*, "The predictive performance of artificial intelligence on the outcome of stroke: A systematic review and meta-analysis," *Frontiers in neuroscience*, vol. 17, p. 1 256 592, 2023.