Stroke Prediction using Bayesian Modeling

Group 10
Zhu Yufan, Lu Yiting, Yu Zizhen

Abstract

Stroke remains a leading cause of mortality and morbidity across the globe. This project seeks to harness Bayesian modeling to enhance stroke prediction, offering precision and a measure of uncertainty. Drawing from health metrics and lifestyle data, our aim is a model not just predictive, but also reflective of the depth and robustness inherent in Bayesian methodologies.

November 2, 2023

1 Introduction

Globally recognized as a significant health challenge [24], stroke's sudden onset and potential for severe consequences demand early and effective prediction mechanisms [8]. Timely interventions can greatly reduce the severity of strokes, emphasizing the need for robust predictive models [16]. Recent advancements in machine learning and data-driven methodologies open avenues for enhanced prediction capabilities [4]. This project seeks to tap into Bayesian modeling, celebrated for its comprehensive approach of amalgamating prior knowledge and observed data [25]. In doing so, we consider a range of health determinants, from age and gender to hypertension and heart disease [5, 14]. While the ideal in healthcare would be a model embracing every possible risk factor [13], we've narrowed our focus to the dataset at hand, balancing depth with feasibility [17].

2 Related Works

Several studies have been conducted on stroke prediction using various statistical and machine-learning models. For instance, the Framingham Stroke Risk Profile (FSRP) is a widely used algorithm that incorporates multiple risk factors to predict the 10-year risk of stroke[18]. However, it has been criticized for its lack of precision and inability to account for interactions between risk factors[29].

Machine learning models such as decision trees, random forests, and neural networks have shown promising results in stroke prediction[2]. These models can handle high-dimensional data and capture complex interactions between variables. However, they often lack interpretability, which is crucial in healthcare applications[3].

In addition to these methods, Mainali et al.[19] discussed the application of machine learning in medicine over the past decade. They highlighted that commercially available machine learning algorithms have already been incorporated into clinical applications for rapid diagnosis. The creation and advancement of deep learning techniques have greatly improved clinical utilization of machine learning tools and new algorithms continue to emerge with improved accuracy in stroke diagnosis and outcome prediction.

Wang et al.[30] conducted a systematic review of machine learning models for predicting outcomes after stroke. They found that the use of machine learning for predicting stroke outcomes is increasing. However, few met basic reporting standards for clinical prediction tools and none made their models available in a way that could be used or evaluated.

Bayesian modeling has been used in various healthcare applications due to its ability to incorporate prior knowledge and provide probabilistic predictions[1]. However, its application in stroke prediction is relatively unexplored. This project aims to fill this gap by developing a Bayesian model for stroke prediction.

This paper aims to contribute to this growing field by developing a Bayesian model for stroke prediction, addressing some of the limitations identified in previous works.

3 Methodology

3.1 Data Collection

The dataset utilized in this study is sourced from Kaggle, titled "Stroke Prediction Dataset" [9]. This dataset aims to assist in predicting the likelihood of a patient experiencing a stroke based on various clinical and lifestyle-related attributes.

The dataset comprises 3245 observations, each providing pertinent information about individual patients. The attributes include:

Table 1: Attributes in the Stroke Prediction Dataset

Attribute	**Description**
id	A unique identifier for each patient.
gender	Categorical variable with values "Male", "Female".
age	Numerical variable indicating the age of the patient.
hypertension	Binary variable, where 0 indicates the absence and 1 indicates the presence of hypertension.
heart_disease	Binary variable, where 0 denotes no heart diseases and 1 signifies the presence of heart disease.
ever_married	Categorical variable with values "No" or "Yes".

Continued on next page

Table 1 – continued from previous page

Attribute	Description
work_type	Categorical variable indicating the type of employment - "Govt_job", "Private", or "Self-employed".
Residence_type	Categorical variable with values "Rural" or "Urban".
avg_glucose_level	Numerical variable representing the average glucose level in the patient's blood.
BMI	Numerical variable indicating the patient's body mass index.
smoking_status	Categorical variable with values "formerly smoked", "never smoked", and "smokes".
stroke	Binary target variable, where 1 indicates the patient had a stroke, and 0 indicates they did not.

3.2 Data Cleaning

The initial dataset, obtained from Kaggle, necessitated a meticulous data-cleaning process due to the presence of missing or ambiguous values in several features.

Firstly, to maintain the integrity of the dataset, rows with missing values in the 'BMI' column were eliminated. This step was crucial as 'BMI' is often a significant predictor in health-related studies and its absence could lead to inaccurate model predictions.

Secondly, the dataset included entries with an 'Unknown' smoking status. These entries could potentially introduce noise into the model, thereby affecting its performance. To prevent this, such rows were discarded.

Lastly, inconsistencies in the 'gender' column were rectified by assigning integer values to the gender categories ('Male' as 0 and 'Female' as 1). Furthermore, rows with undefined gender were removed to ensure clarity and consistency in the dataset.

3.3 Feature Transformation

The data underwent several transformations and encodings to prepare it for modeling.

The 'id' column, which uniquely identifies each patient, was dropped as it does not contribute to the model's predictive power. The 'gender' feature was cleaned by mapping 'Male' to 0 and 'Female' to 1 and removing rows with undefined gender. The 'age' feature was generalized by binning age values into decades.

Binary columns like 'hypertension' and 'heart_disease', which indicate the presence of hypertension or heart disease in the patient, were already suitable for modeling. The binary feature 'ever_married', indicating marital status, was encoded with 'Yes' as 1 and 'No' as 0.

The non-ordinal categorical feature 'work_type', describing the patient's job type, was transformed into two separate binary columns: 'is_private' for private jobs and 'is_self_employed' for self-employment. The original column was then dropped.

The 'Residence_type' feature, specifying the patient's residence type, was binary encoded with 'Urban' as 1 and 'Rural' as 0. The continuous feature 'avg_glucose_level', representing the average glucose level in the patient's blood, was simplified by binning it into three categories: low (below 90), normal (90 to 130), and high (above 130).

The Body Mass Index of the patient, represented by the 'BMI' feature, was categorized into standard BMI classifications: underweight, normal weight, overweight, and obese. The ordinal feature 'smoking_status', indicating the patient's smoking habits, was mapped to ordinal values.

The primary target variable 'stroke' was addressed by oversampling rows where 'stroke' is 1 seventeen times to balance the dataset representation. Finally, to ensure randomness and reduce overfitting risk, the dataset was shuffled with a fixed seed for reproducibility.

3.4 Modeling

3.4.1 Bayesian Modeling

In our pursuit to predict strokes, we employed Bayesian modeling, a probabilistic approach that integrates both prior knowledge and observed data. The structure of our Bayesian network was meticulously designed based on the intricate relationships and dependencies among various health metrics and lifestyle features present in the dataset.

The Bayesian network's architecture integrates both empirical data and established medical knowledge, ensuring its credibility:

The risk of hypertension, heart_disease, and stroke escalates with age, a correlation consistently documented in medical studies [15, 22]. Gender plays a pivotal role in shaping smoking habits and directly correlates with stroke risk, as evidenced by numerous research findings [26, 27]. Established predictors of stroke, such as hypertension, heart_disease, marital status (ever_married), residence_type, employment status (is_private and is_self_employed), avg_glucose_level, and BMI have been duly integrated into our model [10, 21, 28]. Furthermore, a substantiated link between BMI and conditions like hypertension and heart_disease has been considered [12, 11]. Additionally, the model acknowledges the adverse ramifications of smoking_status on hypertension, heart_disease, and stroke risk, drawing from extensive scientific literature [7, 23, 20, 6].

By incorporating these evidence-based relationships into our Bayesian network, we have ensured that our model is not only data-driven but also grounded in scientific knowledge. This lends credibility to our model's structure and reinforces the robustness of the relationships we've introduced. As such, our model serves as a reliable tool for stroke prediction.

The graphical structure of the Bayesian network is illustrated below:

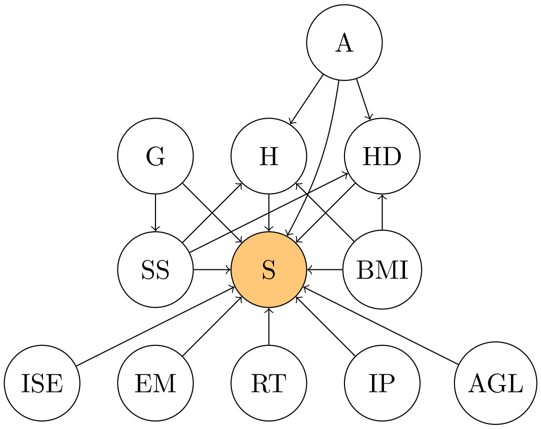

Figure 1: Hierarchical structure of the Bayesian network for stroke prediction.

Legend:

Symbol	Description	Symbol	Description	Symbol	Description
A	age	G	gender	H	hypertension
HD	heart_disease	S	stroke	EM	ever_married
RT	residence_type	AGL	avg_glucose_level	BMI	BMI
SS	smoking_status	IP	is_private	ISE	is_self_employed

Table 2: Legend for the Symbols

Once the structure was defined, we trained the Bayesian model using the Bayesian Estimator with the BDeu prior. After training, the learned Conditional Probability Distributions (CPDs) for each node, such as the "stroke" node, enabled our model to predict stroke likelihood based on various input features.

In addition to our primary Bayesian model, we also developed a simpler and more complex version for comparison. The streamlined model reduces interconnections, while the advanced one introduces added complexities. The intention behind these variants is to measure their performance against our main model. A thorough analysis of these models is presented in Section 4.2.

3.4.2 Neural Networks

Briefly introduce the architecture, training process, and tuning for the neural network model.

3.4.3 Reinforcement Learning

If used for prediction or any other aspect, discuss its architecture and application.

4 Performance Evaluation

4.1 Evaluation Metrics

To comprehensively assess the model's adequacy, we employ the following metrics:

- **Accuracy**: Represents the fraction of correct predictions. Simple but potentially misleading for imbalanced datasets.
- **Precision & Recall**:
  - **Precision**: Measures how many identified positives are genuinely positive. Especially relevant when the cost of false positives is high.
  - **Recall** (Sensitivity): Quantifies how many actual positives our model correctly identifies. It becomes essential when the implications of missing a positive are significant.
- **F1-score**: The harmonic mean of precision and recall, offering a balanced metric, especially in the presence of imbalanced datasets.
- **ROC Curve**: A graphical representation of a model's true positive rate against its false positive rate. The associated **AUC** (Area Under the Curve) provides a scalar measure of performance.

4.2 Comparison of Bayesian Networks with Different Complexities

Detail the differences in performance, interpretability, and other relevant aspects when comparing Bayesian networks of varying complexities.

4.2.1 Model Performance: Proposed Bayesian Network

- **Accuracy**: The proposed model exhibits an impressive accuracy of 97.07%, implying that it correctly predicts the stroke outcome for 97.07% of the test cases.
- **Precision Recall**:

Metric	No Stroke	Stroke
Precision	1.00 (100%)	0.94 (94%)
Recall	0.94 (94%)	1.00 (100%)

Table 3: Precision and Recall for the Proposed Bayesian Network Model

- **F1-score**: The model has an F1-score of 0.97 0.97 for both "No Stroke" and "Stroke" predictions. This harmonized score showcases balanced precision and recall, especially vital given the critical nature of stroke predictions.
- **ROC Curve**:

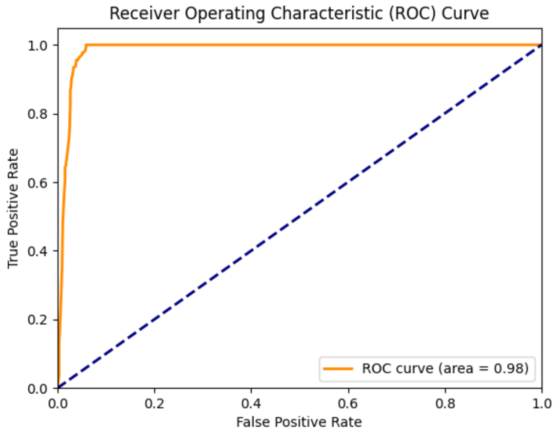

Figure 2: Enter Caption

The Receiver Operating Characteristic (ROC) curve provides insights into the trade-off between a true positive rate and a false positive rate. A model with perfect prediction capability will have an Area Under the Curve (AUC) of 1. Our model's AUC is 0.98 suggesting it has an outstanding discriminative power between positive and negative classes.

4.2.2 Model Performance: Simpler Bayesian Network

- **Accuracy**: The simpler model has an accuracy of 90.36%, which is slightly lower than the 97.07% accuracy of the proposed model.
- **Precision & Recall**:

Metric	**No Stroke**	**Stroke**
Precision	0.99 (99%)	0.84 (84%)
Recall	0.81 (81%)	0.99 (99%)

Table 4: Precision and Recall for the Simpler Bayesian Network Model

Compared to the proposed model, the simpler model has a slightly lower precision for predicting ”Stroke” and a lower recall for predicting ”No Stroke”. This suggests that while both models are effective, the proposed model is more balanced in terms of precision and recall.

- **F1-score**: The F1-score for the simpler model is slightly lower than that of the proposed model for both ”No Stroke” and ”Stroke” predictions. This indicates that the proposed model provides a better balance between precision and recall.
- **ROC Curve**:

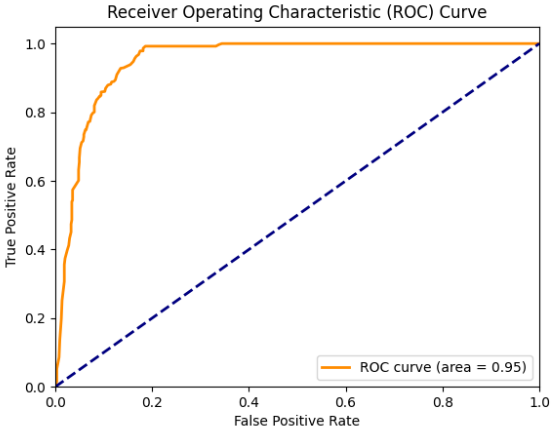

Figure 3: Enter Caption

The Area Under the Curve (AUC) for the simpler model is 0.95, which is slightly lower than the AUC of 0.98 for the proposed model. This suggests that while both models have excellent discriminative power between positive and negative classes, the proposed model performs slightly better.

4.2.3 Model Performance: More Complex Bayesian Network

- **Accuracy**: The more complex model has an accuracy of 79.80%, which is lower than both the simpler model (90.36%) and the proposed model (97.07%). This suggests that the additional complexity may be leading to overfitting, reducing its ability to generalize to unseen data.

- **Precision & Recall**:

Metric	No Stroke	Stroke
Precision	0.84 (84%)	0.77 (77%)
Recall	0.74 (74%)	0.86 (86%)

Table 5: Precision and Recall for the More Complex Bayesian Network Model

Compared to the proposed model, the more complex model has lower precision and recall for both "No Stroke" and "Stroke" predictions. This indicates that the model may be overfitting to the training data, reducing its predictive performance on the test data.

- **F1-score**: The F1-score for the more complex model is lower than that of both the simpler and proposed models for both "No Stroke" and "Stroke" predictions. This further suggests that the additional complexity may be detrimental to the model's performance.

- **ROC Curve**:

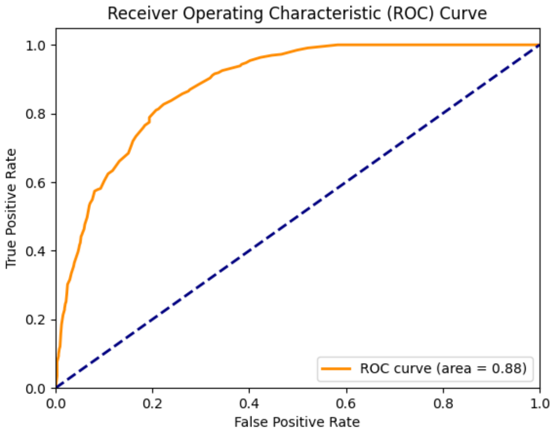

Figure 4: Enter Caption

The Area Under the Curve (AUC) for the more complex model is 0.88, which is lower than both the simpler model (0.95) and the proposed model (0.98). This indicates that while the more complex model still has good discriminative power between positive and negative classes, it is outperformed by the simpler and proposed models.

4.3 Bayesian Network vs. Other Methods

Discuss the performance, advantages, and disadvantages of the Bayesian network in comparison with other methodologies you've utilized, such as neural networks, reinforcement learning, etc.

5 Limitation and Possible Future Enhancement

5.1 Limitations

- **Model Combinations**: Given the myriad ways Bayesian Networks can be structured, it is infeasible to compare all possible combinations within the stipulated time frame.

- **Comparative Analysis**: Time constraints also limited our ability to juxtapose our Bayesian Network with an array of alternative models, potentially missing out on identifying superior modeling techniques.

- **Data Constraints**: The dataset, while robust, may not capture all nuances and variables that influence stroke risk. There might be lurking variables or interactions not considered.

- **Generalizability**: Our model, being data-driven, is based on the specific dataset we used. It might not generalize perfectly to different demographics or regions.

- **Simplifying Assumptions**: Bayesian networks inherently make conditional independence assumptions that might oversimplify real-world intricacies.

5.2 Possible Future Enhancements

- **Expanded Comparisons**: Future work could include comparisons with other probabilistic models, deep learning techniques, and ensemble methods to further improve prediction accuracy.

- **Incorporate Additional Data**: Introducing more variables, perhaps from supplementary datasets or medical literature, might enhance the network's predictive capabilities.

- **Dynamic Bayesian Networks**: Transitioning to dynamic Bayesian networks could better model time-dependent changes and capture temporal relationships among variables.

- **Regularization Techniques**: Implementing regularization can help prevent overfitting, especially as the network's complexity grows.

- **User-friendly Interface**: Developing an intuitive, user-friendly interface could make the model more accessible to healthcare professionals for real-time risk assessment.

6 Responsible AI: State of the Art

6.1 Transparency

Transparency in AI, particularly in the healthcare domain, is pivotal. In our project, using Bayesian networks, we inherently favor transparency. The probabilistic relationships between nodes are interpretable, allowing medical professionals and patients to understand the basis of predictions. This clear understanding can foster trust and enable informed decision-making, rather than treating the model as a black-box.

6.2 Privacy

Handling medical data always raises valid privacy concerns. While our model requires certain health metrics and lifestyle data, it's crucial to ensure that data sourcing respects patient confidentiality. In this project, data from Kaggle was used, assuming prior anonymization. In real-world applications, data should be aggregated and anonymized, ensuring no direct patient identification from the model's inputs or outputs.

6.3 Fairness

Fairness in AI models ensures that predictions don't inadvertently favor or discriminate against any particular group. In the context of our project, biases in data, such as underrepresentation of certain age groups or genders, could lead to skewed predictions. It's essential to ensure that the data adequately represents the diverse population, ensuring equitable stroke predictions across various demographic groups.

6.4 Safety and Robustness

In the realm of healthcare, model safety and robustness are paramount. Misclassifications or false predictions can have direct health implications. Our Bayesian model's high accuracy indicates its reliability. However, continuous validation against new data and real-world scenarios is necessary to ensure that it maintains its robustness and doesn't produce harmful false predictions.

6.5 Human Factors

While our Bayesian model can provide valuable predictions, the human factor can't be sidelined. Medical professionals should use the model as a supplementary tool, combining its insights with their expertise. Relying solely on model predictions without human oversight might overlook nuances or exceptional cases that clinicians can identify.

6.6 Governance and Regulation

As AI becomes increasingly integrated into healthcare, there's a pressing need for robust governance and regulatory frameworks. Our project, while academically oriented, should in real-world applications adhere to local and international health data standards and AI ethics guidelines. Regular audits, ethical reviews, and ensuring the model's adherence to evolving regulations can ensure its responsible and ethical use.

In conclusion, while our project showcases the potential of Bayesian networks in stroke prediction, it's essential to understand and address the myriad of ethical considerations that come with AI's territory in healthcare. Responsible AI is not just about creating effective models but ensuring they are used in ways that prioritize human well-being, fairness, and justice.

7 Conclusion

In this study, we delved into the potential of Bayesian Networks to predict stroke risk, harnessing both prior knowledge and observed data. The results showcased the efficacy of our chosen model, underpinned by both data-driven insights and established medical literature. While the model exhibited promising accuracy and robustness, we also identified areas for improvement and potential future enhancements. The intersection of machine learning and healthcare holds immense promise, and our work underscores the pivotal role Bayesian modeling can play in advancing predictive healthcare. As we move forward, refining and expanding upon this foundation could pave the way for more nuanced and reliable medical predictions, ultimately aiding in better patient outcomes and preventive healthcare strategies.

References

[1] Bayesian modeling in healthcare: A mini-review and some perspectives https://www.sciencedirect.com/science/article/pii/S2352914820300064

[2] Stroke prediction models: A systematic review https://www.ijser.org/researchpaper/Stroke-Prediction-Models-A-Systematic-Review.pdf

[3] Stroke risk prediction with machine learning techniques https://www.mdpi.com/1424-8220/22/13/4670

[4] Early prediction of ischemic stroke using machine learning boosting (2022), https://ieeexplore.ieee.org/document/10205861/

[5] Association, A.H.: How high blood pressure can lead to stroke (2022), https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-stroke

[6] Association, A.H.: Smoking, high blood pressure and your health - american heart association (2023), https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/smoking-high-blood-pressure-and-your-health

[7] CDC: Heart disease and stroke — smoking and tobacco use — cdc (2023), https://www.cdc.gov/tobacco/basic_information/health_effects/heart_disease/index.htm

[8] Clinic, M.: Hypertensive crisis: What are the symptoms? (2022), https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/hypertensive-crisis/faq-20058491

[9] Fedesoriano: Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (2021)

[10] GitHub: Exploratory data analysis of stroke dataset in r (2023), https://github.com/djthorne333/Exploratory-Data-Analysis-of-Stroke-Dataset-in-R

[11] Health, B.P.: Long-term body mass index changes in overweight and obese adults and the risk of heart failure, cardiovascular disease and mortality: a cohort study of over 260,000 adults in the uk (2021), https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-10606-1

[12] in Health, L., Disease: Body mass index is associated with blood pressure and vital capacity in medical students (2023), https://lipidworld.biomedcentral.com/articles/10.1186/s12944-023-01920-1

[13] Health, P.D.: Ideal algorithms in healthcare: Explainable, dynamic, precise ... (2022), https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000006

[14] Healthline: What is the risk of having a stroke by age range? (2022), https://www.healthline.com/health/age-range-for-stroke

[15] Healthline: How are hypertension, heart disease, and stroke related? (2023), https://www.healthline.com/health/high-blood-pressure-hypertension/how-are-hypertension-heart-disease-and-stroke-related

[16] Kaur, M., et al.: Early stroke prediction methods for prevention of strokes (2022), https://www.hindawi.com/journals/bn/2022/7725597/

[17] Library, I.X.D.: Heart disease prediction using stacking model with balancing techniques ... (2022), https://ieeexplore.ieee.org/document/10287281

[18] Mainali, S., Darsie, M.E., Smetana, K.S.: Machine learning in action: Stroke diagnosis and outcome prediction. Frontiers in Neurology **12** (2021), https://www.frontiersin.org/articles/10.3389/fneur.2021.734345/full

[19] Mainali, S., Darsie, M.E., Smetana, K.S.: Machine learning in action: Stroke diagnosis and outcome prediction. Frontiers in Neurology **12** (2021), https://www.frontiersin.org/articles/10.3389/fneur.2021.734345/full

[20] Medicine, J.H.: Smoking and cardiovascular disease — johns hopkins medicine (2023), https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease

[21] medRxiv: From tabulated data to knowledge graph: A novel way of improving the performance of the classification models in the healthcare data (2021), https://www.medrxiv.org/content/medrxiv/early/2021/06/12/2021.06.09.21258123.full.pdf

[22] Network, J.: Age-adjusted mortality rates and age and risk–associated contributions to change in heart disease and stroke mortality (2023), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2790434

[23] NIH: Scientists investigate link between smoking and high blood pressure — nhlbi, nih (2021), https://www.nhlbi.nih.gov/news/2021/scientists-investigate-link-between-smoking-and-high-blood-pressure

[24] Organization, W.H.: World stroke day (2021), https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

[25] Paolucci, I., et al.: Bayesian parametric models for survival prediction in medical applications (2023), https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02059-4

[26] Research, C.: Smoking and gender (2002), https://academic.oup.com/cardiovascres/article/53/3/568/325628

[27] Today, M.N.: Can smoking cause a stroke? risk, statistics, and more (2023), https://www.medicalnewstoday.com/articles/can-smoking-cause-a-stroke

[28] Vidhya, A.: How to create a stroke prediction model? (2021), https://www.analyticsvidhya.com/blog/2021/05/how-to-create-a-stroke-prediction-model/

[29] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D., Bray, B.: A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLOS ONE **15**(6) (2020), https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0234722

[30] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D.A., Bray, B.: A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLOS ONE **15**(6) (2020), https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0234722