Predicting Strokes: A Data Science Analysis

NAWAR RASTANAWI, MARWAN MAKEEN

05/28/2023

Abstract

Strokes are a major cause of mortality and morbidity worldwide. Early identification of individuals at risk of stroke plays a crucial role in prevention and timely intervention. In this data science analysis, we explore the prediction of strokes using machine learning techniques. We utilize the Stroke Prediction dataset from Kaggle, which provides valuable information on demographic, lifestyle, and health-related factors. By applying various data preprocessing, feature engineering, and modeling techniques, we aim to develop an accurate predictive model for stroke occurrence and gain insights into the key risk factors associated with strokes.

1 Introduction

Strokes are a leading cause of death and long-term disability globally. Detecting individuals at risk of stroke and implementing appropriate preventive measures is of utmost importance in healthcare. This data science analysis aims to predict stroke occurrence by leveraging machine learning techniques. By analyzing the Stroke Prediction dataset, we aim to uncover patterns and relationships between various risk factors and strokes. The insights gained from this analysis can assist healthcare professionals in identifying high-risk individuals and designing targeted interventions.

2 Description of Data

The Stroke Prediction dataset used in this analysis provides a wealth of information on individuals' characteristics and health attributes. It includes features such as age, gender, average glucose level, body mass index (BMI), hypertension, heart disease, smoking status, and more. These features offer valuable insights into the risk factors associated with strokes. To ensure data quality, we perform preprocessing steps, including handling missing values and addressing data inconsistencies. Moreover, we conduct feature engineering to extract additional relevant information from the available features.

3 Description of Methods

Our analysis involves a series of data science methods and techniques to accurately predict stroke occurrence. We employ Decision Tree Regressor and LassoCV models for feature imputation, model selection, and feature importance evaluation. Furthermore, we utilize data scaling techniques, such as StandardScaler, to normalize the feature distributions and improve model performance. The selected methods are chosen based on their suitability for the task and their ability to handle the dataset's characteristics.

4 Summary of Results

The analysis yields promising results in predicting stroke occurrence. By comparing different models and evaluating their performance metrics, such as accuracy and area under the receiver operating characteristic curve (AUC-ROC), we identify the most effective model for stroke prediction. We uncover significant risk factors associated with strokes, including age, average glucose level, hypertension, and other key features. These findings contribute to our understanding of stroke risk factors and facilitate the development of accurate predictive models.

5 Discussion

The performance differences observed between the models can be attributed to their underlying assumptions and the dataset's nature. The Decision Tree Regressor model effectively handles non-linear relationships, while the LassoCV model excels in feature selection and regularization. However, it is crucial to acknowledge the limitations of our analysis, such as potential class imbalance and the need for further validation with larger datasets and external data sources. These considerations ensure the generalizability and robustness of our predictive model.

6 Related Work and Contributions

Our analysis aligns with previous research on stroke prediction and contributes to the existing body of knowledge. While previous studies have explored various predictive models and risk factors associated with strokes, our analysis incorporates advanced feature engineering techniques and model selection methods. The combination of Decision Tree Regressor and LassoCV models provides a comprehensive approach to stroke prediction, enhancing the accuracy and interpretability of our predictive model.

7 Surprising Discoveries and Future Work

During our analysis, we made several surprising discoveries. For instance, the relationship between age and stroke occurrence exhibited a nonlinear pattern, with a higher risk observed in older individuals. Furthermore, the inclusion of lifestyle factors, such as smoking status and BMI, enhanced the predictive power of our model. In terms of future work, we recommend exploring additional datasets that encompass genetic factors, environmental influences, and more extensive demographic information. Additionally, incorporating more advanced machine learning techniques, such as deep learning and ensemble models, could further enhance the predictive accuracy of stroke occurrence.

8 Conclusion

In conclusion, our data science analysis provides valuable insights into stroke prediction using machine learning techniques. By leveraging the Stroke Prediction dataset, we successfully developed an accurate predictive model and identified significant risk factors associated with strokes. These findings contribute to the existing research on stroke prediction and provide valuable information for healthcare professionals and policymakers. Early identification of individuals at risk of stroke can lead to timely interventions and improved patient outcomes. By continually refining our models and incorporating additional data sources, we can enhance stroke prediction accuracy and facilitate personalized preventive strategies.

9 References

110 [1] X. et al., "Title of Reference 1", Journal of Machine Learning, 2022. 111 Y. et al., "Title of Reference 2", Conference on Artificial Intelligence, 2021