AIT 664 Module 8 Project Deliverable 4
G01394251

BRAIN STROKE PREDICTION

[REPORT]

Throughout this analysis focused on brain stroke prediction, a comprehensive journey unfolded, aiming to explore various demographic and health-related attributes to predict stroke occurrences. In this final part of the project, I will recap the initial hypotheses, the analysis results, the overall approach, and the lessons learned from this semester-long effort.

Initial Hypothesis:

The hypotheses centered around certain demographic and health-related factors being significant predictors of stroke occurrence. These included age, hypertension status, average glucose levels, and smoking history**.**

(hypothesis recap from deliverable 1)

- Advanced age is positively associated with an increased risk of stroke.
- Individuals with a history of hypertension are more likely to experience strokes.
- Elevated average glucose levels are correlated with a higher likelihood of stroke.
- Individuals with smoking history are also associated with higher chances of stroke.

Initial hypothesis will serve as a guiding framework for our investigation into the dataset. We will explore the data to determine if there is substantial evidence to support or refute these hypotheses. Additionally, we will consider alternative hypotheses to ensure a comprehensive and rigorous analysis. The results of our analysis will help validate or refine these hypotheses and contribute to a better understanding of stroke prediction and risk factors.

Approach Overview:

The analysis journey encompassed multiple stages. It began with data acquisition from Kaggle, followed by meticulous data preparation, cleaning, and exploratory data analysis (EDA) to derive crucial insights. Preprocessing involved balancing the data by performing down-sampling to address class imbalance. Model selection, tuning, and evaluation were executed using a Random Forest Classifier.

Demonstration of Hypothesis:

The analysis largely validated the initial hypotheses. Age, hypertension, high glucose levels, and smoking history emerged as influential factors in stroke occurrences. The refined model demonstrated the strength of these predictors in identifying individuals at risk of strokes.

Lessons Learned:

Several critical lessons surfaced throughout the analysis process.

Reflections on Challenges and Successes:

- Challenges Faced: Dealing with missing data, ensuring data quality, and interpreting the dataset's nuances were primary challenges. Additionally, managing class imbalance in the dataset required specialized techniques.
- Successes: Successful data cleaning, normalization, and feature engineering contributed to meaningful insights. Moreover, the enhancement of predictive models through hyperparameter tuning was a success point.
- Unexpected Discoveries: Some unexpected correlations emerged, such as the association between marriage and increased stroke risk, shedding light on non-traditional factors affecting health outcomes.

Overall Approach:
The analysis encompassed:

1. Comprehensive data collection and preprocessing.
2. Exploratory data analysis highlighting critical associations.
3. Balancing data for modeling accuracy through down-sampling.
4. Model deployment using Random Forest Classifier and hyperparameter tuning.

Effectiveness of Analysis:
The analysis effectively demonstrated correlations between various attributes and stroke occurrence. The predictive model exhibited promising accuracy and precision, laying a strong foundation for stroke risk prediction.

Demonstration of Hypothesis:
The results validated initial hypotheses, showcasing clear relationships between age, hypertension, glucose levels, and stroke incidence.

Improvements and Future Recommendations that I would consider:

- **Enhanced Data Documentation**: Comprehensive documentation of data sources, collection methods, and quality control measures is crucial for future analyses.
- **Further Model Refinement:** Exploring additional algorithms and refining feature engineering methods could enhance predictive models further.
- **Long-term Outcome Analysis:** A longitudinal study tracking stroke occurrences over time could offer more comprehensive insights into causation and recurrence.
- **Collaborative Research:** Collaboration with medical professionals could provide domain-specific insights and enhance practical applicability in healthcare scenarios.

CONCLUSION:

The analysis effectively explored relationships between demographic, health-related factors, and stroke occurrences. Findings validated initial hypotheses, highlighting the role of age, health conditions, and lifestyle factors in stroke prediction. Addressing challenges, leveraging

successes, and outlining future improvements set the stage for continued exploration and refinement in understanding and preventing strokes.

REFERENCE:

[1] Akbasli, I. T. (2022, July 16). *Brain stroke prediction dataset*. Kaggle. https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset