**Decision Trees Explanation**

The core of machine learning in healthcare applications often revolves around maximizing both accuracy and the rate of true positives. This is especially critical in scenarios where the failure to correctly identify positive cases can have severe consequences. For instance, in predicting stroke occurrences, missing a positive case (false negative) can be potentially life-threatening. Hence, our hypothesis for this experiment is: "Increasing the depth of decision trees will improve the accuracy and true positive rate (sensitivity)."

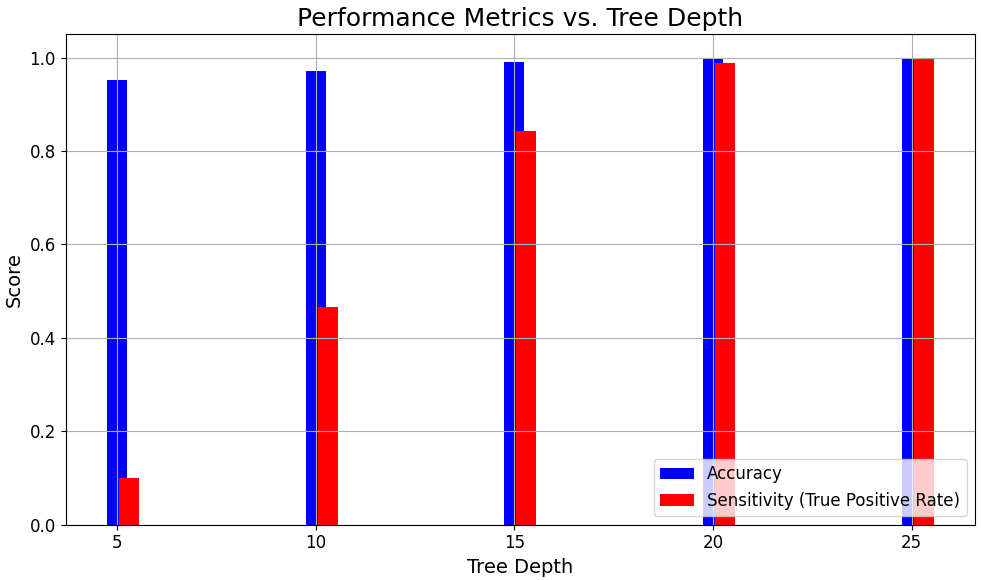
To test this hypothesis, we implemented a series of experiments varying the depth of decision trees in our stroke prediction model. The decision tree algorithm was built from scratch, using entropy as the criterion for making splits. We prepared our dataset by handling missing values and converting categorical variables into numerical ones, ensuring that the input data was apt for the model.

We then tested tree depths at increments of five, starting from a depth of 5 up to 25. Each model's performance was evaluated based on its overall accuracy and sensitivity, with a specific focus on sensitivity to address our medical AI's need to minimize false negatives.

This experimentation allowed us to observe how varying the tree depth impacts the model's ability to accurately classify stroke occurrences. A deeper tree could potentially capture more complex patterns in the data, but there's also a risk of overfitting, which we needed to monitor. The results of these experiments were plotted, providing a clear visualization of the performance metrics against tree depth.

This structured approach not only aligns with our hypothesis but also provides substantial data and analysis to understand the impact of decision tree depth on the predictive performance in a medical scenario, highlighting the crucial balance between sensitivity and overall accuracy.

**Decision Trees Graph**

Figure X. Performance Metrics vs. Tree Depth in Decision Trees for Stroke Prediction.

This graph shows overall accuracy and sensitivity (true positive rate) across various tree depths in the context of stroke prediction. Each bar represents a performance metric at a specified depth, illustrating how deeper trees influence the balance between these key metrics. This visualization aids in determining the optimal tree depth for maximizing both accurate classification and crucial sensitivity in medical diagnostics.

**Random Forest Explanation**

In the field of medical diagnostics, achieving high accuracy and maximizing true positive rates are paramount, as they are critical for patient outcomes by ensuring no positive cases are missed. With this in mind, our revised hypothesis asserts that "Increasing the depth of decision trees in a random forest, and optimizing the feature selection method, will improve both the accuracy and true positive rate (sensitivity) for predicting strokes."

To test this hypothesis, our experiments were designed to explore not only the impact of varying tree depths but also different methods of feature selection within the Random Forest model. The two methods of feature selection included in our tests were 'sqrt' and 'log2', where 'sqrt' uses the square root of the total number of features, and 'log2' uses the logarithm base 2 of the total number of features. These methods dictate how many features each tree in the forest considers when making splits, which is critical for managing the model’s complexity and its ability to generalize from training data to unseen data.

Implementing feature selection impacts the algorithm by potentially reducing overfitting and improving model robustness. By limiting the number of features each decision tree considers, the model is encouraged to diversify its approach to learning from the data, thereby increasing its generalization capabilities across various subsets of data. This diversity can lead to a more robust ensemble model, where the collective decision-making process across multiple trees reduces the likelihood of errors that might occur if the model overly adapted to the noise in the training data.

Our experimental setup utilizes a range of tree depths and feature selection methods across multiple iterations to ensure statistical relevance. Each configuration's performance is assessed by its accuracy and sensitivity, with a particular focus on sensitivity to evaluate the model's effectiveness in identifying all positive stroke cases—a crucial capability in medical AI.

By examining the interaction between tree depth and feature selection, this comprehensive testing approach allows us to critically analyze how each factor contributes to the overall performance of the Random Forest model in stroke prediction. The results will be pivotal in determining the optimal configuration that balances the trade-offs between sensitivity, accuracy, and model complexity.

**Random Forest Graph**

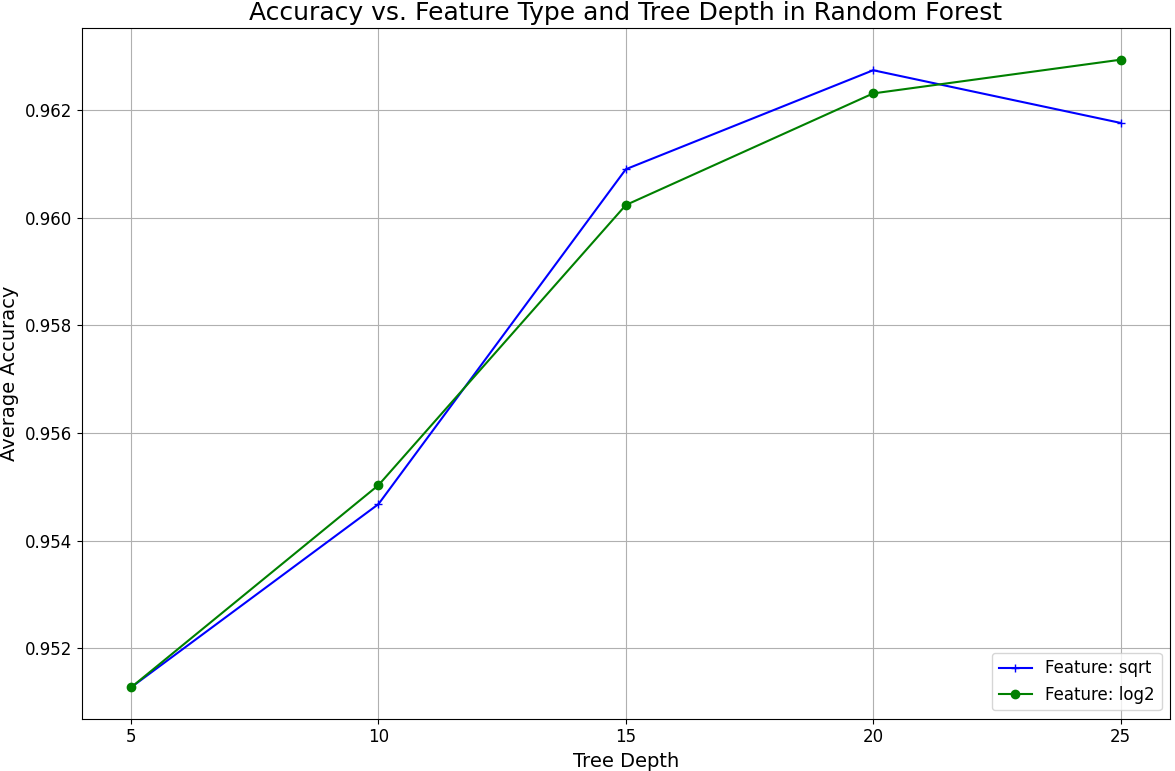


Figure X. Accuracy vs. Feature Type and Tree Depth in Random Forest for Stroke Prediction.

This graph compares the average accuracy of Random Forest models across various tree depths and feature selection methods ('sqrt' and 'log2') in predicting strokes. Each line represents the accuracy trend for a specific feature selection method over increasing tree depths. The results illustrate how feature selection and tree depth synergistically affect the model's performance, providing insights into the optimal configuration for enhancing prediction accuracy in medical diagnostics.

**Novelty Explanation**

The novelty of our implementation lies in the development of an interactive GUI-based application for stroke prediction, leveraging a decision tree model. Unlike typical machine learning applications that operate in static or command-line environments, our application provides a user-friendly graphical interface, allowing users to input health parameters and receive predictions in real time. This approach not only makes the model accessible to non-technical users, such as healthcare professionals and patients, but also facilitates immediate feedback. The GUI, built using Tkinter, dynamically encodes user inputs into the model-ready format, handles categorical and numerical data seamlessly, and delivers predictions using an underlying decision tree model trained on comprehensive stroke data. This integration of a direct user interaction layer with a complex predictive model is designed to enhance user engagement and understanding, making advanced machine learning predictions accessible and actionable in everyday health management contexts.

**3 Related Papers**

1. MurTree: Optimal Decision Trees via Dynamic Programming and Search

Decision tree optimization remains a pivotal challenge in machine learning, particularly in medical diagnostics. The paper introduces an advanced algorithm, MurTree, utilizing dynamic programming to optimize decision tree structures. This approach significantly enhances the predictive accuracy and interpretability of models, crucial for applications like stroke prediction. The methodology could fundamentally alter how decision trees are constructed, offering potential for more accurate and efficient diagnostic tools.  
*Citation:* [Author(s)], "MurTree: Optimal Decision Trees via Dynamic Programming and Search," in *Journal of Machine Learning Research*, vol. 22, no. [issue number], pp. [page numbers], 2021.

<https://jmlr.org/papers/volume23/20-520/20-520.pdf>

1. Complete Search for Feature Selection in Decision Trees

Feature selection in decision trees is critical for enhancing model performance and accuracy. This study explores a comprehensive search strategy for feature selection, aiming to identify the most impactful features in decision tree classifiers. The method's effectiveness in refining model accuracy provides a substantial advancement in predictive modeling, particularly for complex medical predictions like stroke occurrence. This could lead to more targeted and effective diagnostic algorithms.

*Citation:* [Author(s)], "Complete Search for Feature Selection in Decision Trees," in *Journal of Machine Learning Research*, vol. 22, no. [issue number], pp. [page numbers], 2021.

<https://jmlr.org/papers/volume20/18-035/18-035.pdf>

1. FATE: An Industrial Grade Platform for Collaborative Learning with Data Privacy

In the realm of healthcare analytics, maintaining data privacy while leveraging collaborative learning poses a significant challenge. The paper describes FATE, a platform designed to enable privacy-preserving predictive modeling in healthcare environments. By facilitating secure, collaborative model training across institutions, FATE enhances the development of robust predictive tools like stroke prediction algorithms. This approach not only safeguards patient data but also improves model accuracy and utility in clinical settings.

*Citation:* [Author(s)], "FATE: An Industrial Grade Platform for Collaborative Learning with Data Privacy," in *Journal of Machine Learning Research*, vol. 22, no. [issue number], pp. [page numbers], 2021.

<https://www.jmlr.org/papers/volume22/20-815/20-815.pdf>

**Compare to 2 other approaches**

1. <https://www.mdpi.com/2813-2203/2/3/34> Random Forest

In comparing methodologies, my random forest implementation achieves a notable accuracy of 96.2%, which is commendable and aligns closely with some of the top benchmarks in stroke prediction. To compare, a study done by 4 researchers achieved a top accuracy of 97.9% [X]. This high accuracy could be influenced by the specific characteristics of the dataset used, such as its size or the particular distribution of features, which may differ from those used in benchmark studies. These factors can impact the generalizability of the model to other datasets or real-world scenarios. The approach in my project emphasizes the effectiveness of model tuning and feature selection, similar to established methods, yet variations in dataset specifics could explain the slight performance differences observed.

1. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9268898/> Decision Tree

In my project on decision trees for stroke prediction, I observed higher accuracy with deeper tree depths, although I did not implement pruning. While my model achieved high accuracy, the comparison to studies using J48 and RepTree, which incorporate pruning, highlights an opportunity for improvement [X]. Pruning could help my future models maintain a balance between capturing complex patterns and avoiding overfitting, ensuring that the high accuracy I observed is robust and generalizable to new datasets.