STROKE PREDICTION FINAL REPORT

Team member: Bala Adithya Malaraju, Harsha Vardhan Bollineni, Sai Chandra Reddy

Summary of the project:

The Stroke Prediction Project plans to create a reliable prediction model for identifying people who are at risk of having a stroke. The project tackles a crucial healthcare concern by offering early detection capabilities, enabling timely treatments, and encouraging stroke risk reduction techniques using the Stroke Prediction Dataset sourced from Kaggle. With approximately 5,110 observations, the dataset enables robust model building and evaluation. Gender, age, medical history, and lifestyle factors are all important predictors.

The Knime workflow employed Decision tree Analysis, Logistic regression, and Neural networks. Using extensive data preprocessing, feature engineering, and model training, the workflow achieved a Highest False Positive Rate of 0.878. Synthetic Minority Over-Sampling Technique (SMOTE) was used in all the three models to address the class imbalance.

Business Problem/Opportunity:

The Stroke Prediction Dataset offers a chance to use predictive modeling to solve this urgent problem. Our objective is to create a reliable prediction model that can correctly pinpoint those who are at risk of having a stroke. This prediction model's effective creation not only provides potentially life-saving skills, but also helps healthcare professionals make wise judgments and give high-risk people timely interventions. It may encourage people to lead better lives, which will help lower the rates of morbidity and death associated with strokes.

Specific Business Objective:

Specific business objectives for a stroke prediction dataset involve setting clear and measurable goals related to using the data to address the issue of stroke prediction. These goals must be in line with the organization's mission and more general healthcare aims. The following are some specific Business objectives for a stroke prediction dataset:

- Develop educational materials and awareness campaigns based on the findings, with a goal of reaching a certain percentage of the population to promote stroke risk reduction and early intervention.
- Develop a model that can predict strokes at an earlier stage, thus enabling proactive healthcare interventions for individuals at high risk, reducing the severity and impact of strokes.

- Identify specific demographics or subgroups within the dataset (e.g., age groups, genders) that are at a higher risk of stroke and create targeted prevention and awareness campaigns for these populations.

Process followed for selecting and gathering the data:

1. **Data Source Identification:**
   - Define the problem statement as follows: Determine the requirement for a stroke prediction model.
   - Determine the factors that must be predicted (for example, age, gender, and health indicators)
   - Choose data sources that are relevant to the project's goals.

2. **Kaggle Dataset selection:**
   - Because of its relevance to the topic, selected the Stroke Prediction Dataset from Kaggle.
   - Check that the dataset has a large enough number of observations (more than 5,110) to ensure statistical robustness.

3. **Data Exploration:**
   - Begin by exploring the dataset to learn about its structure and variables.
   - Examine the model for missing values, outliers, and any flaws that might degrade model performance.

4. **Relevance of Variables:**
   - Determine the importance of each variable in the dataset to the problem of stroke prediction.
   - Ensure that the factors chosen to have a significant influence on predicting stroke outcomes.

5. **Definition of the Target Variable:**
   Define the Target variable ("Stroke Outcome") as a binary outcome variable indicating whether or not a stroke occurred.

6. **Ethical Considerations:**
   Consider ethical implications related to the data, ensuring that the dataset collection and usage align with privacy and ethical standards.

7. **Data Gathering:**
   - Access reliable and representative information by retrieving the Stroke Prediction Dataset from Kaggle.

- Check the dataset's integrity to avoid biases and mistakes.

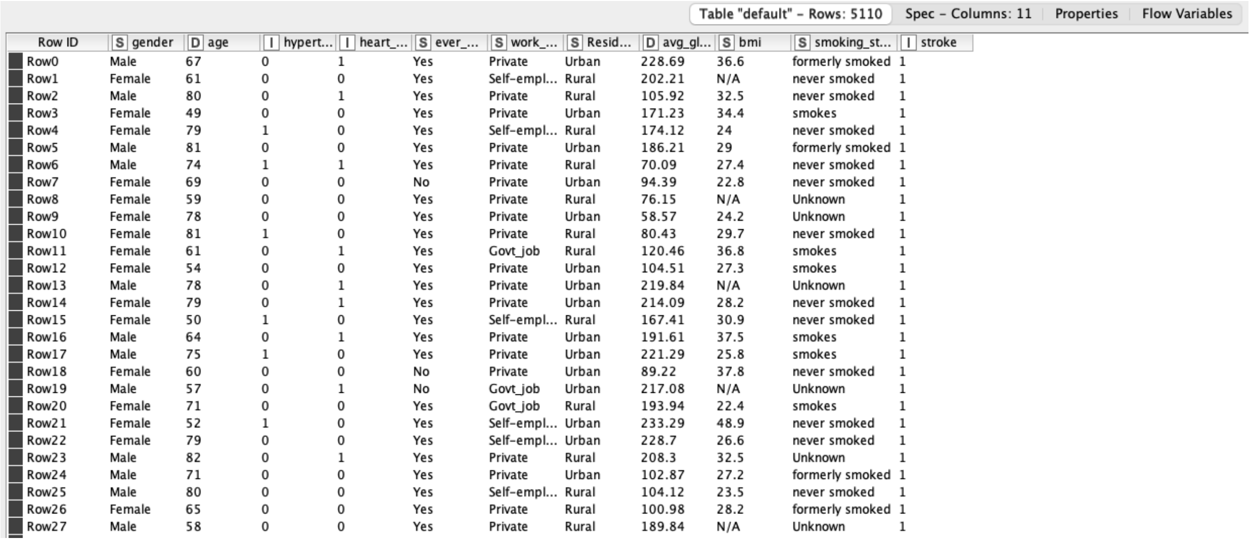
8. **Data Preprocessing:**
   - Based on the nature of the data, handle missing values by imputing or deleting them.
   - To guarantee uniformity among features, normalize or standardize numerical variables.
   - For model compatibility, encode categorical variables suitably.

9. **Data Splitting:** Divide the dataset into training (80%) and validation (20%) sets to facilitate model training and evaluation.

10. **Data Documentation:**
    - Create comprehensive documentation detailing the dataset's origin, variables, and any preprocessing steps applied.
    - Include details about any data modifications performed to enable transparency and repeatability.

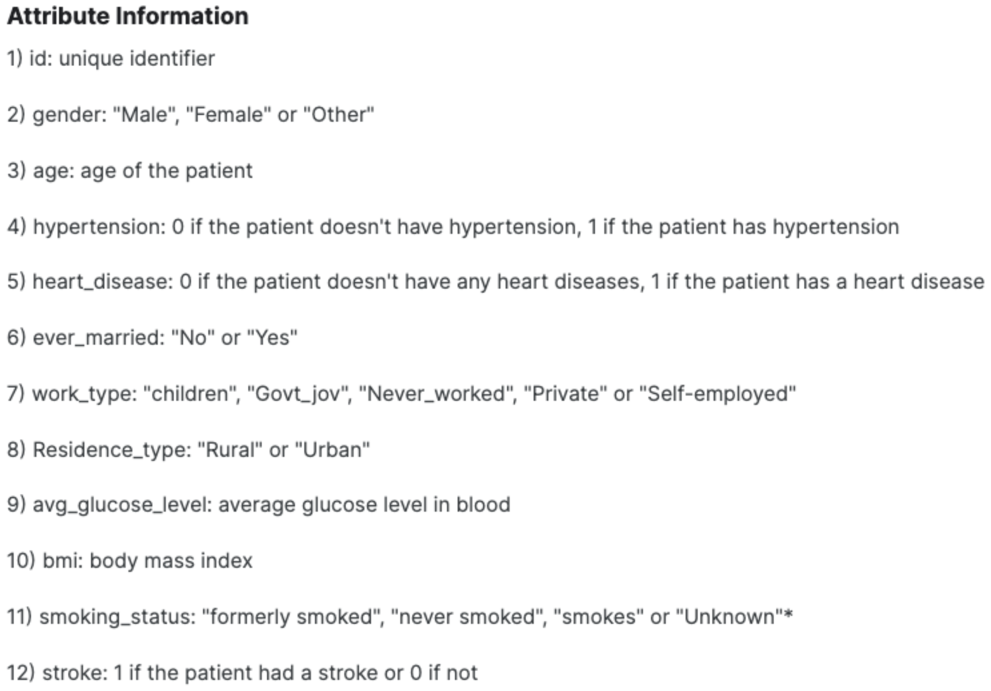
Discussion of preliminary data exploration and findings:

The Stroke Prediction Dataset contains approximately 5,110 observations, resulting in an adequate sample size for analysis. (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)



Row ID	S gender	D age	I hypert...	I heart_...	S ever_...	S work_...	S Resid...	D avg_gl...	S bmi	S smoking_st...	I stroke
Row0	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Row1	Female	61	0	0	Yes	Self-empl...	Rural	202.21	N/A	never smoked	1
Row2	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Row3	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
Row4	Female	79	1	0	Yes	Self-empl...	Rural	174.12	24	never smoked	1
Row5	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
Row6	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
Row7	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
Row8	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
Row9	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
Row10	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
Row11	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
Row12	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
Row13	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
Row14	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
Row15	Female	50	1	0	Yes	Self-empl...	Rural	167.41	30.9	never smoked	1
Row16	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
Row17	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
Row18	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
Row19	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
Row20	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
Row21	Female	52	1	0	Yes	Self-empl...	Urban	233.29	48.9	never smoked	1
Row22	Female	79	0	0	Yes	Self-empl...	Urban	228.7	26.6	never smoked	1
Row23	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
Row24	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1
Row25	Male	80	0	0	Yes	Self-empl...	Rural	104.12	23.5	never smoked	1
Row26	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly smoked	1
Row27	Male	58	0	0	Yes	Private	Rural	189.84	N/A	Unknown	1

- Variables include demographic data (such as age and gender), health indicators (such as hypertension and heart disease), lifestyle variables (such as smoking status), and the binary outcome variable "Stroke."



Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if not

- The KNIME workflow processed stroke prediction data in the preliminary data exploration using CSV Reader.
- Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance for decision tree analysis, Logistic Regression analysis, Neural networks analysis.
- The reason why we used SMOTE because, in the given dataset, we came to know while using the Row Splitter that the percentage of people who didn’t get the stroke (4861 out of 5110) i.e 95.12% is way higher than the people who actually effected by stroke(249 out of 5110) i.e 4.88%. so to address the class imbalance we opted to use SMOTE.
- The distribution of the goal variable, "Stroke Outcome," was investigated to better understand the prevalence of strokes in the dataset.
- Categorical characteristics such as gender, marital status, work type, and smoking status were examined.
- Potential outliers in numerical variables (age, average glucose level, BMI) were identified and assessed for their impact on model training.
- Remarkably, a Logistic regression model using Backward Feature elimination achieved a Highest False Positive Rate of 0.878

Description of data preparation - repairs, replacements, reductions, partitions, derivations, transformations and variable clustering

- **Outlier Treatment (Repairs):** Identified and Handled outliers in numerical variables. The outlier treatment aimed to prevent skewed model training and ensure robustness.
- **Data partitioning (partitions):** The dataset has been split into training and validation sets to facilitate model training and evaluation. The split was made as 80% for training and 20% for validation.
- **Variable Transformations (transformations):** Applied transformations like Logarithm to numerical variables and numeric to string variables to achieve better model as well as to address issues like skewed distributions.

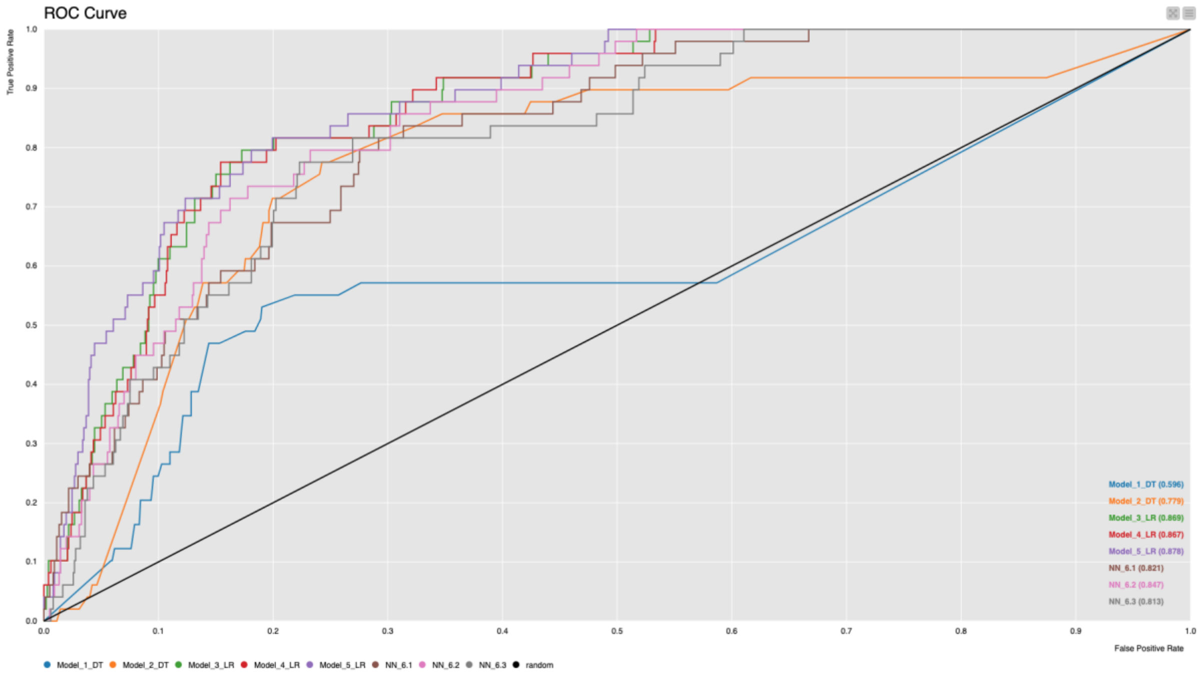
Description of data modeling/analyses and assessments

We had imbalanced data in the dataset, thus we had to balance it by oversampling the negative class with a smote method.

To decrease skewness in logistic regression, we used LOG TRANSFORMATION. The 80-20 split worked perfectly for us.

We did the Decision tree analysis initially, and later we performed the logistic regression model with and without logistic regression, Logistic regression model using Backward feature elimination using SMOTE. We then performed the neural network analysis using SMOTE.

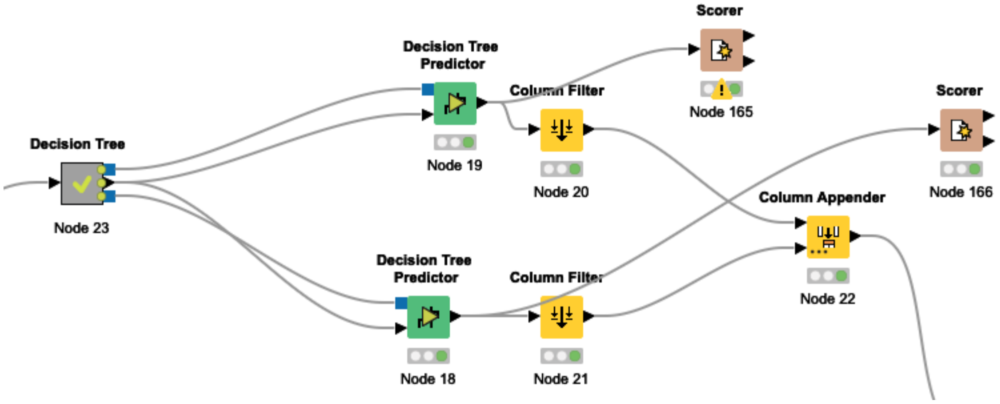
Here is an Assessment of the data modeling techniques used :

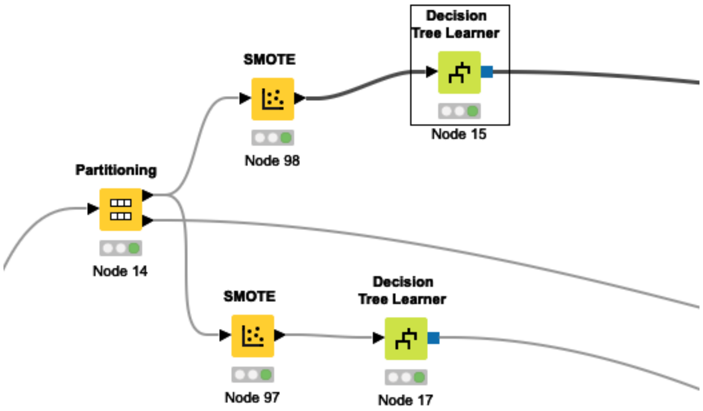


Model Selection & Conclusion:

Model	Accuracy	Recall(sensitivity)		Precision		Specificity		FP Rate – ROC Curve
		0	1	0	1	0	1	
Decision Tree(w/o Pruning)	85.28	0.881	0.286	0.961	0.109	0.286	0.881	0.596
Decision Tree(with Pruning)	81.018	0.82	0.612	0.977	0.146	0.612	0.82	0.779
Logistic Regression (without LOG)	74.755	0.744	0.816	0.988	0.138	0.816	0.744	0.869
Logistic Regression (With LOG)	74.853	0.745	0.816	0.988	0.139	0.816	0.745	0.867
Logistic Regression (Backward Feature)	75.245	0.748	0.837	0.989	0.143	0.837	0.748	0.878
Neural Network -1	72.505	0.721	0.796	0.986	0.126	0.796	0.721	0.821
Neural Network -2	75.049	0.749	0.776	0.985	0.135	0.776	0.749	0.847

1. Decision Tree Models:

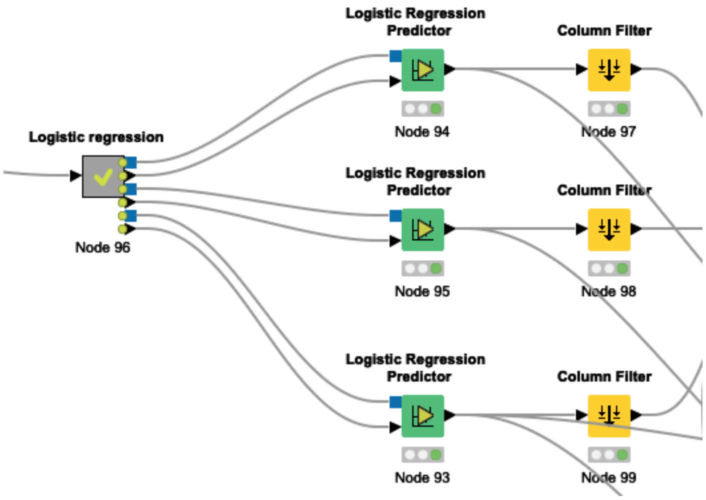


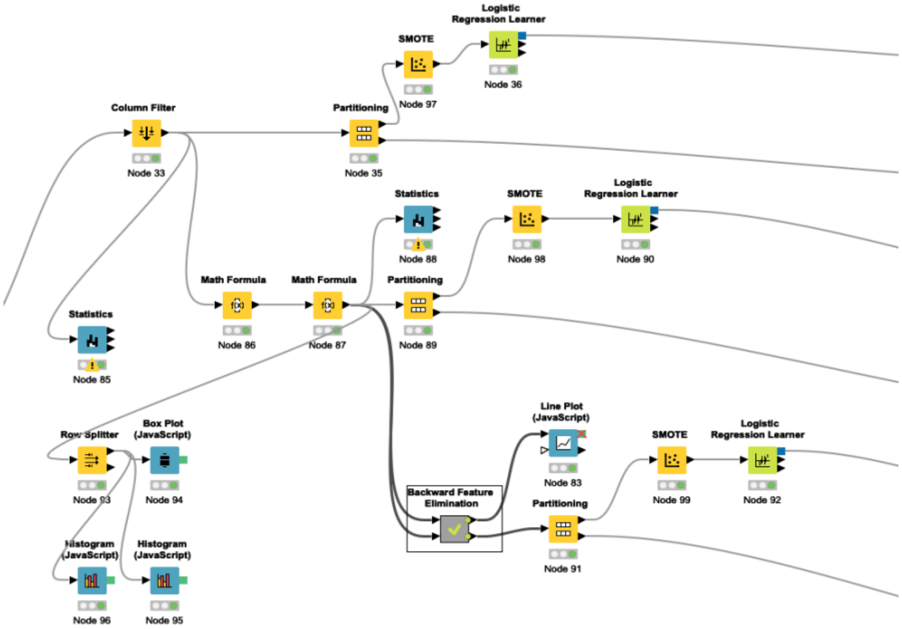


- Without Pruning: This model is built without any restrictions on the depth or size of the tree. It tends to have high accuracy but can overfit the training data, leading to poor generalization on unseen data.

- With Pruning: This model involves limiting the growth of the tree (e.g., setting a maximum depth or minimum number of samples per leaf). Pruning helps to reduce overfitting and can improve the model's performance on new data.

2. Logistic Regression Models:



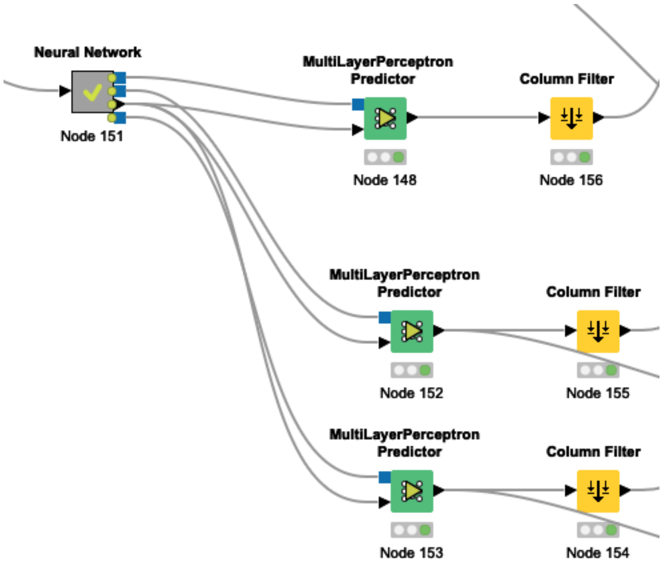


- No Log Transformations: A basic logistic regression model without any transformations. It serves as a baseline to compare other logistic regression models.

- With Log Transformations: This model applies logarithmic transformations to certain features, potentially improving the model's ability to capture non-linear relationships.

- Log Transformation with Backward Feature Elimination: This advanced approach not only applies log transformations but also uses backward feature elimination to remove less significant features iteratively. This model aims to enhance performance by focusing on the most relevant predictors.

3. Neural Network Models:



- Model 1: A more complex model with 35 iterations, 3 hidden layers, and 10 neurons in each layer. This model can capture more complex patterns but requires careful tuning to avoid overfitting.

- Model 2: Similar to Model 2 but with 20 neurons in each hidden layer, increasing the model's capacity to learn from data.

4. Model Evaluation and Comparison:

- Each model is evaluated using metrics like accuracy, precision, recall (sensitivity), and specificity. The ROC curve and AUC (Area Under the Curve) are also used to assess model performance.

- The decision tree model without pruning, despite its high accuracy, may suffer from overfitting, as indicated by lower precision, sensitivity, and specificity compared to other models.

- The logistic regression model with log transformation and backward feature elimination stands out due to its balance in all metrics, including a high false-positive rate in the ROC curve, indicating better generalization.

- Neural network models are evaluated based on their complexity and ability to capture non-linear relationships. The performance of these models can vary significantly based on their architecture and hyperparameters.

5. Model Selection:

Based on the provided performance metrics for the various models, the conclusion that the logistic regression model with log transformation and backward feature elimination is the best choice for stroke prediction in this scenario can be drawn from several key observations:

- **Balanced Performance Across Metrics:** This model demonstrates a well-balanced performance across crucial metrics, including accuracy, recall (sensitivity), precision, and specificity. While its accuracy (75.245%) is not the highest among all models, it offers a good balance between correctly identifying positive cases (sensitivity) and avoiding false positives (specificity).

- **High Precision and Sensitivity:** The model achieves a precision of 0.837 and a sensitivity of 0.748. This indicates that it not only accurately identifies a high proportion of actual stroke cases (sensitivity) but also maintains a high level of precision, meaning that when it predicts a stroke, it is correct a significant portion of the time.

- **Appropriate False Positive Rate:** The false positive rate in the ROC curve for this model is 0.143, which is relatively low compared to other models. This indicates a lower rate of incorrectly predicting strokes in healthy individuals, which is crucial in medical predictions to avoid unnecessary anxiety or medical interventions.

- **ROC Curve Performance:** Although not the highest, the ROC performance of this model is commendable (0.878), suggesting a good trade-off between sensitivity and specificity. This balance is important in medical diagnostics, where both identifying true positives and avoiding false positives are crucial.

- **Contextual Suitability:** In the context of stroke prediction, where the cost of false negatives (failing to identify a real stroke) can be very high, the relatively high sensitivity of this model makes it a suitable choice. Additionally, its precision ensures that the healthcare resources are utilized effectively, avoiding unnecessary treatments based on false positives.

- **Comparative Analysis:** When compared to other models, including decision trees and neural networks, the logistic regression model with log transformation and backward feature elimination stands out for its balanced performance. While some models may excel in one metric, they fall short in others. This model provides a more holistic approach to stroke prediction.

6. Business Implication:

The chosen logistic regression model aids in accurately identifying individuals at risk of stroke, which is crucial for preventive healthcare measures and timely interventions.

This detailed analysis demonstrates the importance of not only considering accuracy but also other performance metrics and the model's ability to generalize when selecting the best model for a specific application like stroke prediction.

Conclusion

In this project, we explored various machine learning models to predict stroke risk, focusing on achieving a balance between accuracy, sensitivity, precision, and specificity. After rigorous testing and analysis, the logistic regression model with log transformation and backward feature elimination emerged as the most effective. This model strikes an optimal balance, offering robust predictive power without overfitting, thanks to its methodical feature selection. It demonstrates high precision, ensuring that stroke predictions are reliable, and maintains a commendable sensitivity rate, crucial for medical diagnostics where missing true stroke cases can have serious consequences. While other models, like decision trees and neural networks, showed promise in certain aspects (e.g., higher accuracy in some cases), they didn't provide the same level of balanced performance across all key metrics. The logistic regression model's ability to minimize false positives while accurately identifying true stroke cases makes it particularly suitable for practical healthcare applications, where both identifying at-risk individuals and avoiding unnecessary medical interventions are equally important. This project highlights the importance of a nuanced approach to model selection, where a balance of various performance metrics, rather than a singular focus on accuracy, guides the decision-making process in a critical domain like healthcare.