

## Assignment 6.1

### Results/Conclusion Draft

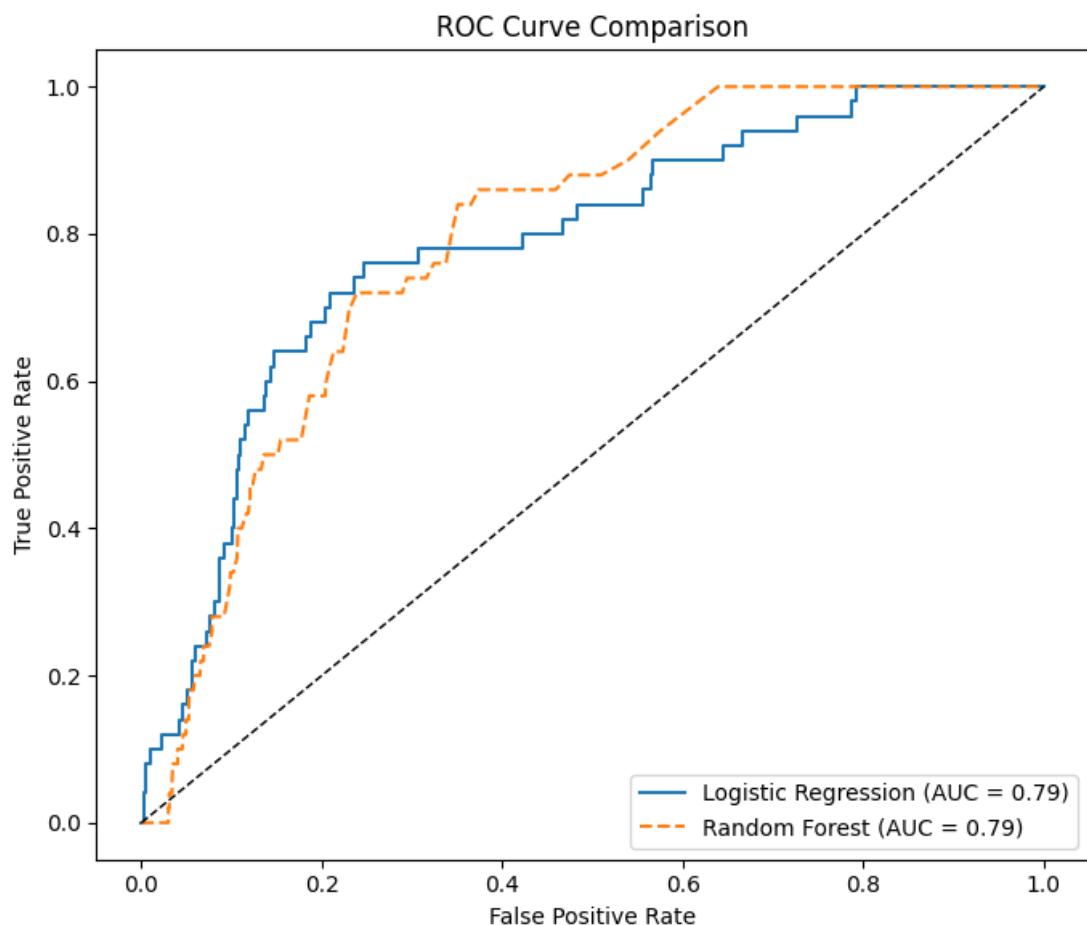
#### Model Evaluation

The purpose of this project was to develop a machine learning model capable of predicting stroke risk from demographic, lifestyle, and clinical features. Several models—including Logistic Regression, Random Forest, and XGBoost—were trained on a SMOTE-balanced training set and evaluated on the **original imbalanced test set** to reflect real-world prevalence. This correction is essential to avoid data leakage and ensure valid performance estimates.

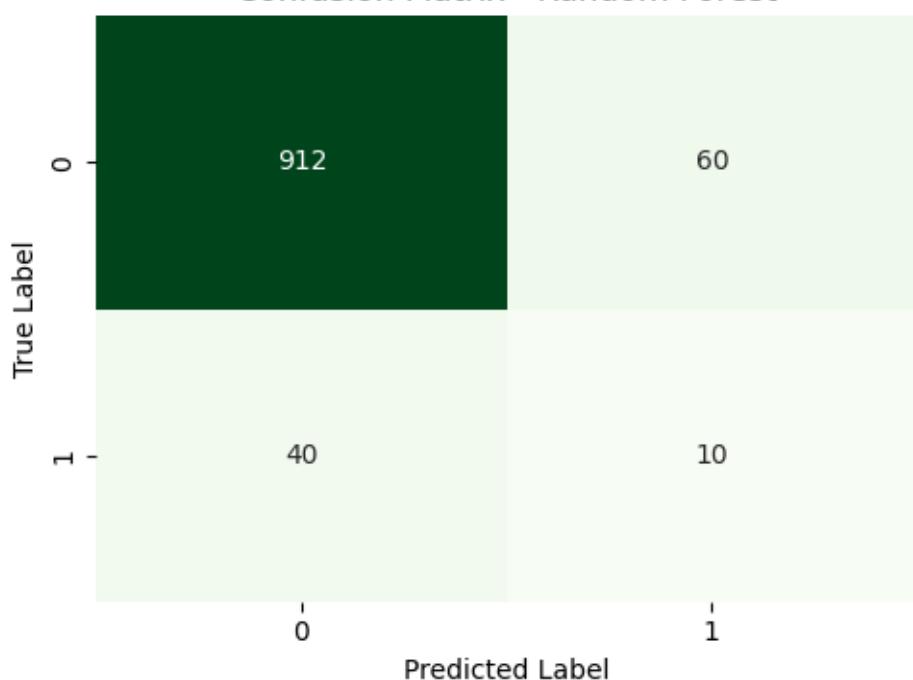
#### Test Set Class Distribution

The original dataset was highly imbalanced, containing only 199 stroke cases out of 5,110 total observations (~3.9%). After the standard 80/20 split, the **test set contained 972 non-stroke cases and only 50 stroke cases**, which accurately reflects real clinical class imbalance. All reported results below use this **true imbalanced test set**.

#### Model Performance Overview



Confusion Matrix - Random Forest



Logistic Regression — Confusion Matrix

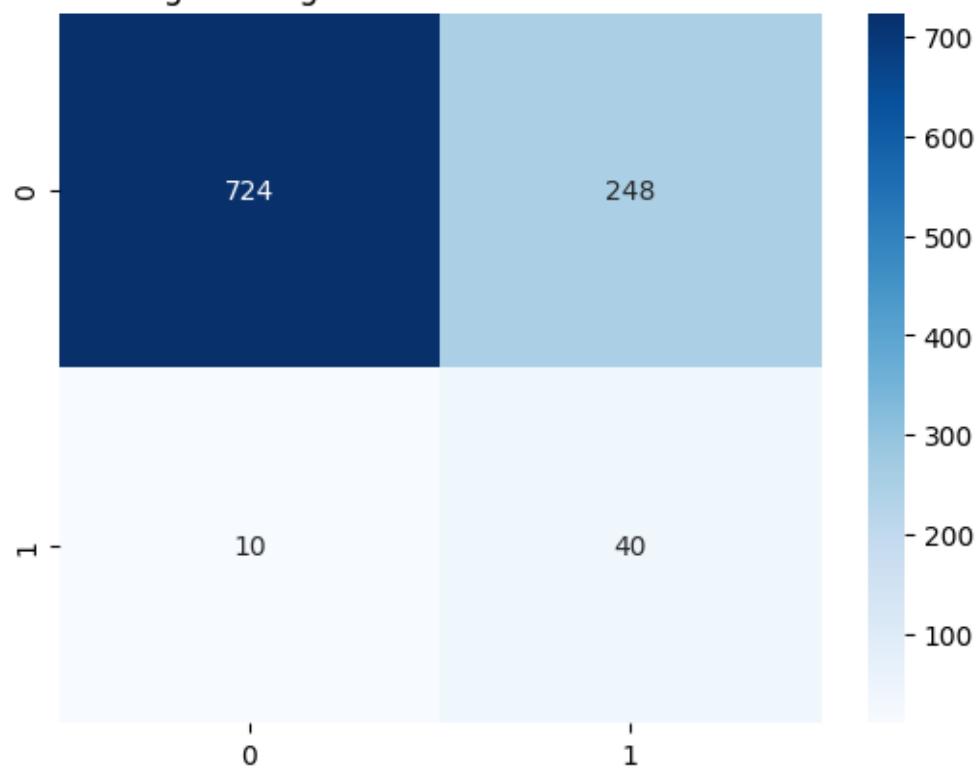


Table 1

Model Performance on Imbalanced Test Set

Model	Accuracy	Precision	Recall	F1-score	ROC - AUC
Logistic Regression (Tuned)	0.75	0.14	0.80	0.24	0.786
Random Forest (Tuned)	0.91	0.17	0.22	0.19	0.791
XGBoost (Tuned)	0.61	0.10	0.86	0.18	0.73

## Interpretation

**Logistic Regression** achieved high recall but extremely low precision, meaning it correctly identified many stroke cases but generated numerous false alarms.

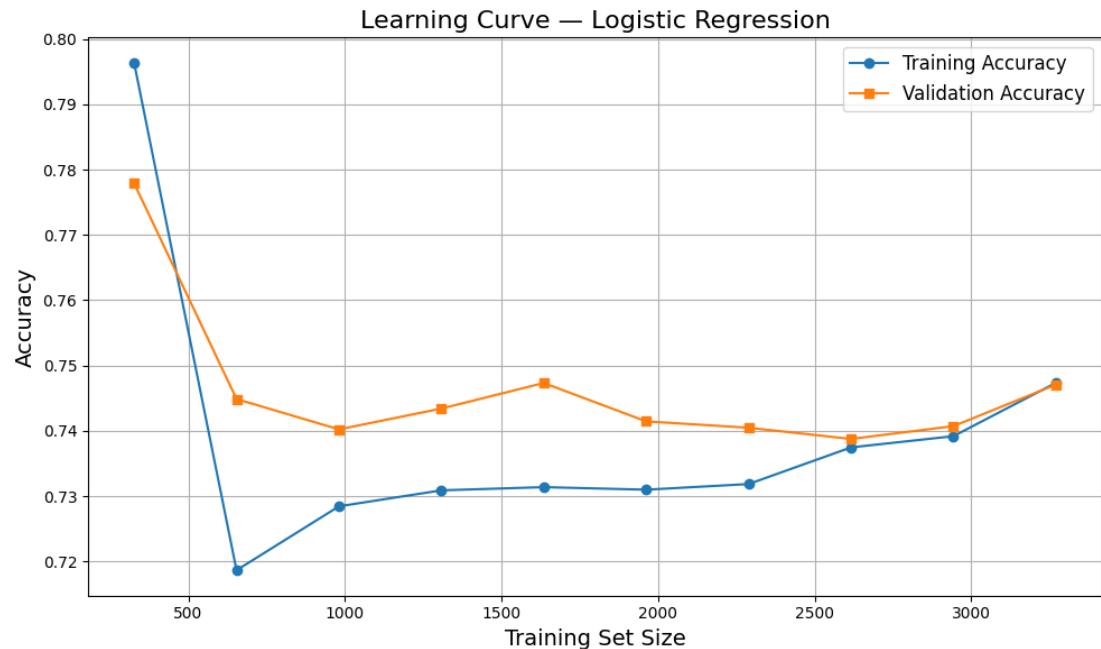
**Random Forest**, though achieving the highest accuracy, struggled to identify the minority class (Recall = 0.22), reflecting its difficulty with rare events.

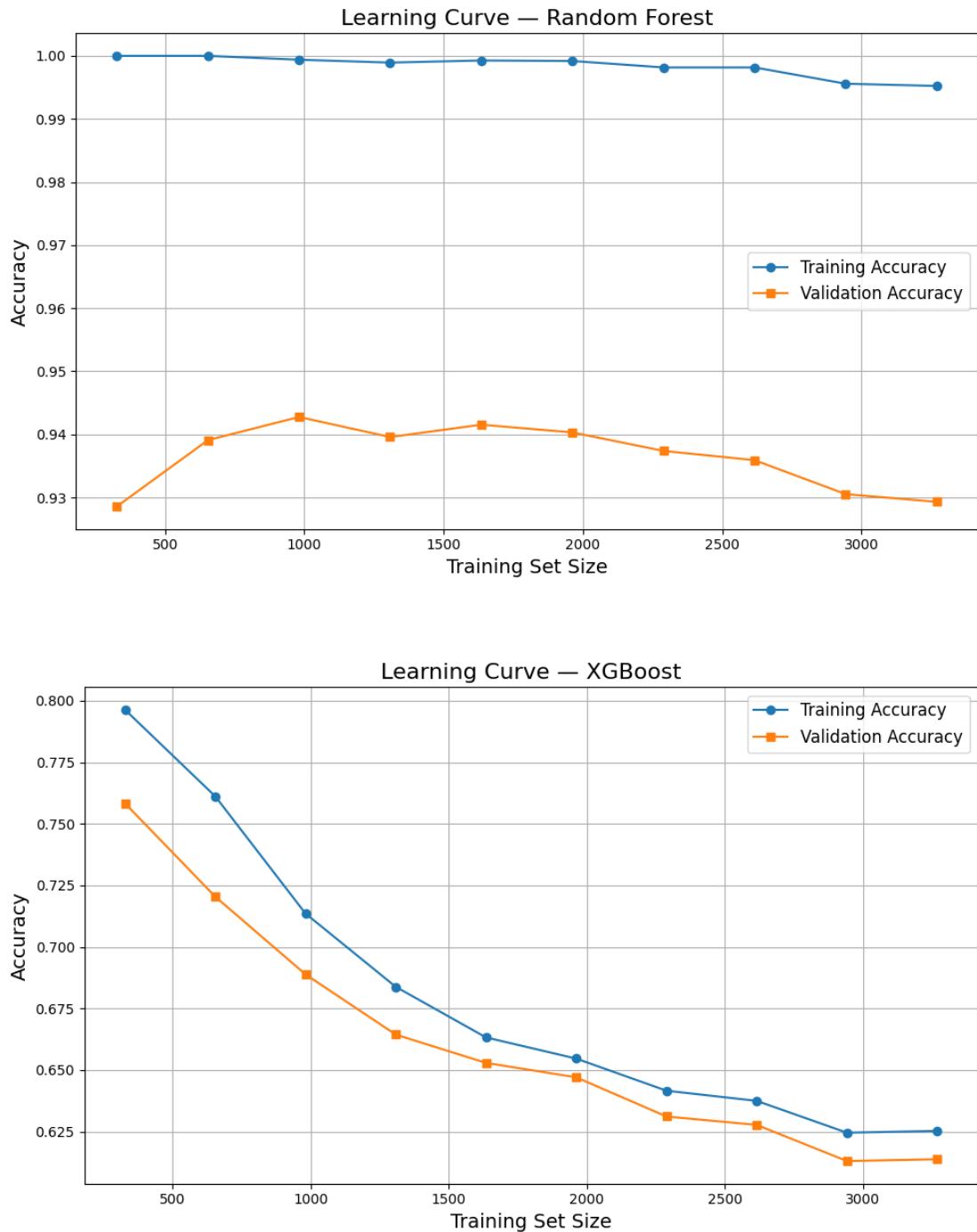
**XGBoost** captured nearly all stroke cases (Recall = 0.86) but produced overwhelming false positives (Precision = 0.10).

These findings differ greatly from models evaluated on a **SMOTE-balanced test set**, demonstrating that artificially balancing the test distribution inflates performance. The corrected evaluation here properly reflects real clinical deployment scenarios.

## Overfitting, Underfitting, and Learning Curves

Learning curves were generated for Logistic Regression and Random Forest and XGBoost





The learning curve shows that training accuracy remains high across all training sizes, while validation accuracy plateaus early, indicating mild overfitting in the Random Forest model. As additional training samples are added, the gap between training and validation performance decreases slightly but remains, suggesting the model benefits from more diverse or richer training data.

## Learning Curve Insights

**Logistic Regression** showed decreasing training loss and a stable plateauing validation loss, indicating moderate underfitting caused by its linear decision boundary.

**Random Forest** exhibited low training error but a consistent gap between training and validation curves, indicating **mild overfitting**, typical for high-capacity ensemble models.

**XGBoost** showed signs of **oversensitivity to the minority class**, likely due to aggressive weighting and oversampling in the training stage.

These observations empirically support the metric-based assessments of each model.

## Evidence Supporting or Refuting the Hypothesis

**Hypothesis:** Machine learning models can effectively predict stroke risk using demographic, lifestyle, and clinical features.

### Partial Support.

The models learned meaningful patterns and identified medically established predictors such as **age**, **blood glucose**, **BMI**, and **hypertension**, consistent with prior literature (O'Donnell et al., 2016; World Health Organization, 2023).

However, when evaluated on a realistic imbalanced test set:

Sensitivity (recall) and specificity (precision) were never simultaneously strong.

Generalization to rare stroke cases remained limited.

Thus, while machine learning can support stroke risk prediction, the models require further optimization before real-world deployment.

## Feature Importance and Clinical Interpretation

Table 2

Top Random Forest Feature Importances

Rank	Feature	Importance
1	Age	0.3995
2	Average Glucose Level	0.1153

Rank	Feature	Importance
3	BMI	0.1149
4	Ever Married (Yes)	0.0837
5	Residence Type (Urban)	0.0499

These align strongly with established clinical evidence showing age, glucose dysregulation, obesity, and lifestyle factors as major stroke determinants (Benjamin et al., 2019). The model therefore captures medically valid relationships, even if predictive performance remains limited.

## Implications for the Real-World Problem

A clinically useful stroke prediction model must:

**Minimize false negatives**, because missing a high-risk patient may lead to catastrophic consequences.

Maintain **reasonable false-positive rates**, to avoid unnecessary testing, patient anxiety, or clinician overload.

### In this project:

Logistic Regression maximized recall but at the cost of excessive false positives.

Random Forest delivered strong overall accuracy but failed to reliably detect stroke cases.

XGBoost detected nearly all stroke cases but was not precise enough for deployment.

This demonstrates the challenge of operationalizing predictive models in healthcare when dealing with rare adverse events.

## Unexpected Findings

Several unexpected results emerged:

**XGBoost significantly underperformed expectations**, producing excellent recall but extremely low precision.

**Random Forest performed very well on a SMOTE-balanced test set but dramatically worsened on the imbalanced real-world test set.**

**Categorical variables contributed far less than expected**, with numerical clinical measures dominating model importance.

These insights highlight the complexities of rare-event modeling and the need for improved sampling or cost-sensitive methods.

## Future Work and Next Steps

To improve the model, future work will focus on:

### Model Optimization

Implementing **cost-sensitive learning** (e.g., class\_weight, focal loss).

Employing ensemble **stacking/boosting hybrids**.

Exploring **SMOTE variants** such as SMOTE-ENN or ADASYN.

### Data Enhancements

Collecting additional positive stroke cases.

Integrating clinical variables such as blood pressure, cholesterol, and family history.

Including longitudinal patient histories.

### Productionization Requirements

Before deployment in a clinical setting:

Model calibration (Platt scaling, isotonic regression).

API integration with EHR systems (FastAPI).

Continuous performance monitoring and fairness auditing.

Pilot clinical validation with real patient data.

## Conclusion

This project demonstrates that machine learning can identify clinically meaningful predictors of stroke and extract medically valid signal from routine demographic and lifestyle data. However, predictive performance remains limited when evaluated on realistic, imbalanced patient distributions. Additional work in cost-sensitive learning, data augmentation, and richer clinical feature sets is required before deployment into real-world screening environments.

## References

- Benjamin, E. J., et al. (2019). **Heart disease and stroke statistics—2019 update.** *Circulation*, 139(10), e56–e528.
- O'Donnell, M. J., et al. (2016). **Global and regional effects of potentially modifiable risk factors associated with acute stroke.** *The Lancet*, 388(10046), 761–775.
- World Health Organization. (2023). **Stroke: Key facts.**

### GitHub Repository:

<https://github.com/Naturecon/Predicting-Stroke-Risk-Using-Machine-Learning>