COMP0172 ARTIFICIAL INTELLIGENCE FOR BIOMEDICINE AND HEALTHCARE COURSEWORK 1 REPORT

1. Introduction

This coursework report focuses on predicting stroke in patients using machine learning applied to clinical data. Stroke has long been a leading cause of death globally [10, 22, 23].

The dataset, likely sourced from Kaggle [26], consists of tabular records indicating whether patients have experienced a stroke, along with their physiological, social, and lifestyle data. The objective is to develop a model to predict stroke risk.

Machine learning models can assist physicians by predicting the disease, allowing early intervention and better resource allocation. They can identify complex, non-obvious data patterns beyond human expertise. Additionally, models like decision trees highlight important predictive factors, aiding diagnosis and treatment. Moreover, these models could be integrated into personal devices, such as smartphones, for accessible and low-cost healthcare outside clinics, as seen in other medical applications [8].

Formally, we use a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ represents a vector of D features and $y_i \in \{0, 1\}$ is the binary target. The aim is to learn a function $f : \mathbb{R}^D \to \{0, 1\}$ from a family of functions Γ, mapping inputs to the target. Training minimises the empirical risk of the loss function $L(y, f(\mathbf{x}))$ to find optimal parameters θ for f. The model is trained on $\mathcal{D}_{\text{train}}$, validated on $\mathcal{D}_{\text{val}}$, and tested on $\mathcal{D}_{\text{test}}$.

2. Exploratory Data Analysis

The dataset contains 5110 records with 12 columns: an ID, a binary target (stroke occurrence), and 10 features. The features are a mix of categorical (gender, hypertension, heart disease, marital status, work type, residence type, smoking status) and numerical data (age, BMI, glucose level).

The data is split into 80% training, 10% validation, and 10% testing sets to prevent data leakage. Validation data evaluates models during training, while the test set assesses the final model.

Figure 1 indicates a highly unbalanced target distribution, with a 1:20 stroke to no-stroke ratio. Although challenging, this imbalance reflects real-world medical data [13].

Key data observations:

- The ID column lacks informativeness and can be removed.
- The BMI column has 119 missing values.
- "Unknown" entries exist in smoking status.

Insights from Figures 2 and 3:

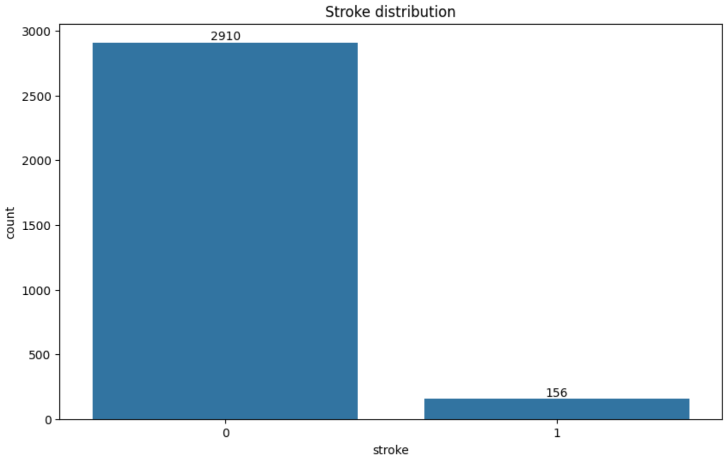


Figure 1. Train data target distribution: 0 (no stroke), 1 (stroke).

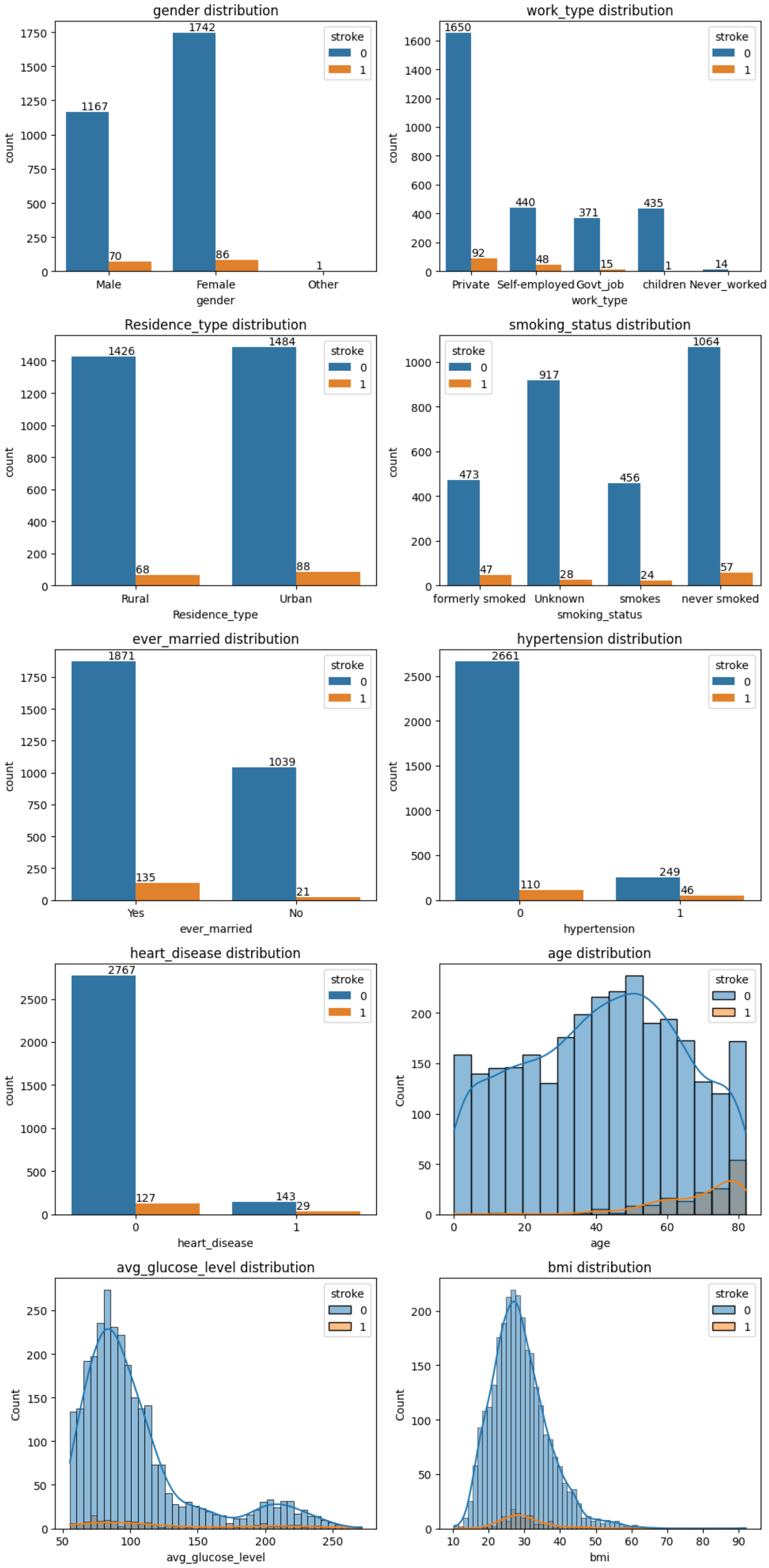


FIGURE 2. Distribution of 10 features in train data. Blue: no stroke, Orange: stroke.

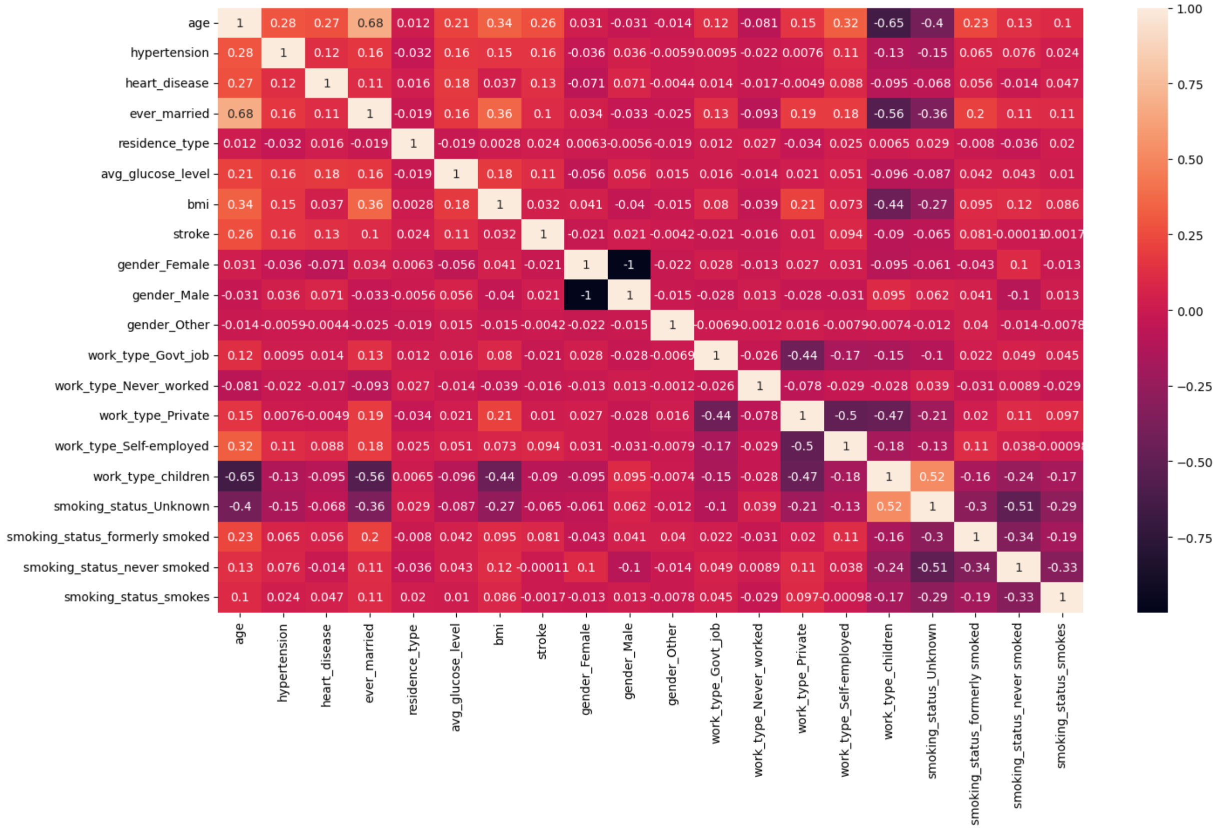


FIGURE 3. Correlation matrix of features and target, with one-hot encoding. Ignore diagonals of complementary one-hot columns despite high correlations.

- Unknown smoking status may be due to social stigma, with younger individuals or child-related jobs less likely to disclose.
- Age correlates with irreversible features over time, such as physiological decline or marital status.
- Stroke shows slight positive correlation with age, body metrics, hypertension, heart disease, glucose level, and BMI.

The data reveals intriguing patterns, notably around age. Stroke is mainly correlated with age and general health metrics, consistent with the understanding that older individuals and those with health issues are at higher risk.

3. METHODS

3.1. **Data Preprocessing.**

Initially, missing BMI values are imputed with the mean. The ”Unknown” smoking status serves as its own feature, as non-disclosure may indicate a revealing pattern worth modeling. Categorical data are either binary or one-hot encoded, including hypertension and heart disease, which are already encoded. Gender, work type, and smoking status undergo one-hot encoding into multiple columns. The ID column and the sparse ”gender_Other” column are removed.

Standardisation (z-score normalisation) is applied, beneficial for models like logistic regression and those using distances [9, 14]. Min-Max normalisation is also considered for gradient descent models like MLP (multi-layer perceptron) to prevent feature dominance due to magnitude.

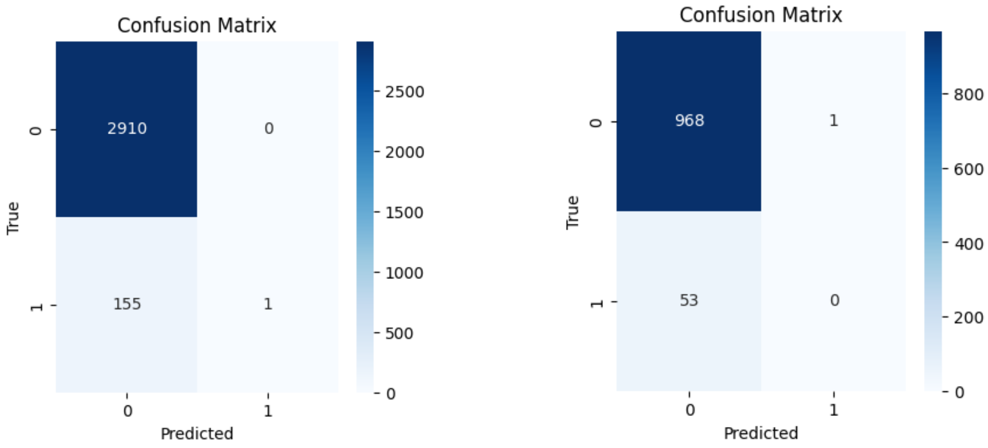
Additional optimisation strategies in data processing for the model are discussed in Section 3.3.

3.2. **Model Selection.**

Logistic Regression serves as the baseline due to its simplicity and interpretability for this binary classification task.

To identify the best-performing model, a variety of models were chosen following a literature review [3, 7, 17, 21, 24]. These include:

- Baseline (Logistic Regression)



(A) On Train Data (B) On Validation Data

Figure 4. Confusion matrices of Baseline model at vanilla setting.

- Stochastic Gradient Descent Classifier (SGD) with Logistic Regression
- Mult-Layer Perceptron (MLP)
- Support Vector Machine (SVM)
- K-Nearest Neighbours (KNN)
- (Gaussian) Naive Bayes
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Experiment Pipeline:

(1) Build and fit the preprocessor using only train data, applying transformations on validation and test sets.
(2) With a random seed, initialise models from sklearn (and xgboost) libraries with default settings. Train on train data and evaluate on validation data.
(3) Using baseline results, explore other models, sampling strategies for class imbalance, and encoding strategies. Develop a pivot model addressing key issues.
(4) Further optimise each model within the pivot setting using Hyper-Parameter Optimisation (HPO). Conduct two HPO rounds, then refine the hyper-parameter space for differing from pivot.
(5) With the optimised pivot model, test improvements in data processing and model selection. Run HPO for best model performance.
(6) Select and optimise the decision threshold for the best model, evaluating it on test data.

3.3. **Optimisation Strategies.**

For class imbalance, majority class is undersampled [24] or the minority class is oversampled using SMOTE [4], as random oversampling is less effective [16, 20].

Both one-hot and label encoding are considered for categorical data, with label encoding also proving effective [7, 24].

Imputation methods include mean and median (statistical) [7], and decision tree regressor (predictive) [21]. Mean is the default.

Feature selection is performed to eliminate redundancy and noise [1]. Feature importance from HPO pivot models guides selection [15, 18], using various aggregation methods and thresholds.

Decision thresholds are fine-tuned to potentially improve certain metrics, such as recall over precision, beyond the default 0.5 threshold.

4. Results and Discussion

4.1. **Metrics.**

The baseline model achieves an accuracy of 94.47% and ROC AUC of 0.7995. However, if we look at Figure 4, the model is clearly biased towards negative class, as expected from the class imbalance.

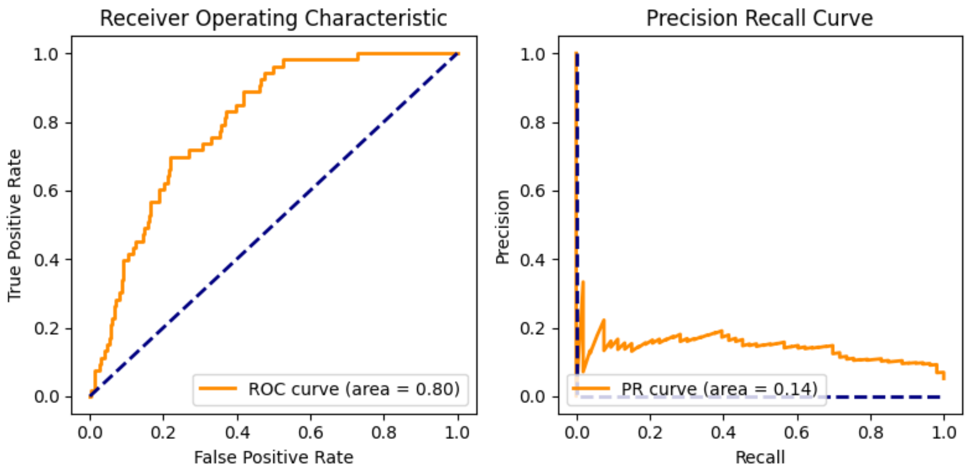
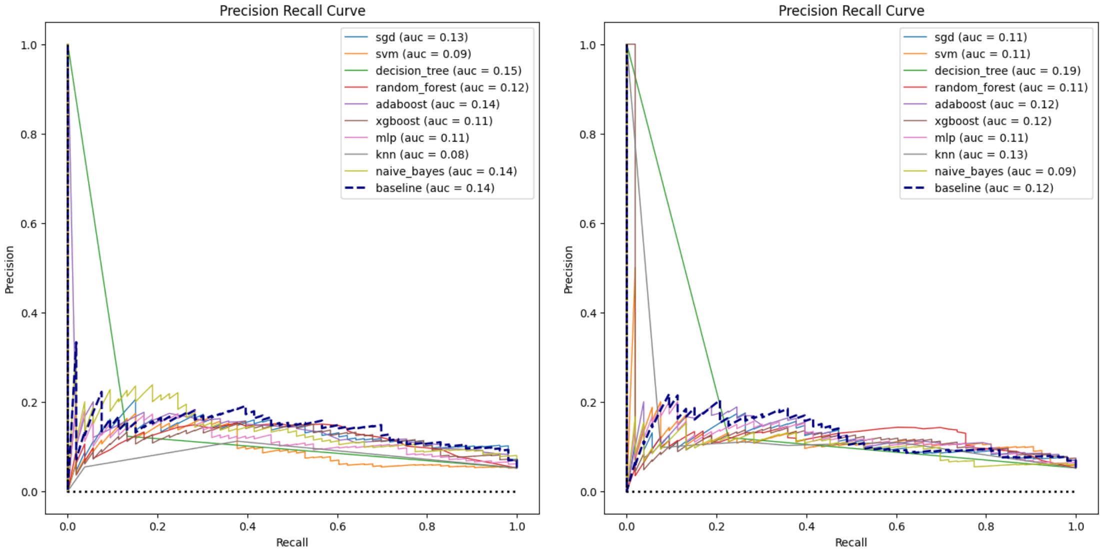


FIGURE 5. ROC and PR Curves of Baseline model at vanilla setting.



(A) Vanilla Setting

(B) Pivot Setting: SMOTENC + One-hot Encoding

FIGURE 6. PR Curves of models in vanilla and pivot settings.

Therefore, regular accuracy and ROC AUC (Receiver Operating Characteristic Area Under the Curve) are insufficient for evaluation. We use F1-Score, which looks after the precision and recall on the same class. And recall, which is crucial in this medical context, where missing a stroke patient is more dangerous than falsely diagnosing a healthy patient. The latter can be corrected with further tests, but the former can be fatal. To balance precision and recall, PR curve AUC is included alongside F1-Score and recall, thus ROC and PR in Figure 5.

Precision-Recall (PR) curve AUC is considered alongside F1-Score and recall. While F1-Score and recall are point estimates, PR curve AUC offers a broader evaluation. These metrics guide model evaluation through the HPO rounds.

4.2. **Pivot Setting.**

After testing various models, ”one-hot encoding” and ”SMOTENC oversampling” were chosen as the pivot setting, making minimal assumptions on categorical data with SMOTENC suitable for mixed data types [4]. This setting shows increased PR curve AUC (Figure 6). Results in Figure 6 show a slight PR curve AUC improvement. Decision tree PR AUC might be overly optimistic due to limited threshold points.

With the pivot setting, we explored different sampling and encoding strategies, observing that SMOTENC performs better than other variants, and undersampling and label encoding also show promise (Figure 7). Vanilla has unstable performances across models, but generally worse than pivot setting.

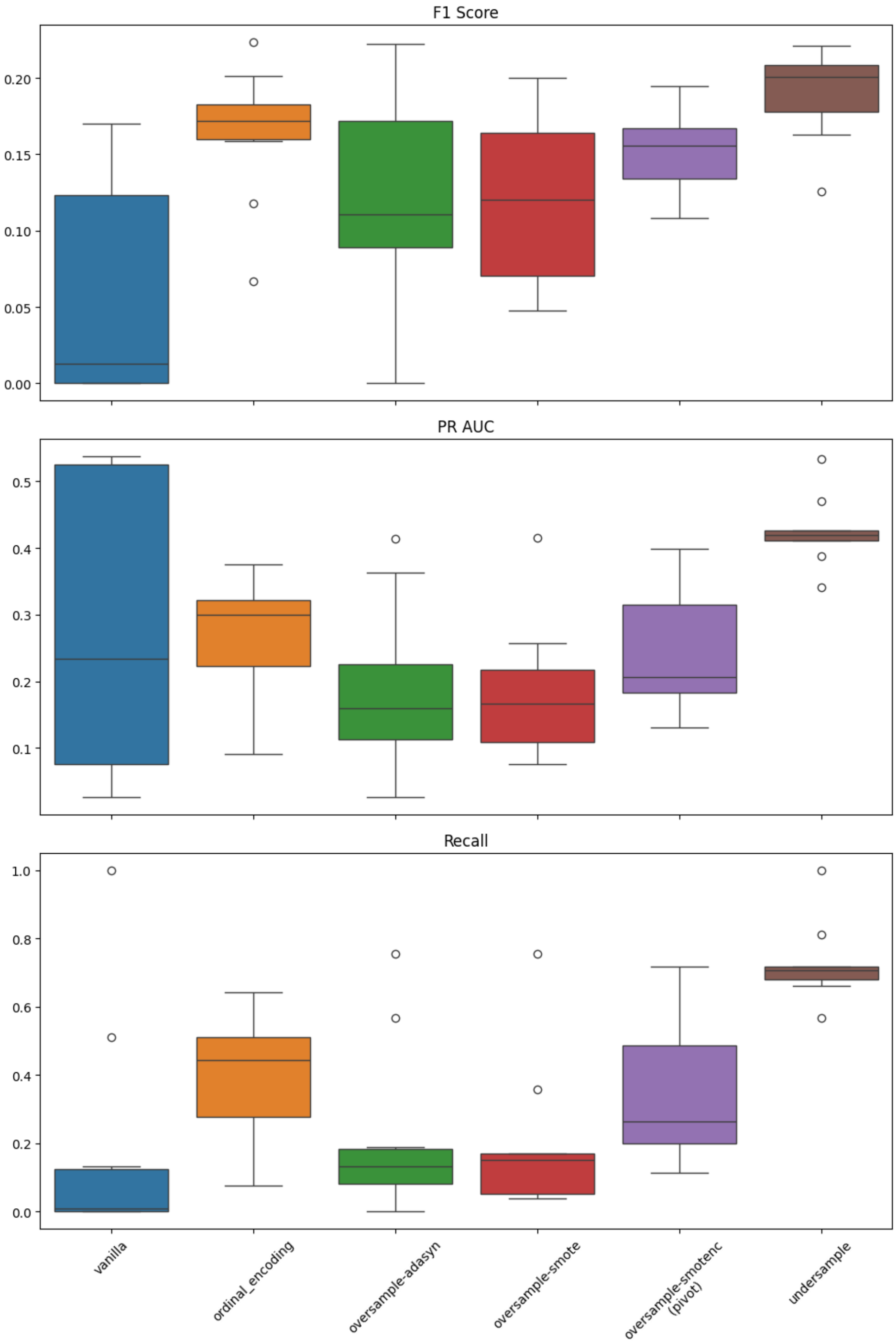


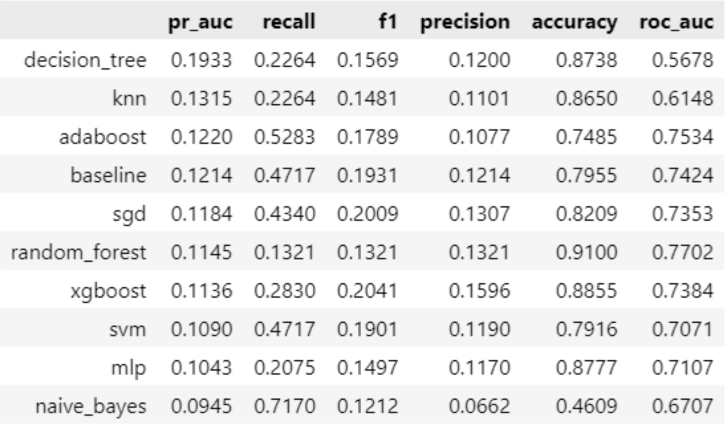
FIGURE 7. Performance metrics across settings before HPO. Each box represents the distribution of 10 models' performance.

4.3. **Optimisation.**

Conducting HPO on the pivot setting and further data processing optimisations yielded no significant gains (Figure 8).

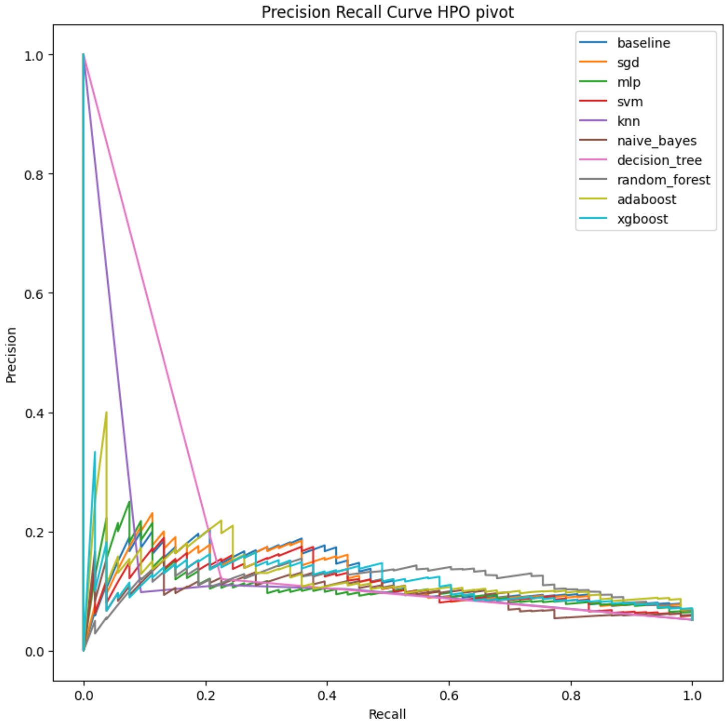
Figure 9 highlights performance improvements, notably in models employing undersampling and feature selection, suggesting a reduction in redundant information aids focus on significant data. Experimentation explored the following combinations:

- Sampling: SMOTENC or undersample
- Encoding: one-hot-encoding or label encoding



	pr_auc	recall	f1	precision	accuracy	roc_auc
decision_tree	0.1933	0.2264	0.1569	0.1200	0.8738	0.5678
knn	0.1315	0.2264	0.1481	0.1101	0.8650	0.6148
adaboost	0.1220	0.5283	0.1789	0.1077	0.7485	0.7534
baseline	0.1214	0.4717	0.1931	0.1214	0.7955	0.7424
sgd	0.1184	0.4340	0.2009	0.1307	0.8209	0.7353
random_forest	0.1145	0.1321	0.1321	0.1321	0.9100	0.7702
xgboost	0.1136	0.2830	0.2041	0.1596	0.8855	0.7384
svm	0.1090	0.4717	0.1901	0.1190	0.7916	0.7071
mlp	0.1043	0.2075	0.1497	0.1170	0.8777	0.7107
naive_bayes	0.0945	0.7170	0.1212	0.0662	0.4609	0.6707

(A) Scores



(B) PR curve

Figure 8. Performance of 10 models at pivot setting after HPO.

- Imputation: mean, median, or decision tree regressor
- Feature selection: by absolute mean or root mean square

In decision tree regressor, a tree model is trained on age, glucose level, bmi, hypertension, and heart disease, and predict on missing bmi values, as these are most likely related to bmi.

In feature selection, the values are normalised to sum 1 in each model, then aggregated by taking the absolute mean or root mean square. This can be visualised in Figure 10. The threshold is set to 0.05 and 0.085, resulting 10 and 5 features, respectively, where there is a clear difference in importance. We notice the high relevancy of age, and smoking status have a considerable impact in the prediction.

Finally, the best setting for each of the model is selected based on PR AUC, and a final round of decision threshold optimisation is performed. 100 thresholds are tried in the decision boundary, scored with F1-Score, due to its class sensitivity. Recall is especially emphasised beyond, by setting a constraint that requires models to have higher recall than 0.7 to update decision threshold.

Results of final models are shown in Figure 11. We also allow a looser constraint of recall 0.5 to see the difference in performance. Indeed, decreasing the constraint compromises recall for overall performance due to increase in precision.

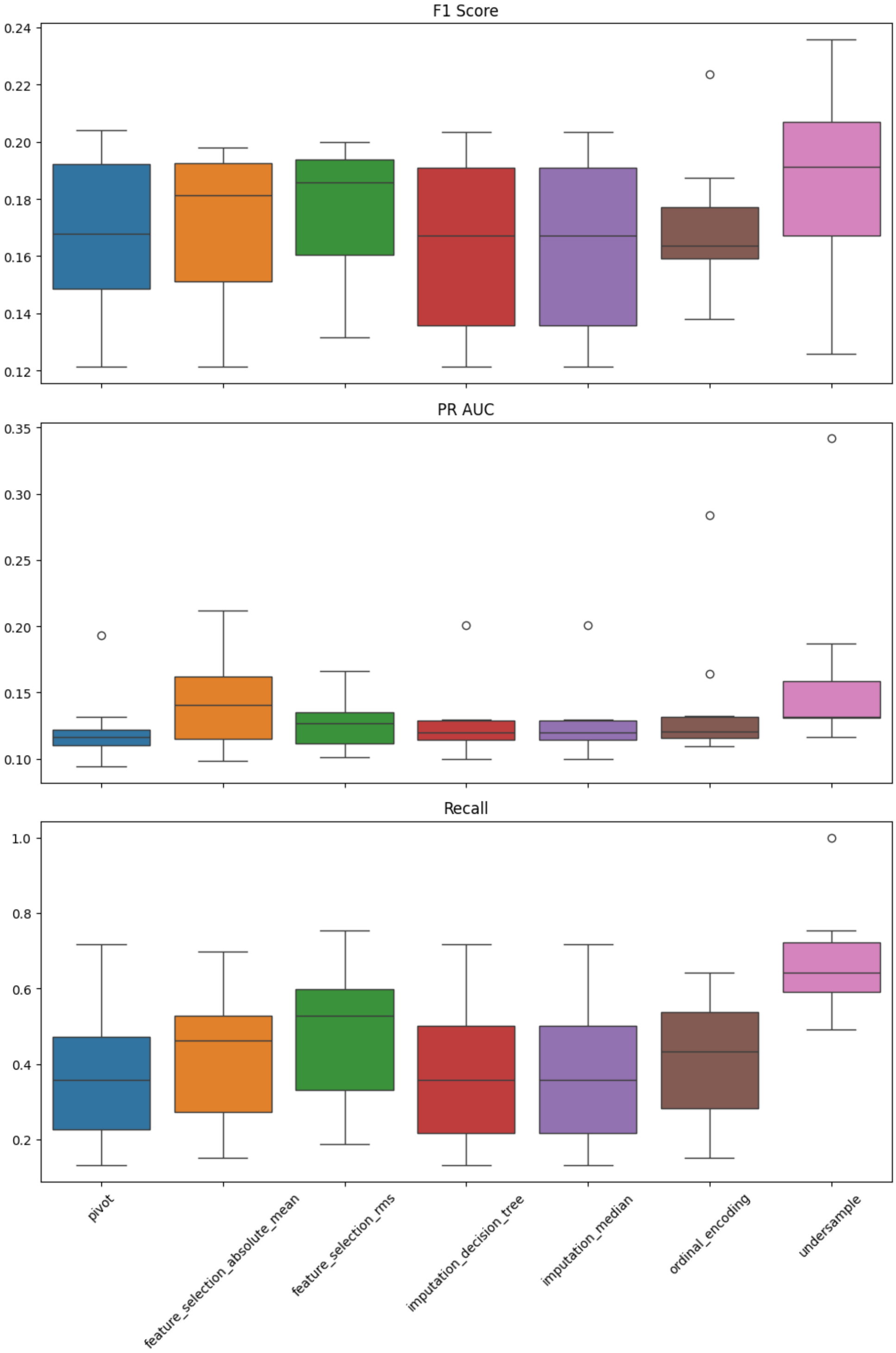


Figure 9. Box plots where each box represents the distribution of 10 models' performance. Each setting is HPO optimised.

In the end, our model of choice is XGBoost, which has the best performance in terms of F1-Score in constraint version and top-3 performance in the looser version. The final performance of the model on the test data is shown in Table 1. We clearly observe a huge increase in recall, at expense of some false positives that decreases the accuracy. This is a trade-off that we are willing to make, as explained earlier.

Finally, looking at the confusion matrices at 12 to more straightforwardly understand the performance of the model. Considering that only 40 out of 1022 patients had a stroke, the model is able to predict 33 of them correctly, leaving only 7 patients undiagnosed. Moreover, if the model's precision is not satisfactory,

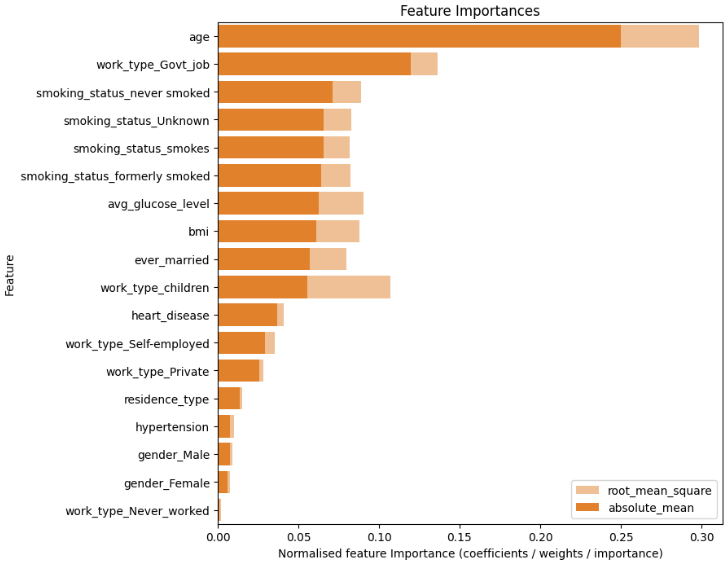
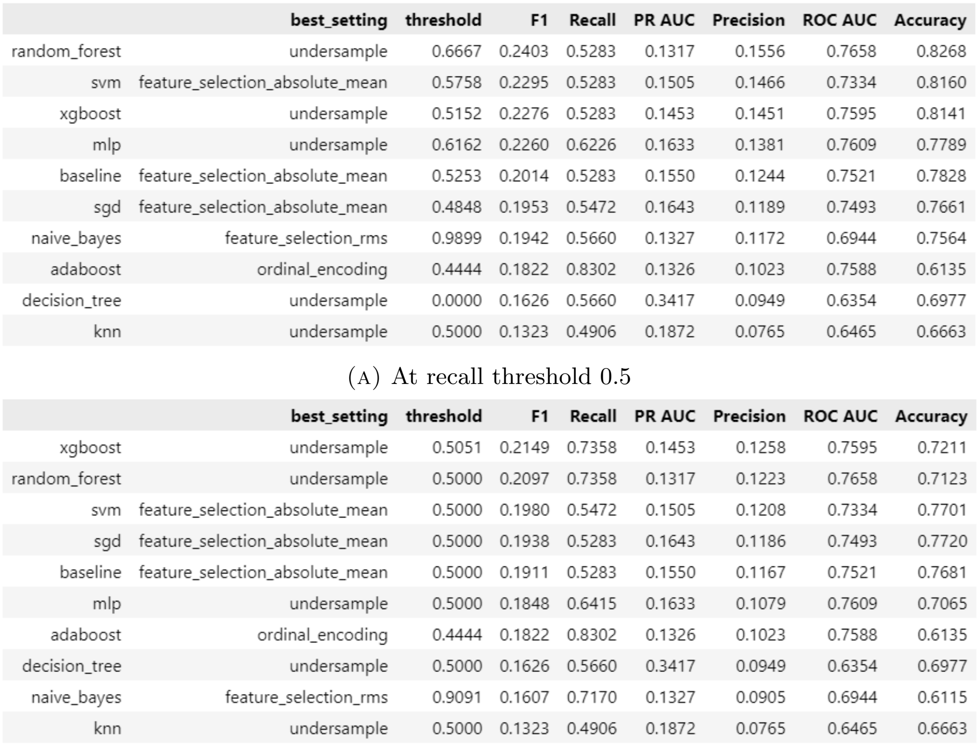


Figure 10. Feature importance of models after HPO.



	best_setting	threshold	F1	Recall	PR AUC	Precision	ROC AUC	Accuracy
random_forest	undersample	0.6667	0.2403	0.5283	0.1317	0.1556	0.7658	0.8268
svm	feature_selection_absolute_mean	0.5758	0.2295	0.5283	0.1505	0.1466	0.7334	0.8160
xgboost	undersample	0.5152	0.2276	0.5283	0.1453	0.1451	0.7595	0.8141
mlp	undersample	0.6162	0.2260	0.6226	0.1633	0.1381	0.7609	0.7789
baseline	feature_selection_absolute_mean	0.5253	0.2014	0.5283	0.1550	0.1244	0.7521	0.7828
sgd	feature_selection_absolute_mean	0.4848	0.1953	0.5472	0.1643	0.1189	0.7493	0.7661
naive_bayes	feature_selection_rms	0.9899	0.1942	0.5660	0.1327	0.1172	0.6944	0.7564
adaboost	ordinal_encoding	0.4444	0.1822	0.8302	0.1326	0.1023	0.7588	0.6135
decision_tree	undersample	0.0000	0.1626	0.5660	0.3417	0.0949	0.6354	0.6977
knn	undersample	0.5000	0.1323	0.4906	0.1872	0.0765	0.6465	0.6663

(A) At recall threshold 0.5

	best_setting	threshold	F1	Recall	PR AUC	Precision	ROC AUC	Accuracy
xgboost	undersample	0.5051	0.2149	0.7358	0.1453	0.1258	0.7595	0.7211
random_forest	undersample	0.5000	0.2097	0.7358	0.1317	0.1223	0.7658	0.7123
svm	feature_selection_absolute_mean	0.5000	0.1980	0.5472	0.1505	0.1208	0.7334	0.7701
sgd	feature_selection_absolute_mean	0.5000	0.1938	0.5283	0.1643	0.1186	0.7493	0.7720
baseline	feature_selection_absolute_mean	0.5000	0.1911	0.5283	0.1550	0.1167	0.7521	0.7681
mlp	undersample	0.5000	0.1848	0.6415	0.1633	0.1079	0.7609	0.7065
adaboost	ordinal_encoding	0.4444	0.1822	0.8302	0.1326	0.1023	0.7588	0.6135
decision_tree	undersample	0.5000	0.1626	0.5660	0.3417	0.0949	0.6354	0.6977
naive_bayes	feature_selection_rms	0.9091	0.1607	0.7170	0.1327	0.0905	0.6944	0.6115
knn	undersample	0.5000	0.1323	0.4906	0.1872	0.0765	0.6465	0.6663

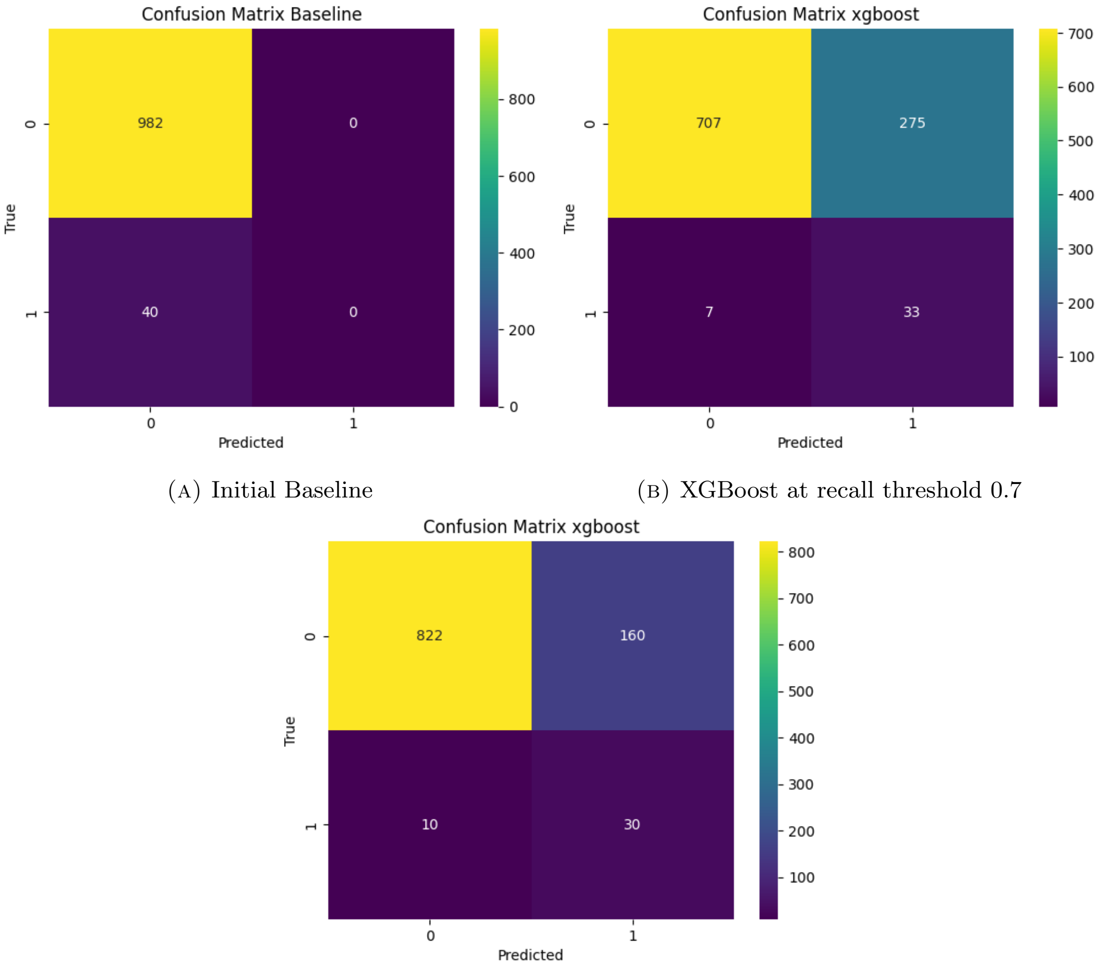
(B) At recall threshold 0.7

Figure 11. Performance of models at best setting and threshold.

Model	Setting	Recall Constraint	Threshold	F1 Score	Recall	Precision	Accuracy
Baseline	vanilla	0	0.5000	0.0000	0.0000	0.0000	0.9609
XGBoost	undersample	0.5	0.5152	0.2609	0.7500	0.1579	0.8337
XGBoost	undersample	0.7	0.5051	0.1897	0.8250	0.1071	0.7241

Table 1. Performance metrics of baseline and final model.

recall threshold could be reduced to lead a decision boundary that reduces false positives from 275 to 160, at the cost of only 3 more false negatives.



(A) Initial Baseline

(B) XGBoost at recall threshold 0.7

(C) XGBoost at recall threshold 0.5

Figure 12. Confusion matrices on test data

4.4. **Future Work.**

Despite respectable performance, the low F1-Score and convex PR curve suggest unresolved data imbalance issues, a frequent challenge [5]. Future solutions could consider advanced undersampling methods like clustering-based undersampling [19, 21] or data augmentation using GANs [11] or transfer learning [6].

Feature selection holds potential for further improvements. Techniques like recursive feature elimination [12] and permutation importance [2] warrant exploration. We should also investigate global optimisations, exploring unsampled + label encoding and advanced models, such as deeper neural networks [25] or hybrid ensembles [21].

5. Conclusion

This report explored multiple models and data preprocessing strategies, coupled with hyper-parameter optimisation. The final XGBoost model achieved a recall of 0.825 on a predominantly negative dataset. Experimentation details provide insights into the stroke phenomena, hinting at future improvement avenues.

Deploying the model in clinical settings presents challenges. Trust is paramount; even with interpretable feature importance, the model's accuracy might be questioned by both patients and physicians. Misdiagnosis is costly and risky, while over-reliance on the model could also lead to serious consequences.

Ultimately, a model is only as effective as the data it is trained on. Discrepancies often exist between training and deployment phases due to for example data shifts and concept drifts. This can affect generalisation and degrade performance over time, a significant challenge in machine learning deployment, especially in medical fields like stroke prediction.

References

[1] Yazan Abdel Majeed, Saria S Awadalla, and James L Patton. Regression techniques employing feature selection to predict clinical outcomes in stroke. *PLoS One*, 13(10):e0205639, 2018.

[2] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[3] Ahmet Kadir Arslan, Cemil Colak, and Mehmet Ediz Sarihan. Different medical data mining approaches based prediction of ischemic stroke. *Computer methods and programs in biomedicine*, 130:87–92, 2016.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[6] Jie Chen, Yingru Chen, Jianqiang Li, Jia Wang, Zijie Lin, and Asoke K Nandi. Stroke risk prediction with hybrid deep transfer learning framework. *IEEE Journal of Biomedical and Health Informatics*, 26(1):411–422, 2021.

[7] Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, Md Mahfujur Rahman, M Shamim Al Mamun, and M Shamim Kaiser. Performance analysis of machine learning approaches in stroke prediction. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*, pages 1464–1469. IEEE, 2020.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[9] Cheng Fan, Meiling Chen, Xinghua Wang, Jiayuan Wang, and Bufu Huang. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in energy research*, 9:652801, 2021.

[10] Valery L Feigin, Michael Brainin, Bo Norrving, Sheila Martins, Ralph L Sacco, Werner Hacke, Marc Fisher, Jeyaraj Pandian, and Patrice Lindsay. World stroke organization (wso): global stroke fact sheet 2022. *International Journal of Stroke*, 17(1):18–29, 2022.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

[13] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

[14] Joseph M Hilbe. Logistic regression. *International encyclopedia of statistical science*, 1:15–32, 2011.

[15] Mohamed Sobhi Jabal, Olivier Joly, David Kallmes, George Harston, Alejandro Rabinstein, Thien Huynh, and Waleed Brinjikji. Interpretable machine learning modeling for ischemic stroke outcome prediction. *Frontiers in neurology*, 13:884693, 2022.

[16] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on artificial intelligence*, volume 56, pages 111–117, 2000.

[17] Josh. Predicting a stroke [shap, lime explainer and eli5]. Kaggle Notebook, https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5, 2021.

[18] Seong-Hwan Kim, Eun-Tae Jeon, Sungwook Yu, Kyungmi Oh, Chi Kyung Kim, Tae-Jin Song, Yong-Jae Kim, Sung Hyuk Heo, Kwang-Yeol Park, Jeong-Min Kim, et al. Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke. *Scientific reports*, 11(1):20610, 2021.

[19] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[20] Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.

[21] Tianyu Liu, Wenhui Fan, and Cheng Wu. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial intelligence in medicine*, 101:101723, 2019.

[22] World Health Organization. The top 10 causes of death. Website, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2024.

[23] World Stroke Organization. Impact of stroke. Website, https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke, 2024.

[24] Gangavarapu Sailasya and Gorli L Aruna Kumari. Analyzing the performance of stroke prediction using ml classification algorithms. *International Journal of Advanced Computer Science and Applications*, 12(6), 2021.

[25] M Sheetal Singh and Prakash Choudhary. Stroke prediction using artificial intelligence. In *2017 8th annual industrial automation and electromechanical engineering conference (IEMECON)*, pages 158–161. IEEE, 2017.

[26] Federico Soriano. Stroke prediction dataset. Website, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, 2021.