Comparative Evaluation of Machine Learning Models for Binary Classification in Healthcare Datasets

Emanoel Agbyani

Abstract

This study evaluates the effectiveness and robustness of multiple machine learning classifiers applied to high-stakes healthcare prediction tasks. Using five binary classification datasets from the UCI Repository: Autism; Maternal Health Risk; Doctor Visits; Cirrhosis; and Thyroid Cancer. We compare Logistic Regression, Support Vector Machines, K-Nearest Neighbor [2], Random Forests, and Gradient Boosted Trees [1] under three data partitioning strategies: 70/15/15; 80/20; and 50/50. Hyperparameters are optimized via grid search cross-validation, and models are assessed using accuracy and related classification metrics. The results demonstrate that ensemble based methods, particularly Random Forests and Gradient Boosted Trees, consistently achieve strong performance and exhibit greater stability across datasets and splits. Meanwhile, simpler models such as Logistic Regression remain competitive on datasets with more linear decision boundaries. Performance varies substantially by dataset which highlights the importance of dataset specific evaluation and model selection in medical applications. Overall, the findings underscore the trade off between predictive performance and interpretability; due to this, it emphasizes the need for careful experimental design when deploying machine learning models in healthcare contexts.

1 Introduction

Throughout this class, we have studied and worked on a multitude of machine learning methods. Furthermore, we have been taught that we can apply these methods to a variety of disciplines and domains. The methods conducted within this are [1]: Logistic Regression; Support Vector Machine Classifier; Random Forest Classifier; K-Nearest Neighbor; and Gradient Boosted Trees. In this study, we will be applying these methods to a high-stakes domain like the healthcare industry and medical field. Compared to other domains, the medical field is consistently filled with inconsistencies alongside complex situations, factors, and variables that are always hard to account for. Due to this, there are huge discrepancies between datasets, which makes it more difficult to apply proper machine learning methods and techniques. From this, we can see why choosing a wide variety of machine learning methods and techniques will be optimal in this situation rather than just linear or only ensemble methods. From the UCI Repository, we are able to study 5 different datasets with a binary classification task: Autism; Maternal Health Risk; Doctor Visits; Cirrhosis; Thyroid Cancer. Each binary classification task represents a realistic medical prediction problem: disease screening; risk stratification; or outcome prediction. Thus, having accurate classification can have significant implications for patient care and resource allocation.

2 Methodology

2.0 Learning Algorithms and Training Procedure

We compare model performance within and across the classifiers using the following experimental procedure. In our healthcare classification project, we evaluate six classifiers across

multiple data partitioning strategies to assess model stability and generalization. Specifically, we experiment with three different data partitions: 70/15/15, 80/20, and 50/50. For each experiment, hyperparameter tuning is performed using grid search cross validation on the training set to identify the optimal model configurations. Once the best hyperparameters are selected, each classifier is retrained on the full training portion of the data and evaluated on the held-out test set. Afterwards, accuracy and performance is then reported using appropriate classification metrics.

2.0.1 Logistic Regression

The Logistic Regression parameter grid explores a wide range of regularization strengths, *C*, to balance underfitting and overfitting across diverse healthcare datasets. Multiple *solvers* are included to ensure compatibility with different penalty terms and dataset sizes. Meanwhile, varying *max_iter* values help guarantee convergence, especially for higher-dimensional medical data. The inclusion of different *penalty* types: l1, l2, elastic net, and None. These different penalty types allow the model to handle multicollinearity and perform implicit feature selection, which is valuable for interpretability in healthcare applications.

2.0.2 Support Vector Machine Classifier

The SVM parameter grid evaluates multiple *kernel* functions to capture both linear and nonlinear relationships commonly present in clinical and demographic features. Similarly to Logistic Regression, different values of the regularization parameter *C* balance underfitting and overfitting across diverse healthcare datasets. The *gamma* options enable flexibility in how individual data points influence the decision boundary. This is particularly important when modeling complex medical patterns.

2.0.3 Random Forest Classifier

For Random Forests, the grid varies the number of trees, *n_estimators*, to study the trade off between model stability and computational cost. Different *max_features* settings control feature randomness and reduce correlation among trees. Meanwhile, enabling or disabling *bootstrap* and *oob_score* allows assessment of generalization performance and robustness. These choices help adapt the model to heterogeneous healthcare datasets with mixed feature types.

2.0.4 K-Nearest Neighbor

The KNN parameter grid focuses on selecting an appropriate number of neighbors, *n_neighbors* to balance sensitivity to noise and generalization. Using the *weight* parameter, both uniform and distance-based weighting schemes are included to assess how proximity should influence classification decisions. Furthermore, multiple *algorithm* choices are explored to ensure computational efficiency across datasets of varying sizes and dimensionality.

2.0.5 Gradient Boosted Trees

Similar to Random Forests, the Gradient Boosted Trees grid examines different numbers of boosting stages, *n_estimators* and learning rates to control the bias variance trade-offs. Additionally, varying the *max_depth* of trees allows the model to capture interactions of differing complexity. These parameters are particularly important for achieving strong predictive performance while minimizing overfitting in high-stakes healthcare classification tasks.

2.0.6 Data Partitions

We decided to choose three different data partitions: 70/15/15, 80/20, and 50/50. The latter two are selected because they are the most commonly used data splits in machine learning and provide a clear benchmark for comparing model performance. Despite being in a high-stakes domain, it is important to use this as a baseline when comparing different models in different datasets. The 70/15/15 split is included to explicitly separate training, validation, and testing sets.

Due to this, it provides more reliable hyperparameter tuning while preserving an unbiased evaluation on unseen data. Furthermore, using multiple partitioning strategies enables us to assess the robustness and stability of the models.

3 Experiments

3.1 Dataset Preprocessing

Autism The Autism Dataset or Autistic Spectrum Disorder Screening Data for Children within the UCI Repository, is a dataset with 20 Features featuring A Score, age, gender, ethnicity, jaundice, autism, country_of_res, used_app_before, result, age_desc, relation, and class. We are predicting based on the autism feature or column. Despite this, there were some null values within the ethnicity and relationship column. The relationship column is what relationship this person has to the autistic patient. Due to the difficult nature of trying to fill in the values for ethnicities, we decided to drop all the null values of 43 within the ethnicity column and we are left with one null value in the age column. Due to it only being one value, it was concluded that minimal model performance would be affected if we drop this one value. From this, we went from a total of 292 instances or entries to 248, which means we preserved a majority of the original dataset.

Maternal Health Risk The Maternal Health Risk Dataset within the UCI Repository, is a dataset with 1013 instances or entries and 6 Features featuring Age, Systolic BP (blood pressure), Diastolic BP (blood pressure), BS (Blood Glucose Levels), BodyTemp, HeartRate, and RiskLevel. We are predicting based on the RiskLevel feature or column. The only categorical or non-numerical column is the RiskLevel column, which has values of ‘low risk’, ‘mid risk’, and

‘high risk’. Due to this, we used a one shot encoding of values of 1, 2, and 3 to indicate the following risk levels.

Doctor Visits The Doctor Visit Dataset or National Poll on Healthy Aging (NPHA) within the UCI Repository, is a dataset with 714 instances or entries and 14 Features featuring Number_of_Doctors_Visited, Age, Physical_Health, Mental_Health, Dental_Health, Employment, Factors Keeping Patients from Sleeping, Prescription_Sleep_Medication, Race, and Gender. We are predicting based on the Number_of_Doctors_Visited feature or column. Due to all these values having integer values, no further preprocessing was deemed necessary. We are predicting based on the Number of Doctors Visited feature or column.

Cirrhosis The Cirrhosis Dataset or Cirrhosis Patient Survival Prediction within the UCI Repository, is a dataset with 418 instances or entries and 17 Features featuring Number of Days, Status, Drug, Age, Sex, Variety of Physical Symptoms, Liver Functions, Enzymes, Hematologic Elements, and Stage. We are predicting based on the Stage feature or column. When checking for Null or Nan Values, we saw there were about 105 values missing in the Drug column. Due to the dataset coming from a Mayo Clinic study, it was decided that trying to predict missing would not be a feasible strategy and concluded to drop all the missing values. Furthermore, the values in the age column were based on how many days old the patients were. Due to this, we made sure to convert them back into years old. Upon closer inspection of the categorical variables, we decided to split them into two groups before recombining in a final dataset: the One Shot Encoding and Fake Integer Groups. For the One Shot Encoding Group, these were either Yes No Questions or categorizing a patient into three statuses. Due to this, the mapping given consisted of 0’s, 1’s, and 2’s. For the Fake Integers Group, we decided to de-stringify the numbers and make them into integers. Despite this, we saw 30 and 28 null values for Cholesterol and Tryglicerides. Due to

this, we decided to further drop them as predicting cholesterol and triglyceride presence within a patient would be difficult and could disrupt maintaining true model performance on the overall dataset. Due to this, we were left with 282 Entries left.

Thyroid Cancer The Thyroid Cancer Dataset orDifferentiated Thyroid Cancer Recurrence within the UCI Repository, is a dataset with 383 instances or entries and 16 Features featuring Age, Gender, Smoking, Variety of Enzymes, Variety of Tests, Stage, Response, and Recurred. We are predicting based on the Recurred feature or column. All of the columns were string values besides Age which was an integer. Due to this, we had to process them into some form of integer value through One Shot Encoding. A majority of the values were either yes and no, or different stages of their column. Due to this, we could apply a 0, 1, 2, and etc. to them.

3.2 Tables

Best Performing Parameter Across All Datasets

Model	Best Avg Acc	Trial	Best Hyperparameters
Boosted Gradient Forest	95.67	Trial 2	{'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 1000}
K-Nearest Neighbor	96.52	Trial 2	{'bootstrap': True, 'max_features': 'sqrt', 'n_estimators': 50, 'oob_score': True}
Logistic Regression	96.24	Trial 3	{'C': 0.05, 'max_iter': 100, 'penalty': 'l1', 'solver': 'saga'}
Random Forest Classifier	96.52	Trial 2	{'bootstrap': True, 'max_features': 'sqrt', 'n_estimators': 50, 'oob_score': True}
Support Vector Machine Classifier	94.22	Trial 2	'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}

Best Performing Model per Dataset

Dataset	Best Model	Avg Acc	Trial	Best Hyperparameters
Autism	Logistic Regression	83.10	Trial 2	{'C': 0.001, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}
Cirrhosis	Support Vector Machine Classifier	53.50	Trial 1	{'C': 0.05, 'gamma': 'scale', 'kernel': 'linear'}
Doctor Visits	Logistic Regression	51.51	Trial 2	{'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
Maternal Health Risk	Boosted Gradient Forest	84.07	Trial 3	{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 100}
Thyroid Cancer	Random Forest Classifier	96.52	Trial 2	{'bootstrap': True, 'max_features': 'sqrt', 'n_estimators': 50, 'oob_score': True}

Best Performing Hyperparameters per Model

Model	Best Avg Acc	Count	Best Hyperparameters
Boosted Gradient Forest	95.67	1	{'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 1000}
K-Nearest Neighbor	96.52	1	{'bootstrap': True, 'max_features': 'sqrt', 'n_estimators': 50, 'oob_score': True}
Logistic Regression	96.24	3	{'C': 0.05, 'max_iter': 100, 'penalty': 'l1', 'solver': 'saga'}
Random Forest Classifier	96.52	7	{'bootstrap': True, 'max_features': 'sqrt', 'n_estimators': 50, 'oob_score': True}
Support Vector Machine Classifier	94.22	2	'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}

Average Model Accuracy per Data Partition: 70/15/15

Model	Trial 1	Trial 2	Trial 3	Avg Acc
Boosted Gradient Forest	71.47	71.09	70.34	70.94
K-Nearest Neighbor	70.13	72.23	68.80	70.25
Logistic Regression	69.07	70.07	70.13	69.76
Random Forest Classifier	71.10	70.16	68.92	69.76
Support Vector Machine Classifier	69.92	72.04	68.70	70.02

Average Model Accuracy per Data Partition: 80/20

Model	Trial 1	Trial 2	Trial 3	Avg Acc
Boosted Gradient Forest	72.13	72.19	70.58	71.63
K-Nearest Neighbor	72.04	70.78	69.38	70.73
Logistic Regression	68.54	67.91	67.61	68.02
Random Forest Classifier	71.17	69.96	68.68	69.93
Support Vector Machine Classifier	68.76	70.75	66.00	68.34

Average Model Accuracy per Data Partition: 50/50

Model	Trial 1	Trial 2	Trial 3	Avg Acc
Boosted Gradient Forest	70.32	69.36	69.55	69.74
K-Nearest Neighbor	68.84	70.58	68.88	69.43
Logistic Regression	67.38	69.06	68.4	68.28
Random Forest Classifier	67.49	70.29	66.47	68.68
Support Vector Machine Classifier	69.79	69.86	68.41	69.36

4 Discussion and Conclusion

The experimental results demonstrate that model performance varies substantially across both datasets and data partitioning strategies. Overall, ensemble-based methods such as Random Forest Classifier and Gradient Boosted Trees tended to achieve the strongest performance; meanwhile simpler linear models remained competitive on certain datasets.

Across all datasets, the Random Forest Classifier and K-Nearest Neighbor models achieved the highest average accuracy of 96.52%. This indicates their ability to capture complex patterns in heterogeneous healthcare data. Also, the Boosted Gradient Forest performed strongly and achieved a best average accuracy of 95.67%; suggesting that boosting is effective at correcting residual errors across datasets. In contrast, the Support Vector Machine achieved slightly lower peak performance. Due to its sensitivity to feature scaling and kernel selection in diverse medical datasets, this is the reasoning behind its poor performance.

When examining the best performing model per dataset, results varied significantly. Logistic Regression performed best on the Autism and Doctor Visits datasets; due to this, it suggests that these problems may be well approximated by linear decision boundaries and benefit from regularization for feature selection and interpretability. The Support Vector Machine performed best on the Cirrhosis dataset; although, overall accuracy of 53.50% was relatively low. This reflects the inherent difficulty of this prediction task and possible class imbalance or limited predictive signal. For the Maternal Health Risk dataset, the Boosted Gradient Forest achieved the highest accuracy. The Random Forest Classifier performed exceptionally on the Thyroid Cancer

dataset. It achieved the highest accuracy of 96.52%, which suggests that ensemble tree methods are particularly effective for structured clinical data with complex interactions.

Furthermore, analysis of best performing hyperparameters per model reveals that Random Forests demonstrated the greatest stability. Additionally, this is further supported with the same hyperparameter configuration emerging as optimal across multiple trials. This consistency suggests robustness and reliable generalization across different healthcare tasks. Also, Logistic Regression showed repeatable optimal configurations; due to this, it reinforces its reliability as a baseline model. In contrast, Boosted Gradient Forest and Support Vector Machine Classifier exhibited more variability in optimal hyperparameters. This may be due to their sensitivity to dataset characteristics and training conditions.

When comparing average model accuracy across data partitions, the 80/20 split consistently produced the highest average accuracies across models, followed by the 70/15/15 split, with the 50/50 split yielding the lowest overall performance. This trend indicates that increased training data generally improves model learning and generalization. This consensus is the general train of thought. Despite this, performance differences across splits were relatively moderate and within expected values. This suggests that the models are reasonably robust to changes in data availability.

In conclusion, these findings highlight that no individual model outperforms others across all healthcare datasets. Ensemble methods, Random Forests and Gradient Boosted Trees, consistently demonstrated strong performance and robustness; due to this, it makes them well suited for high-stakes domain tasks such as medical classification tasks. However, simpler

models such as Logistic Regression remain valuable due to their interpretability and competitive performance on certain datasets. These results emphasize the necessity of dataset specific evaluation and careful experimental design when applying machine learning methods within the medical field and healthcare environment as a whole.

6 Potential Problems

Despite the results, there are several limitations and potential problems that should be considered when interpreting this study. First, many of the healthcare datasets used may suffer from class imbalance and bias. Due to the nature of a high-stakes domain of the healthcare industry, it is difficult to capture consistent 80%+ Accuracy throughout all datasets. Due to this, the potential of finding a performing model across all datasets is miniscule. Second, dataset size and feature quality vary across tasks. Not all of the datasets chosen were classifying on whether or not a patient had their respective disease or diagnosis, but rather there were Doctor Visits and Maternal Health Risk which are not diseases or a diagnosis. Furthermore, more complex conditions such as cirrhosis may limit the models' ability to learn meaningful patterns and contribute to lower predictive performance.

Additionally, the study relies primarily on accuracy as the main evaluation metric. With the 5 Datasets * 5 Models * 3 Data Partitions * 3 Trials, we would get 225 Uniquely tuned Models. Due to this, actively documenting Confusion Matrixes and F1 Scores would help reinforce which models performed better; but rather, the inconsistent dataset features and not only classifying whether a person has a diagnosis or disease would already reinforce the expectation that there is not one single model that outperforms other models.

Finally, the results may be sensitive to data partitioning strategies and hyperparameter tuning procedures. Although multiple splits were evaluated to mitigate this issue, the findings may not fully generalize to external populations or real world clinical settings. Future work should include external validation, fairness analysis, and domain specific evaluation metrics to address these challenges.

7 Bonus Points

Due to the scope and depth of its empirical evaluation across multiple healthcare prediction tasks, the project merits bonus points.. Specifically, the study conducted a comprehensive comparison of five distinct machine learning classifiers: Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, and Gradient Boosted Trees. These models span linear, non-parametric, and ensemble learning methods. Due to this, this diverse set of models enables meaningful analysis of trade offs between interpretability, robustness, and predictive performance in high-stakes medical domains.

In addition, the study was performed on five different real world healthcare datasets obtained from the UCI Machine Learning Repository: Autism, Maternal Health Risk, Doctor Visits, Cirrhosis, and Thyroid Cancer. These datasets represent varied clinical objectives. These include disease screening, risk stratification, and outcome prediction; additionally, these differ significantly in feature types, dataset size, and difficulty. This diversity strengthens the generalizability and rigor of the experimental findings.

Beyond model and dataset diversity, extensive experimentation was conducted using multiple data partitioning strategies and systematic hyperparameter tuning via grid search cross validation. Furthermore, validation sets were not discussed in this class. This resulted in a large

number of trained and evaluated models, enabling robust conclusions about model stability and performance consistency.

Ultimately, the breadth of classifiers, the number of datasets, and thorough experimental design constitute a substantial empirical effort that exceeds baseline project requirements and justifies consideration for bonus points.

8 References

[1] Rich Caruana Caruana and Alexandru Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms

[2] Rich Caruana and Nikos Karampatziakis and Ainur Yessenalina, An Empirical Evaluation of Supervised Learning in High Dimensions