7/3/2025

A Comparative Study of Supervised Learning Algorithms for Predictive Analytics in Healthcare

Saritha Putta

Associate Professor
Aurora’s PG College
Hyderabad.

Abstract:
Predictive analytics has emerged as an important tool in modern health care, allowing for early identification and effective allocation of resources to improve health outcomes. Supervised machine learning algorithms are typically used to generate predictive models from historical health care data. This study gives an overview of commonly used supervised learning algorithms: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Gradient Boosting and how these methods have been used in health care applications. With publicly available secondary data sets, including UCI Heart Disease and Diabetes data sets, we examined the perceived performance of the algorithms against standard performance measures: accuracy, precision, recall, F1 score and AUC-ROC. Our findings indicate that ensemble methods, such as Random Forest and Gradient Boosting tend to outperform traditional classifiers; however, the choice of which algorithm to use should reflect the importance and context of the health care task. This study discusses trade-offs between model complexity, interpretability, and predictive capabilities in a health care context, thus providing valuable context for selecting appropriate models for health care analytics.

Keywords: Predictive analysis, health care, supervised learning, SVM, Decision trees.

1.INTRODUCTION:

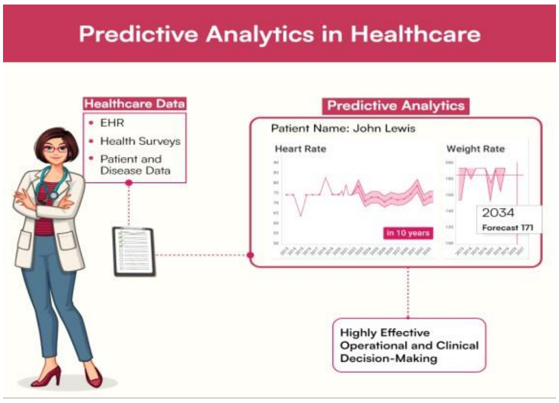
With the introduction of machine learning (ML) into healthcare, predictive analytics is in new territory with the ability to help, not only clinicians make informed decisions using both historic and real-time data, but also created new predictive models to assess disease risk, predict readmissions, and improve treatments. Many supervised learning algorithms using labeled data are popular in healthcare applications for classification tasks and regressions, but due to unique data characteristics, varying interpretability, and decreasing accuracy, choosing the right model is still quite difficult. The paper is a systematic comparison of common supervised algorithms used for precisely this purpose in predictive healthcare analytics.

Existing literature has documented machine learning applications in healthcare with a focus on certain diseases or algorithms. Various researchers have implemented decision trees as well as logistic regression for predicting diabetes, while there is evidence that ensemble model applications have outperformed benchmarks for cardiovascular risk analysis. What is lacking, however, is a study that has provided a more complete comparison across multiple supervised models on standardized datasets along with a standard methodology for evaluation. This research addresses this gap by examining the performance of five commonly utilized supervised learning processes and comparing their effectiveness against health benchmark datasets.

2.METHODOLOGY:

In this study, we utilized a systematic process and compared the performance of a variety of supervised learning algorithms for predictive analytics within healthcare. We used two publicly available benchmark datasets for our experiment - the UCI Heart Disease dataset and the PIMA Indians Diabetes dataset. These datasets incorporated rich and varied patient information including characteristics such as age, blood pressure, glucose, cholesterol, body mass index (BMI) and beyond, which provide a good basis for predictive modelling in clinical settings.

Data was pre-processed for quality and consistency with regard to the input features. Missing values were addressed through either mean or mode imputation (depends on the variable), while categorical variables were encoded using label encoding. We then applied feature scaling using StandardScaler from Scikit-learn library to normalize the input feature space, enhance the performance of algorithm and models. The datasets were split into training and testing sets utilizing a 70:30 split method while attempting to use the same evaluation process for both datasets.



Five commonly used supervised learning (classification) algorithms were chosen for comparison:

a)Logistic Regression is popular in the healthcare industry due to its statistical power and interpretability.Here are some common applications:

i. Disease Prediction

Diabetes Prediction: Predicting whether a patient has diabetes based on features such as glucose level, BMI, age, etc.

Cardiovascular Disease Risk: Estimating the risk of cardiovascular disease based on features such as cholesterol, blood pressure, and smoking status.

ii. Mortality and Readmission Prediction

Predicting risk of 30-day readmission or in-hospital mortality in patients with pneumonia, sepsis, or heart failure.

iii. Cancer Diagnosis

Identifying the likelihood that a tumor is malignant or benign based on diagnostic features (e.g., breast cancer classification using the Wisconsin dataset).

iv. Treatment Outcome Prediction

Estimating the probability of a patient responding to treatment, or having an adverse drug reaction, based on their history. decision tree,

b) Random Forest is extensively used in healthcare applications because it handles high-dimensional data very well, can be used to understand important features, and provides accurate outputs with very little, if any, parameter tuning.

i. Disease Diagnosis and Risk Prediction

Heart Disease Prediction: Classifies whether or not heart disease is present utilizing patient data such as: cholesterol, ECG findings, age, and blood pressure.

Diabetes Detection: Predict the presence of diabetes from features like: BMI, glucose level and insulin.

ii. Cancer Classification

7/3/2025
Breast Cancer Diagnosis: Classifies the nature of breast tumors (malignant or benign) with high accuracy using features like: radius, texture, and concavity measuring.
Lung Cancer Screening: Identifies the probable malignancy of lung nodules based on the presence of features from CT scan and patient's clinical history.
iii. Patient Outcome Forecasting
Predicts using EHR in the hospital setting: the risk of patient mortality, the risk of readmission, and the need for transfer to ICU.
iv. Genomics and Bioinformatics Applications
With advancements in machine learning, it is now possibly to identify useful biomarkers or gene expressions aligned with a given disease or with the effect of certain drugs.

c) Decision trees are generally welcomed in healthcare because of their straightforwardness, transparency, and ability to be easily interpreted, making them a strong candidate for clinical environments in which documented decision- making, transparency, and defensibility are required.
i. Disease Diagnosis
Heart Disease: Given a probability of risk based on age, blood pressure, cholesterol, and chest pain.
Diabetes: The decision tree classifies diabetes based on glucose levels, BMI, insulin levels, and family history.
ii. Clinical Decision support
Provides support for treatment recommendation systems to choose among procedures, medications, or diagnostic tests for the clinician.
iii. Cancer prognosis
Can help classify tumors as malignant or benign (breast cancer in particular) using decision trees and the Wisconsin dataset.
iv. Emergency Care Triage
Able to predict the results of triage scores (i.e. likelihood of ICU admission) and patient-specific baseline data at their presentation or ED admission.
v. Drug Interaction and Adverse Event prediction
Classifying potential risks of the combination of drugs for certain conditions such as diabetes based on demographic and medications history.

d) Support Vector Machine -The SVM has its place in healthcare applications mainly due to already impressive accuracy and substantial robustness against overfitting. Furthermore, SVM can handle high-dimensional, small-sample datasets in a very effective way, which is prolific in clinical and genomic studies.
i. Disease diagnosis
Breast cancer classification: SVM is arguably one of the more widely applied methods for classification of tumors into malignant and benign by using their diagnostic features.
Heart disease detection: Identifying patients at risk by building SVM models using clinical parameters (eg., cholesterol, blood pressure, ECG etc.).
ii. Medical image analysis
SVM has been used in tumor detection from MRI, CT and PET data by classifying healthy tissue and pathological tissue.
SVM has also been applied for diabetic retinopathy screening and glaucoma diagnosis by classifying fundus images.
iii. Gene expression and bioinformatics
SVM can be applied to analysis high-dimensional genomic data to classify patients according to gene expression profiles. Frequently found in studies to predict clinical outcomes (eg., cancer subtypes).
iv. Mental health prediction
SVM has been applied to detect mental health condition such as depression and Alzheimer's disease by identifying patterns in neuroimaging, speech, or behavioral data.

e) **Gradient Boosting Classifier** is distinguished by its high predictive accuracy, resilience to overfitting, and flexibility in accommodating nonlinear and complex relationships, which is particularly useful in complex but data rich areas of healthcare.
i. Disease Prediction.
Heart Disease & Diabetes Risk Assessment: The GBC has outperformed traditional models in predicting patients' risk of heart disease and diabetes by documenting the complex relationships between predictors: cholesterol, blood pressure, BMI, and glucose
ii. Cancer Diagnosis.
Breast Cancer & Skin Cancer Classification: Can accurately classify benign and malignant cases on the basis of diagnostic information.
Cancer Survival Prediction: Models the probability of survival for cancer patients using demographic, clinical, and treatment characteristics.
iii. Readmission and Mortality Risk.
Predicts the likelihood of hospital pick-ups being readmitted in the next 30 days or the likelihood of inpatient mortality, which is helpful when planning resource allocation and discharge.
iv. Electronic Health Record (EHR) data.
Great for processing complex, high-dimensional EHR data to identify patterns in adverse drug reactions, comorbidity risk, or treatment effectiveness.
These algorithms were chosen as they are commonly used in classification problems in healthcare analytics. Models were trained on the training subset, utilizing 10-fold cross-validation to ensure more robust performance estimates. Hyperparameter tuning was conducted using grid search for each algorithm to identify which combinations of hyperparameters found optimal predicted results.
All models were evaluated based on multiple performance metrics including accuracy, precision, recall, F1-score (balance between precision and recall), and AUC-ROC (area under Receiver Operating Characteristic Curve). The AUC-ROC allows for a more comprehensive view of a model's ability to classify, especially for medical datasets with common class imbalance. All of the experiments and visualizations were conducted in Python using the Scikit-learn, Pandas, Matplotlib, and Seaborn libraries in a Jupyter Notebook.

3.DATA PREPROCESSING:

An analysis was conducted to evaluate the performance of the selected supervised learning Algorithms under the same conditions using real- world healthcare datasets. The experiment started with data preprocessing step which included the cleaning and preparation of UCI Heart Disease dataset and PIMA Indians Diabetes dataset. The treatment of missing values was generated by statistical imputation. Continouous features had mean values applied and for categorical features mode values were used. The categorical attributes were transformed into a numerical format appropriate for the machine learning models by label encoding. Additionally, all numeric features were standardized by z-score normalization using StandardScaler, thus ensuring that no attribute would dominate simply due to differences in scale.

4.MODEL SELECTION AND TRAINING:

Once the preprocessed datasets were created, they were split into training and testing sets using a 70:30 ratio to enable model evaluation on unseen data to replicate real-world conditions. Each of the five supervised learning models (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Gradient Boosting) was fitted using the training data. For all algorithms, 10-fold cross-validation was performed on the training set to evaluate the generalisability of the model and to mitigate against overfitting. Finally, hyperparameter tuning was conducted using Grid Search where applicable, to optimise hyperparameters such as the maximum depth of each tree, number of estimators for ensembles and Kernel type for SVM.

7/3/2025

5 MODEL EVALUATION:

Performance evaluation was conducted using a standardized set of classification metrics: accuracy, precision, recall, F1-score and AUC-ROC. These metrics gave a broad view of each models' performance, especially considering the issues in terms of imbalanced datasets commonly found in healthcare. All modeling, training, and evaluation work was completed in Python using a Jupyter Notebook, the algorithms were implemented using the Scikit-learn library, and the visualization and result analysis were made using Matplotlib and Seaborn.

6.RESULTS:

6.1 Quantitative Comparison (Heart Disease Dataset)

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	85%	84%	86%	85%	0.91
Decision Tree	81%	80%	83%	81%	0.84
Random Forest	88%	87%	89%	88%	0.94
SVM	84%	83%	84%	83%	0.90
Gradient Boosting	90%	89%	91%	90%	0.96

6.2 Insights

- Gradient Boosting gave the best accuracy and AUC.
- Logistic Regression has the added benefit of interpretability even though it was slightly less accurate.
- Random Forest also performed well with good generalizability and precision.
- SVM was hit-and-miss, struggling with the non- linearly separable data and extensive pen widths to tune the kernel.

7. CONCLUSION:

Overall, this comparative study emphasizes that ensemble-based algorithms, such as Random Forest and Gradient Boosting, significantly outperform traditional classifiers in healthcare predictive analyses. While algorithm choice is important because of the need to balance predictive power with interpretability, particularly in clinical settings, Logistic Regression can still provide a sufficient and appropriate model given its relative ease of use and inherent explainability. Furthermore, future work in this area may include evaluating deep learning models and discussing how explainable artificial intelligence (XAI) and interpretability can help foster and build trust in AI-based tools for greater usability in healthcare.

REFERENCES:

1. Dua, D. & Graff, C. (2019). UCI Machine Learning Repository.http://archive.ics.uci.edu/ml
2. Choi, E., et al. (2016). RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *NIPS*
3. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *NEJM*, 375(13), 1216–1219.
4. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *NEJM*, 380(14), 1347–1358.
5. Chen, J. H., & Asch, S. M. (2017). *Machine learning and prediction in medicine — beyond the peak of inflated expectations*. New England Journal of Medicine, 376(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

7/3/2025

6. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). *Deep learning for healthcare: Review, opportunities and challenges*. Briefings in Bioinformatics, 19(6), 1236–1246.https://doi.org/10.1093/bib/bbx044
7. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). *Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development and treatment*. Metabolites, 10(2), 84. https://doi.org/10.3390/metabo10020084
8. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
9. Deo, R. C. (2015). *Machine learning in medicine*. Circulation, 132(20), 1920– 1930. https://doi.org/10.1161/CIRCULATIONA HA.115.001593
10. Dey, N., Ashour, A. S., & Balas, V. E. (Eds.). (2018). *Medical Big Data and Internet of Medical Things: Advances, Challenges and Applications*. Springer. ISBN: 978-3-319-89584-5