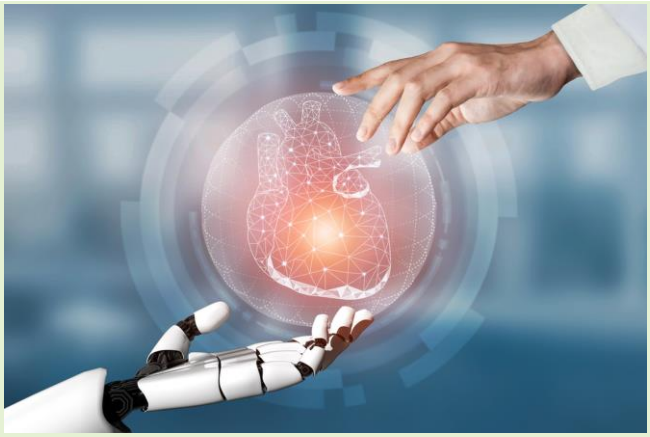
Building a Predictive Medical Diagnostic System using Advanced Machine Learning Models

Abstract— **This project was conducted to develop a predictive medical diagnostic system using machine learning techniques on a dataset of symptoms and diseases (prognosis). Disease prediction is a crucial aspect of healthcare, enabling accurate and timely diagnosis to improve patient outcomes. The dataset consists of 132 symptoms and 41 possible disease outcomes. The initial phase involved data preprocessing and exploratory data analysis to identify imbalance in the dataset, missing, and inconsistent values and target encoding as required for model development.**



Three advanced ensemble models— AdaBoost, LightGBM, and XGBoost— were employed to build the predictive models. The models were evaluated using performance metrics, such as accuracy, precision, recall, F1 score, ROC-AUC score, and computational efficiency such as model training and prediction time. Hyperparameter tuning was specifically performed on AdaBoost model to optimize its parameters. The models achieved high accuracy—approximately 98%—likely due to the training dataset being significantly larger than the testing dataset.

This report outlines the methodology, model development, hyperparameter tuning process, and results, providing insights into model performance for disease prediction based on symptoms.

Keywords— AdaBoost, Classification, Confusion Matrix, Disease Diagnosis, Ensemble Models, Healthcare, LightGBM, Machine Learning, Predictive Analytics, XGBoost.

1. Introduction

Machine learning refers to creating and using models that are learned from data. Machine learning algorithms apply various optimization, statistical, and probabilistic techniques to learn from data that was generated from past experiences and deploy it in decision-making (Uddin et al., 2019). Typically, the goal of machine learning is to use existing data to develop models that can be used to predict various outcomes for new data. For this project, we will focus on supervised machine learning, this is because the model is trained on labelled data.

The emergence of machine learning techniques has led to advancements across various domains, including finance, marketing, transportation, and healthcare.

2. APPLICATION

In recent times, machine learning (ML) has had tremendous impacts in the healthcare industry. Diseases and health-related problems can have a negative impact and sometimes may lead to death if ignored. Therefore, early detection of these maladies and intervention can mitigate these effects. Traditional diagnostic methods can have many limitations such as time consumption, labour-intensive, and often, data is too large and/or complex to generate insights and hidden patterns. The emergence of machine learning has proven to be revolutionary in different ways, for example, by improving diagnostic accuracy through model predictions, enabling personalized treatment plans by analyzing patient records and clinical trials results, and decision-making by analyzing available data to unravel hidden patterns and relationships in the data. This analysis would be beneficial to hospitals, public health specialists, research institutes, animal health organizations, and private sectors.

3. AIMS AND OBJECTIVES

The project aims to develop a predictive medical diagnostic system using advanced machine learning models on a labelled dataset of 132 symptoms (features) and 41 disease outcomes. The study aims to examine the ability of machine learning algorithms to predict disease outcomes based on patterns and relationships among the symptoms in the dataset.

The scope of the study includes implementing complete predictive analytics pipeline from data preprocessing to model

development and to evaluate model performance using established performance metrics to identify the most efficient model for real-world applications.

RESEARCH QUESTIONS:

1. How effective are machine learning models in predicting medical diagnoses?
2. How does hyperparameter tuning impact the performance of ensembles models?
3. What does exploratory data analysis (EDA) play in understanding the relationships between symptoms and diseases?

4. DATASET DESCRIPTION

The dataset is designed for predicting medical diagnoses based on patient symptoms. Each row represents a unique patient record, where their symptoms are listed alongside the corresponding medical diagnosis (prognosis).

4.1 COLUMNS OVERVIEW

Symptom Columns: There are 132 columns, each representing a specific symptom. These columns contain binary values which suggests some form of preprocessing has been carried out on the dataset:
1 indicates the presence of a symptom and 0 indicates the absence of a symptom.
These symptoms could range from physical signs (like "skin rash") to physiological conditions (like "joint pain") and other indicators of disease. Examples of symptom columns:
Itching: Whether the patient reported itching.
skin_rash: Whether the patient had a skin rash.
shivering: Whether the patient experienced shivering.
stomach_pain: Indicates if stomach pain was present.
Acidity: Indicates acidity symptoms.
The dataset includes other symptoms that cover a wide range of medical conditions, ensuring comprehensive coverage of potential diseases.
Target Column (prognosis): The target column specifies the disease diagnosed for the patient. This is a categorical column with 41 unique disease classes, representing various medical conditions. These are the output variables we aim to predict based on the symptoms.

4.2 GENERAL OVERVIEW OF THE DATA

Size: The training dataset has 4,920 rows (patient records) and 134 columns (patients' symptoms and prognosis), while the testing dataset has 42 rows and 133 columns.

Feature Types: The symptom columns are binary features (values: 0 or 1).

Target Variable: The prognosis column is categorical, containing labels for 41 different diseases.

Class Balance: The classes (diseases in prognosis) are evenly distributed, meaning each disease is represented by an equal number of records in the training data. This avoids class imbalance issues during model training.

Purpose of the Dataset: This dataset is designed to train machine learning models to predict the disease outcomes (prognosis) based on symptoms. For example, if a patient has symptoms like "itching", "skin rash", and "shivering", the model might predict "Fungal Infection" as the most likely disease. By analyzing the patterns and combinations of symptoms, the model will identify and differentiate among 41 diseases.

4.3 KEY INSIGHTS FOR FEATURE USAGE

Each symptom column is a feature used by the model to learn patterns associated with diseases.

The combination of symptoms is critical. For instance, "itching" alone may suggest multiple diseases, but when combined with "skin rash", it may narrow down the possibilities.

The balanced nature of the dataset ensures the model isn't biased toward over-predicting common diseases.

5. PREDICTIVE ANALYTICS

Predictive analytics is the use of data to predict future trends and events. It uses historical data to forecast potential scenarios that can help drive strategic decisions. (*Harvard Business School Online, 2021*). In this project, a systematic data pipeline was implemented to transform the raw data into useful information. The project workflow included the following

5.1 DATA SOURCING

The dataset used for this project comprises of 132 symptoms (features) and 41 disease outcomes (target). The dataset comprised of 4920 rows, with each record representing a unique patient's record. These datasets were obtained from accumulated patients' records to represent a comprehensive set of disease-symptom relationships.

5.2 DATA PREPROCESSING

Handling Missing and Inconsistent Values: The dataset was checked for missing values and inconsistencies and upon observation, none were found, this step basically confirms the quality of the data.
Target Encoding: The target variable (prognosis) in the dataset was label encoded to convert categorical data (disease names) into numerical values which can be understood by machine learning models.
Data Reshuffling: The data was shuffled to improve randomness and to avoid bias during model training.

5.3 EXPLORATORY DATA ANALYSIS(EDA)

Thorough exploratory data analysis was conducted on the dataset to understand the characteristics of the dataset. The data was checked for class imbalance, symptom prevalence, and correlation analysis.

Symptoms Distribution: A clear visualization showing symptoms distribution throughout the dataset.

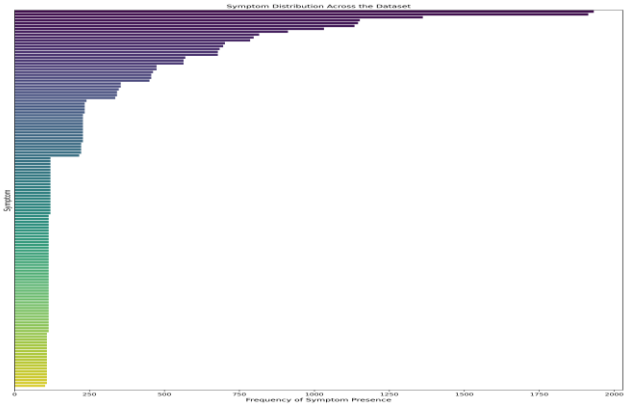


Figure 1: Symptom frequency across the dataset.

The plot above shows the frequency of symptoms across the dataset. With fatigue as the most frequently reported and fluid overload as the least frequently reported.

Class Imbalance: The dataset is relatively balanced as all the classes were equally represented. Upon EDA, it was inferred that the data was balanced, and disease occurrence were evenly distributed throughout the dataset, with each disease occurring 120 times

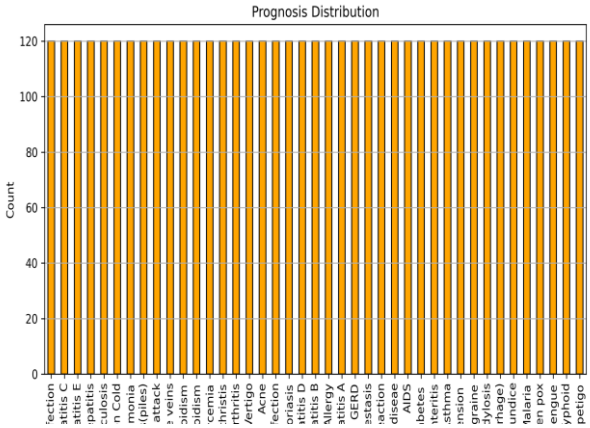


Figure 2: Class (prognosis) Distribution across the Dataset

The figure 2 above shows the count of each class (prognosis) category across the dataset. It can be observed from the plot, that there is no class imbalance in the dataset.

Correlation: Correlation heatmaps were used to show the relationship between symptoms and prognosis in the dataset to provide a clearer insight into their association.

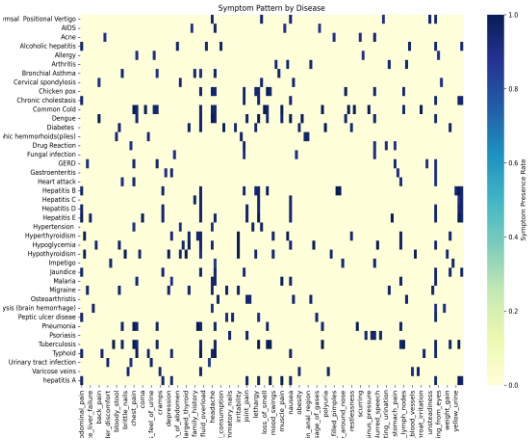


Figure 3: Symptoms Patterns by Disease

The heatmap in figure 3 above shows the correlation between prognosis categories and symptoms. The rows represent prognosis and the columns represent symptoms.

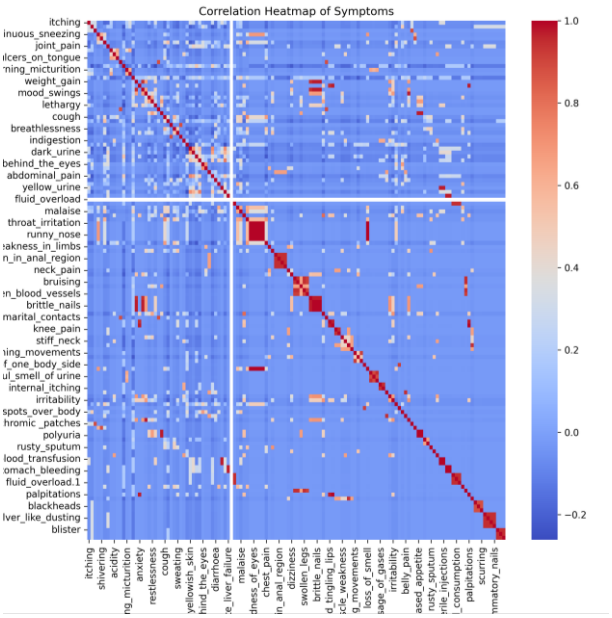


Figure 4: Symptoms Correlation Heatmap

The plot in figure 4 shows correlations among symptoms.

5.4 MODEL DEVELOPMENT

Model Selection: Advanced ensemble models—AdaBoost, LightGBM, and XGBoost— were selected for the project.

Hyperparameter Tuning: Grid Search was employed to optimize hyperparameters for the AdaBoost model, focusing on the number of estimators and learning rate. This tuning improved accuracy and overall model performance.

Model Training and Testing: The training data was used to fit the models, with hyperparameter tuning applied to the AdaBoost model to get the best parameters for the data.
Model Evaluation: The performance of the models was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. As an added measure of model performance, training and prediction times were recorded to compare computational efficiency across models.

6 MOTIVATION FOR MODEL SELECTION

While LightGBM is excellent at handling huge datasets with quicker training periods, XGBoost is renowned for its accuracy and feature importance assessments and AdaBoost is widely used for addressing binary classification challenges. The high dimensionality of symptom data is a good fit for all three models, which provide accurate predictions in a range of situations.

6.1 ADABOOST CLASSIFIER

AdaBoost classifier is a meta-estimator first fits a classifier on the original dataset, after which it fits more copies of the classifier on the same dataset with the weights of instances that were incorrectly classified changed so that later classifiers concentrate more on challenging cases.
(*Freund & Schapire, 1995)*
The fundamental idea behind AdaBoost is to fit a series of weak learners—that is, models that are just marginally superior to random guessing, like tiny decision trees—to iteratively altered datasets. The final prediction is then generated by combining all their predictions using a weighted majority vote (or sum). (*Scikit-learn, n.d.)*
Since its inception, AdaBoost has become a widely adopted technique for addressing binary classification challenges.

6.2 LIGHTGBM CLASSIFIER

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed for faster training speed and higher efficiency, lower memory usage, better accuracy, and is capable of handling large-scale data. *(LightGBM Developers, n.d.).*

6.3 XGBOOST CLASSIFIER

XGBoost is a scalable ensemble technique that has demonstrated to be a reliable and efficient machine learning challenge solver. It is used for two reasons: execution speed and model performance. Execution speed is critical when working with large datasets, as it enables the development of models that are both efficient and high-performing. Additionally, model performance is crucial for ensuring accurate predictions and outpacing other approaches. *(Simplilearn, n.d.)*

7 RESULTS AND DISCUSSIONS

The performance of the three advanced ensemble models—AdaBoost, LightGBM, and XGBoost— were evaluated using multiple performance metrics which include accuracy, precision, recall, F1-score, and ROC-AUC. Also, training and prediction time for each model were calculated to assess computational efficiency.

Models	AdaBoost	AdaBoost Optimized	LightGBM	XGBoost
Accuracy	0.1429	0.9762	0.9762	0.9762
Precision	0.1226	0.9878	0.9878	0.9878
Recall	0.1463	0.9878	0.9878	0.9878
F1-Score	0.1232	0.9837	0.9837	0.9837
ROC-AUC score	0.6146	1.0000	1.0000	0.9991
Training time	1.53	3.81	2.40	0.24
Prediction time	0.04	0.04	0.02	0.06

Table 1: Models Evaluation and Computational Efficiency Results
The table above shows the evaluation metrics used to evaluate the models' performances.

7.1 PERFORMANCE METRICS FOR CLASSIFICATION TASKS

After hyperparameter tuning, AdaBoost achieved high performance comparable to LightGBM and XGBoost, which indicates that the default parameters of the model performed poorly on the dataset. Upon investigation, it was observed that the learning rate parameter was too high, leading to poor generalization. Moreover, a ROC-AUC of 0.6146 shows that the model has moderate predictive power and has room for improvement.

Accuracy: LightGBM, optimized AdaBoost, and XGBoost all have an accuracy of 0.9762 which suggests that the models correctly predicted the outcome for 97.62% of the instances in the dataset. A high accuracy may not always suggest a good model.

Precision: The models show a precision of 0.9878 which shows that out of all the predictions made, 98.78% of them belong to the true class, with very few false positive errors.

Recall: The high value for recall suggests that the model correctly identified most of the actual positive cases in the dataset, highlighting the models' abilities to minimize false negatives.

F1-Score: The resulting F1-Score is relatively high, since the F1-score is the harmonic mean of recall and precision. A high F1-score is usually a sign of a robust and reliable model.

Receiver Operating Characteristic-Area Under the Curve (ROC-AUC): This measures the ability of a binary classifier to distinguish between the positive and negative classes. Both AdaBoost, XGBoost, and LightGBM have a high ROC-AUC score, which indicates that the models can effectively

distinguish between classes, demonstrating robustness in multi-class classification.

7.2 COMPUTATIONAL EFFICIENCY

Training Time: While LightGBM and AdaBoost trained quickly, XGBoost was the fastest to train with training time of 0.24 seconds, which makes it more suitable for real-time applications or frequent retraining.

Prediction time: For prediction time, all the models demonstrated high efficiency, this suggests that all three models are suitable for real-time prediction tasks.

7.3 CONFUSION MATRIX

The confusion matrix was used to visualize the classification performance of the models in greater detail. The visualization gives a better understanding of the predicted labels against the true labels. Each cell in the matrix represents the number of instances classified into a specific category.

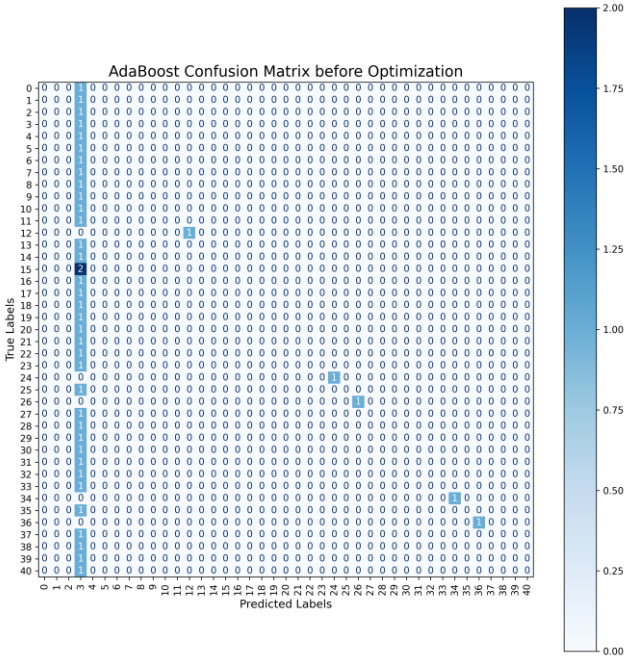


Figure 5: Confusion Matrix for AdaBoost before Hyperparameter Tuning.

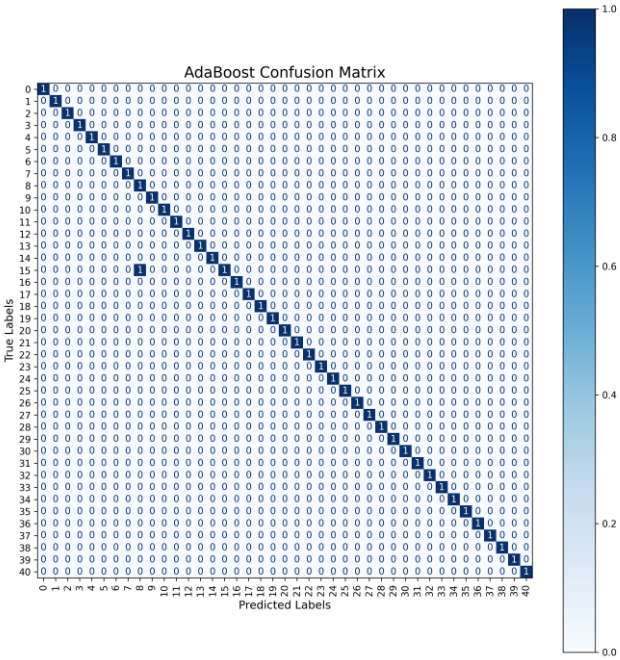


Figure 6: Confusion Matrix for AdaBoost after Hyperparameter Tuning.

8. CONCLUSION AND FUTURE WORK

The predictive medical diagnostic system using machine learning was successfully developed using a complete analytical pipeline.

The models performed optimally on the testing dataset with equal accuracy, precision, recall, and F1-score of 97.6%, 98.8%, 98.8%, and 98.4% respectively. XGBoost, although had a slightly less ROC-AUC score showed the fastest training time, making it ideal for scenarios requiring frequent retraining. These results demonstrate the potential of machine learning in the healthcare industry to improve diagnostic efficiency, thereby supporting medical decision-making.

The dataset used in this project was relatively small and balanced with symptoms evenly distributed, which may not mirror real-world healthcare data.

To further improve this project, it would be beneficial to test the models with a relatively larger dataset and perform validation with diverse and external datasets to test their robustness and efficiency.

Future work could involve incorporating the model into real-world applications to be used by healthcare professionals in real-time diagnostics.

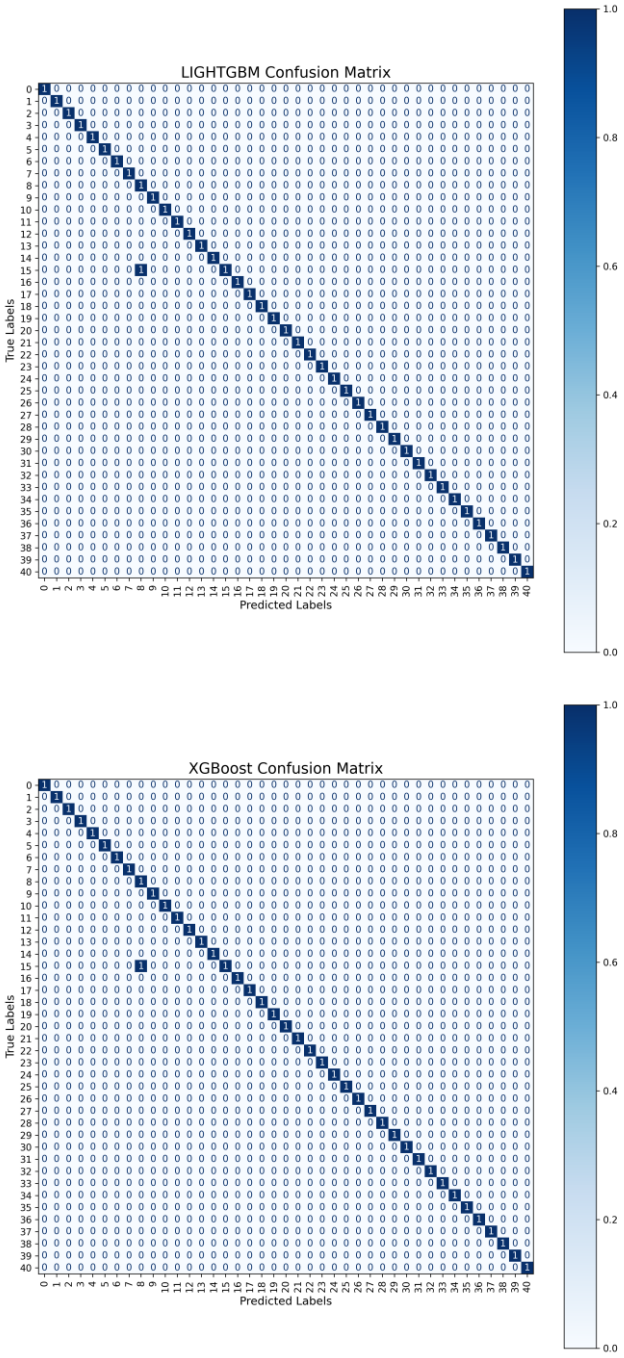


Figure 8: Confusion Matrix for XGBoost

Figure 5 shows off-diagonal values which indicates that there are visible misclassifications by the model with only 5 points along the diagonal, these points are where the model made correct classifications. Figures 6, 7, and 8 shows diagonal dominance which indicates that the model achieved near-perfect classification.

The confusion matrices for LightGBM, optimized Adaboost, and XGBoost shows strong diagonals which reveals that most of the predicted values align with the true values. However, only one disease was misclassified by each of the models possibly due to overlapping symptom patterns in the dataset.

REFERNCES

S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

Harvard Business School Online. (2021). *What is predictive analytics? Uses and examples*. Retrieved December 13, 2024, from https://online.hbs.edu/blog/post/predictive-analytics

Y. Freund, R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", 1995

Scikit-learn. (n.d.). *1.11. Ensemble methods: AdaBoost*. Retrieved December 13, 2024, from https://scikit-learn.org/stable/modules/ensemble.html#adaboost

LightGBM Developers. (n.d.). *LightGBM: A fast, distributed, high-performance gradient boosting (GBDT, GBRT, GBM or MART) framework*. Retrieved December 13, 2024, from https://lightgbm.readthedocs.io/en/stable/

Simplilearn. (n.d.). *What is XGBoost algorithm in machine learning?*. Retrieved December 13, 2024, from https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article#:~:text=XGBoost%20is%20a%20boosting%20algorithm,until%20the%20model%20stops%20improving.