Course Project Phase Two

Professor Corey

Junyu Li, Lanjun Wang

Apr 14th, 2025

Phase Two

For the binary classification task, we use the Breast Cancer Wisconsin dataset. The target variable in this dataset indicates whether a tumor is malignant (1) or benign (0). Based on the exploratory data analysis in Phase 1, we observed some overlap between classes and varying degrees of feature correlation. To address this, we selected four models for evaluation: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost.

We begin with Logistic Regression as our baseline model due to its simplicity, speed, and interpretability. Logistic Regression models the log-odds of the binary outcome and produces probabilistic predictions, which is especially useful in medical diagnostics. Although it assumes linear separability, it often performs well and serves as a solid benchmark. We plan to tune its regularization method (L1 loss and L2 loss) to control overfitting, as well as the regularization strength (C), which determines how heavily penalties are applied to model weights.

Next, we use the Support Vector Machine model, which is well-suited for high-dimensional data. It finds the optimal decision boundary by maximizing the margin between classes. Due to class overlap, we will start with a linear kernel, and consider switching to RBF if needed. We will tune the regularization parameter C to balance margin and error, and gamma if using a non-linear kernel to control decision boundary complexity and avoid overfitting.

We also include the Random Forest model. It is a robust ensemble method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It handles noisy and overlapping features well, which suits the nature of this dataset. We plan to tune the n\_estimators and max\_depth to control model complexity and maintain good generalization.

Finally, we include XGBoost, a gradient boosting method that builds trees sequentially. Compared to Random Forest, XGBoost focuses on correcting previous errors, which is helpful in datasets with overlapping and subtle patterns like ours. It also offers advanced regularization, making it less prone to overfitting. We will tune the learning rate, max\_depth, and regularization terms (reg\_alpha and reg\_lambda) to optimize performance.

For metrics, we will use recall to minimize false negatives, which is critical in medical diagnosis, and ROC AUC score to assess the model’s ability to distinguish between classes across all thresholds. These metrics are well-suited for our slightly imbalanced dataset and the high cost of misclassifying malignant cases.

For the multiclass classification task, we use the Data Science Job Salaries dataset. The target variable is experience\_level, which contains four classes: Entry-level (EN), Mid-level (MI), Senior-level (SE), and Executive-level (EX). Based on the exploratory data analysis in Phase 1, we observed that the dataset includes both categorical and numerical features, contains minimal correlation between features, and is somewhat imbalanced with the senior-level class being the majority. These observations guided our model selection.

We begin with the Decision Tree Classifier, which is a simple and interpretable model capable of handling both categorical and numerical data without the need for scaling. It recursively partitions the feature space to maximize class purity at each node. This method is flexible for datasets with mixed feature types and non-linear decision boundaries. However, due to its tendency to overfit, we will tune max\_depth to control tree complexity and min\_samples\_split to avoid overly specific splits on small data subsets.

Next, we use the K-Nearest Neighbors (KNN) algorithm, which classifies a data point based on the majority vote of its nearest neighbors. KNN is useful in capturing local patterns and works well when data points with similar attributes tend to belong to the same class, which may occur in experience classification based on roles and locations. We will apply feature scaling due to its sensitivity to distance metrics. We will tune the number of neighbors and the weighting method to balance performance and generalization.

We also include the Random Forest Classifier, an ensemble method that aggregates predictions from multiple decision trees. It reduces overfitting by combining diverse models and performs well with low signal-to-noise ratios, which is helpful given the limited feature correlations in this dataset. We plan to tune n\_estimators (number of trees) and max\_depth to balance bias and variance.

Finally, we include the XGBoost Classifier, a gradient boosting algorithm that builds trees sequentially to correct errors from previous iterations. XGBoost is robust to imbalanced data and excels at capturing subtle patterns, making it a strong candidate for our task. We will tune the learning\_rate to control model updates, max\_depth to manage complexity, and regularization terms reg\_alpha and reg\_lambda to reduce overfitting.

For evaluation, we will use the macro-averaged F1-score, which gives equal weight to each class regardless of frequency, ensuring fair evaluation across the imbalanced classes. We will also use accuracy to assess overall correctness. These metrics offer a comprehensive view of model performance on this multiclass task.