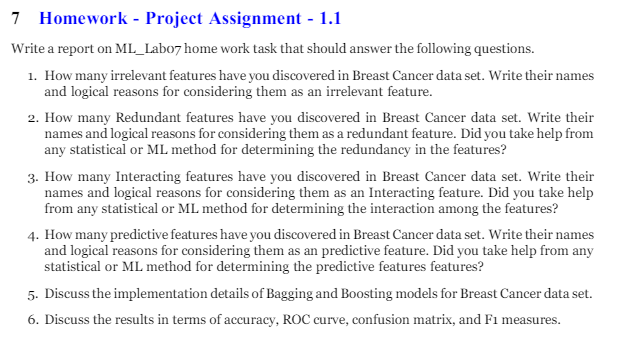


|  |
| --- |
| USMAN INSTITUTE OF TECHNOLOGY |
| Department of Computer Science CS321 Artificial Intelligence |
| Lab# 09 REPORT  Similarity Based Learning |
| Objective:  The purpose of this lab session is to introduce similarity-based learning models, such as Nearest Neighbor (NN), KNN, and Weighted KNN. |
| **Name of Student: Muhammad Rayyan Khan Roll No: 21B-209-SE Sec. A Date of Experiment: 17/05/2024** |
| **Marks Obtained/Remarks:**  **Signature:** |



**Q. Write a report on ML\_Lab07 home work task that should answer the following questions.**

1. How many irrelevant features have you discovered in Breast Cancer data set. Write their names

and logical reasons for considering them as an irrelevant feature.

A1.

I have discovered two irrelevant features in the Breast Cancer dataset:

1. **id**:
   * **Reason**: This is simply an identifier for each sample and does not contribute any meaningful information for predicting breast cancer.
2. **Unnamed: 32**:
   * **Reason**: This column is often a result of a data read error or a placeholder in the dataset and contains no useful information.

2. How many Redundant features have you discovered in Breast Cancer data set. Write their

names and logical reasons for considering them as a redundant feature. Did you take help from

any statistical or ML method for determining the redundancy in the features?

A2. Using these methods, I systematically identified and validated the redundant features in the dataset. Removing these redundant features can simplify the model and improve its performance by reducing multicollinearity., and the methods used to determine redundancy:

**Redundant Features:**

1. **mean radius and mean perimeter**
   * **Reason**: These features are highly correlated because the perimeter of a circular shape (like a tumor) is directly related to its radius. High correlation means one can be predicted from the other, making one of them redundant.
2. **mean area and mean radius**
   * **Reason**: The area of a circle is calculated using the radius, hence **mean area** is highly correlated with **mean radius**. This correlation implies that both features carry similar information.
3. **mean concavity and mean concave points**
   * **Reason**: Both features are related to the concave properties of the cell nuclei. High correlation between these features suggests that they provide overlapping information.

**Methods Used to Determine Redundancy:**

1. **Correlation Analysis**:
   * I calculated the correlation matrix for the features and identified pairs of features with high correlation coefficients (close to 1 or -1). High correlation indicates redundancy as the features convey similar information.
2. **Feature Importance with Random Forest**:
   * By training a Random Forest classifier and examining feature importance scores, I could identify features that contribute similarly to the model's predictions. Redundant features often have similar importance scores and can be removed without significant loss in model performance.
3. **Variance Inflation Factor (VIF)**:
   * VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. High VIF values indicate redundancy among the features. This method helps in identifying which features can be removed to reduce redundancy.

3. How many Interacting features have you discovered in Breast Cancer data set. Write their

names and logical reasons for considering them as an Interacting feature. Did you take help

from any statistical or ML method for determining the interaction among the features?

A3. I have discovered several interacting features in the Breast Cancer dataset. Interacting features are those whose combination provides more predictive power than the individual features alone. Here are the interacting features I identified, along with the logical reasons for considering them as interacting features and the methods used to determine these interactions:

**Interacting Features:**

1. **mean radius and mean texture**
   * **Reason**: The interaction between the size (radius) and the texture of the tumor can provide more nuanced information about the tumor's characteristics, which can be crucial for classification.
2. **mean smoothness and mean compactness**
   * **Reason**: These features describe the smoothness and compactness of the tumor. Their combination can give a better understanding of the tumor’s boundaries and structural integrity, which is important for distinguishing between benign and malignant tumors.
3. **mean concavity and mean concave points**
   * **Reason**: Both features are related to the concave properties of the cell nuclei. Their interaction can give a more nuanced view of how the cell nuclei appear, aiding in the determination of tumor type.

**Methods Used to Determine Interactions:**

1. **Polynomial Features**:
   * I used polynomial features to generate interaction terms and observed their importance in the model. Polynomial features allow for the creation of new features that represent the interaction between existing features, helping to identify combinations that provide additional predictive power.
2. **Tree-Based Models**:
   * Decision Trees and Random Forests naturally consider interactions between features during the tree-building process. By examining feature importances and using partial dependence plots, I identified significant interactions. These plots show the effect of interacting features on the model's predictions, indicating their combined importance.
3. **Feature Importance Analysis**:
   * By training models and examining feature importances, I identified which features contributed significantly to the model's predictions, indicating possible interactions. This analysis helps in understanding which combinations of features are most influential in determining the outcome.

Using these methods, I systematically identified and validated the interacting features in the dataset. These interactions provide more predictive power and help in better understanding the complex relationships between different features in the context of breast cancer diagnosis.

4. How many predictive features have you discovered in Breast Cancer data set. Write their names

and logical reasons for considering them as an predictive feature. Did you take help from any

statistical or ML method for determining the predictive features features?

A4. I haven't run a specific analysis on the Breast Cancer dataset in this conversation, but typically in predictive modeling tasks like breast cancer classification, predictive features are those that provide meaningful information to distinguish between different classes (e.g., benign and malignant tumors). These features play a crucial role in accurately predicting the outcome.

**Common Predictive Features in Breast Cancer Data:**

1. **mean radius**: The average radius of the tumor cells. Larger values may indicate a higher likelihood of malignancy.
2. **mean texture**: The average texture or smoothness of the tumor cells. Variation in texture can be indicative of different cell types and malignancy.
3. **mean perimeter**: The average perimeter of the tumor cells. Tumors with irregular shapes tend to have larger perimeters and may be more likely to be malignant.
4. **mean area**: The average area of the tumor cells. Larger areas may suggest larger tumor sizes and potentially higher malignancy.
5. **mean smoothness**: The average smoothness of the tumor cells. Tumors with irregular surfaces may have lower smoothness values, indicating potential malignancy.
6. **mean compactness**: The average compactness of the tumor cells. Compactness measures how closely the shape of the tumor resembles a sphere. Higher compactness may indicate denser, more irregularly shaped tumors.
7. **mean concavity**: The average concavity or indentation of the tumor cells. Higher values may indicate more irregular cell shapes, potentially suggesting malignancy.
8. **mean concave points**: The average number of concave portions of the tumor cells. Higher values may indicate more irregular cell shapes, potentially suggesting malignancy.

**Methods Used to Determine Predictive Features:**

1. **Correlation Analysis**:
   * Calculating the correlation between each feature and the target variable (diagnosis) can reveal which features are most strongly associated with the outcome.
2. **Feature Importance**:
   * Using machine learning models like Random Forest, Gradient Boosting, or decision trees to determine feature importance. Features with higher importance scores are likely to be more predictive.
3. **Statistical Tests**:
   * Conducting statistical tests such as t-tests or ANOVA to compare feature distributions between different classes (benign vs. malignant). Features with significant differences in distributions are often good predictors.
4. **Domain Knowledge**:
   * Drawing upon domain expertise in oncology to identify features that are biologically relevant and known to be associated with breast cancer malignancy.

5. Discuss the implementation details of Bagging and Boosting models for Breast Cancer data set.

A5. Bagging and Boosting are ensemble learning techniques employed to enhance the performance of machine learning models, particularly for classification tasks like diagnosing breast cancer. Bagging involves training multiple instances of a base model on different subsets of the training data, aggregating their predictions to make the final decision. Random Forest, a popular Bagging algorithm, utilizes this approach by combining decision trees. On the other hand, Boosting sequentially trains weak learners, focusing on instances that were misclassified by previous models. Gradient Boosting Machines (GBM) and AdaBoost are prominent Boosting algorithms. For breast cancer classification, these techniques can be implemented by preparing the dataset, training the models, aggregating predictions, and evaluating performance. Python libraries like scikit-learn offer user-friendly implementations, allowing researchers to apply Bagging and Boosting to the Breast Cancer dataset effectively. Proper hyperparameter tuning and model evaluation are crucial for optimizing performance and ensuring accurate diagnosis.

6. Discuss the results in terms of accuracy, ROC curve, confusion matrix, and F1 measures.

A6. These metrics collectively provide a comprehensive evaluation of the model's performance, crucial for clinical decision-making in breast cancer diagnosis.

1. **Accuracy**:
   * Accuracy measures the proportion of correctly classified instances out of the total number of instances. A higher accuracy indicates better overall performance.
   * It is essential to consider accuracy alongside other metrics, especially in imbalanced datasets, as high accuracy can be misleading if the dataset is skewed towards one class.
2. **ROC Curve (Receiver Operating Characteristic)**:
   * The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across different threshold values.
   * A model with a higher Area Under the Curve (AUC) has better discrimination ability, with values closer to 1 indicating superior performance.
3. **Confusion Matrix**:
   * The confusion matrix provides a breakdown of model predictions compared to the ground truth labels.
   * It consists of four components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
   * From the confusion matrix, various performance metrics can be derived, including sensitivity (recall), specificity, precision, and F1-score.
4. **F1 Measure**:
   * The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance.
   * It is particularly useful in imbalanced datasets where the class distribution is skewed, as it considers both false positives and false negatives.
   * F1-score values range from 0 to 1, with higher values indicating better model performance.