Diagnostic Systems Classifier

February 13, 2024

1 Rule-based Classifier

When constructing the rule-based classifier, we leveraged insights from the Eickhoff paper, which outlines key characteristics of malignant cells compared to benign ones. As stated in the assignment, it was up to us to define the appropriate variables and the notion of abnormality based on the supplied dataset. Therefore, translating the insights we gained from the Eickhoff paper, we decided to take the highest observed values of all the benign cell features as the thresholds for our rule-based classifier.

```
def fit_rule_based_classifier(self, df):
    features = df[df.malignant == 0].drop(columns=["id", "malignant"])

    max_values = features.max()

    thresholds = max_values.to_dict()

    # Define a function to classify based on thresholds
    def classify(row):
        for feature, threshold in thresholds.items():
            if row[feature] > threshold:
                return 1  # malignant
        return 0  # benign

    df["predicted_diagnosis"] = df.apply(classify, axis=1)
```

The model effectively distinguishes between malignant and benign cases, achieving an accuracy of 93.32%.

2 Random Forest Classifier

The implementation of the random forest classifier was done using the sklearn framework. The training was conducted using the 30 real-valued features, divided into a test/training split. The random forest algorithm learns to discern patterns that differentiate malignant from benign cases using the data from the training set, to then be able to predict malignant or benign cases in the test data.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

self.random_forest_classifier = RandomForestClassifier()

self.random_forest_classifier.fit(X_train, y_train)

y_pred = self.random_forest_classifier.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
```

The accuracy score from the test data yielded a result of 96.49%. While this method yields a quite high accuracy, its interpretability may be limited compared to the rule-based approach. However, it provides valuable insights into feature importance, highlighting which characteristics play a significant role in the classification, this is shown in Figure (1).

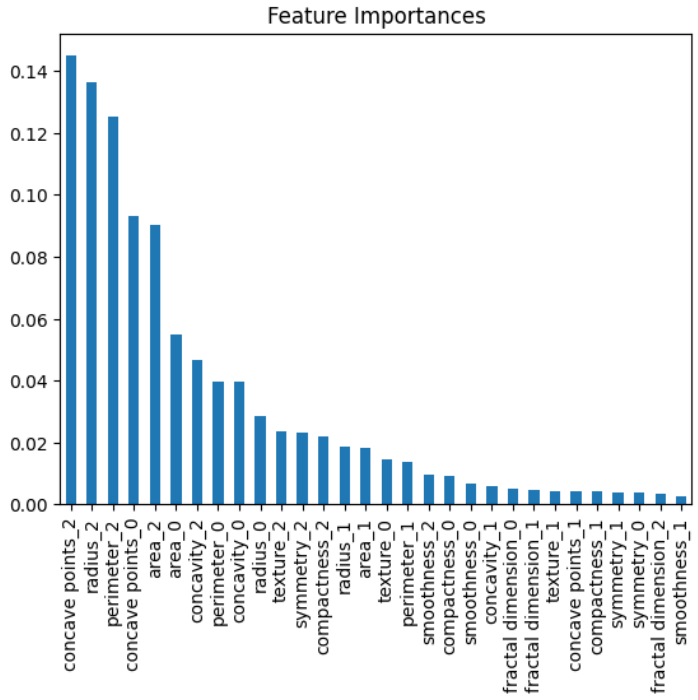

Figure 1: Feature importance from the random forest algorithm.

3 Custom-designed Classifier

As the task stated that the custom classifier should strike a balance between interpretability and classification performance, we decided to build upon the two existing models as they already demonstrated good accuracy.

To do this, we combined predictions of the rule-based and the random forest using a simple voting mechanism with an if-else statement. If the random forest prediction was malignant, we used it; otherwise, we took whatever the prediction from the rule-based model was. By doing so, the custom classifier seeks to capture the nuanced relationships between features while maintaining transparency in decision-making.

```
def custom_classifier(self, df, y_predicted, X_test):

    test_indices = X_test.index

    rule_base_predictions = df.loc[test_indices, 'predicted_diagnosis']

    random_forest_predictions = y_predicted

    predictions_combined = []

    for pred_rule_based, pred_rf in zip(rule_base_predictions,
        random_forest_predictions):
        if pred_rf == 1:
```

```
            predictions_combined.append(pred_rf)
        else:
            predictions_combined.append(pred_rule_based)

    return predictions_combined
```

Initially, we believed that leveraging the high accuracy of both models would complement each other. However, this model yielded the same accuracy as the random forest model, i.e., 96.49%.

4 Conclusions

Overall, by implementing and evaluating these classifiers, we gained valuable insights into their strengths and limitations. While the random forest model demonstrates high accuracy, its interpretability may be more limited compared to the rule-based approach. Nevertheless, it's important to note that this model still retains some interpretability, as valuable insights regarding feature importance can be extracted, as demonstrated above. On the other hand, the rule-based classifier may offer more transparency and interpretability, as it relies on intuitive rules rooted in medical knowledge.

The custom-designed classifier attempts to balance these aspects by bridging these two approaches, leveraging the strengths of each to enhance interpretability. We consider this statement to still be true, even though the model did not significantly outperform the individual models in terms of accuracy.