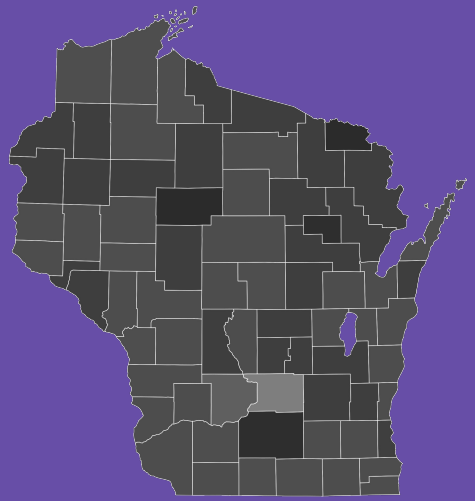


Wisconsin Diagnostic Breast Cancer Insights

Shareek Shaffie & Mohammad Nusairat

Overview

Develop a predictive model that can distinguish between benign and malignant tumors using machine learning techniques.

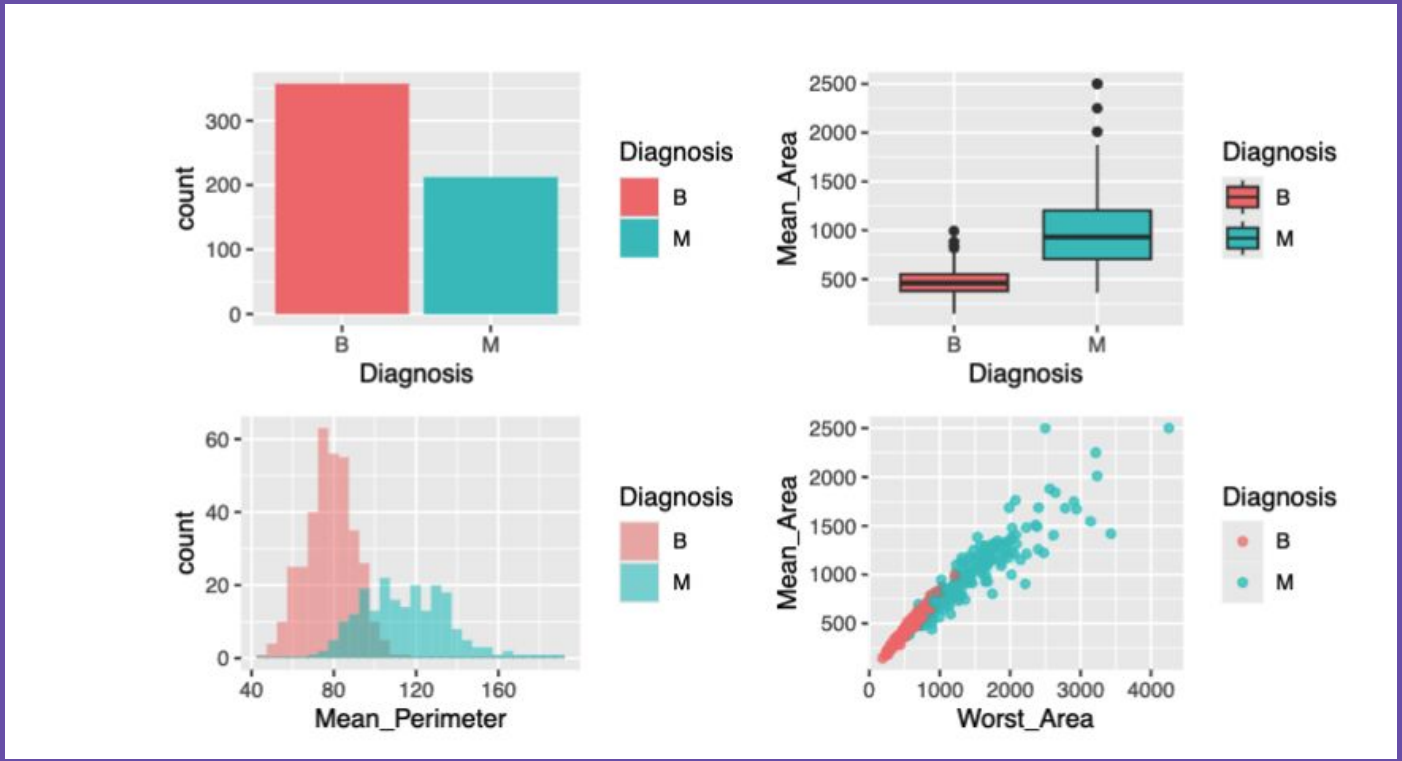
Identify the most significant cellular features that contribute to malignancy

Build a model with high accuracy, precision, and recall.

Steps

- **Clean**
  N/A Values, Factoring. (569 observations 32 attributes)
- **Analyze**
  Histograms, Barcharts, Boxplots.
- **Models**
  Logistic Regression, Decision Trees, Random Forest, KNN, SVM
- **Evaluate and Conclude**
  Best Model

How do we analyze the data?



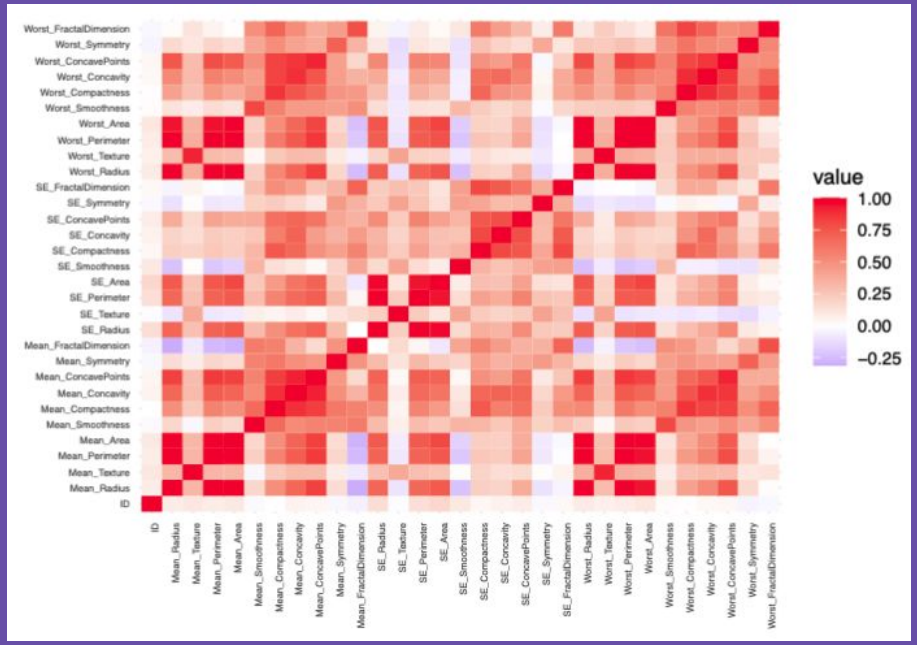


Table 1: Highly Correlated Features

Feature1	Feature2	Correlation
Mean_Perimeter	Mean_Radius	0.9978553
Mean_Area	Mean_Radius	0.9873572
Mean_ConcavePoints	Mean_Radius	0.8225285
Worst_Radius	Mean_Radius	0.9695390
Worst_Perimeter	Mean_Radius	0.9651365
Worst_Area	Mean_Radius	0.9410825

Logistic Regression

- Backward Stepwise Selection
- 7 VIF Checks

Accuracy = 0.97
Precision for Malignant = 0.96
Recall for Malignant = 0.98.

```
Call:
glm(formula = Diagnosis ~ Mean_Area + Mean_Smoothness + Mean_FractalDimension +
    SE_Radius + SE_Texture + SE_Concavity + SE_ConcavePoints +
    SE_FractalDimension + Worst_Texture + Worst_Smoothness +
    Worst_Compactness, family = binomial(link = "logit"), data = wdbc.data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.727e+01  1.021e+01  -4.631 3.63e-06 ***
Mean_Area              1.690e-02  3.793e-03   4.454 8.42e-06 ***
Mean_Smoothness        4.362e-01  5.437e+01   0.008 0.993599
Mean_FractalDimension  8.251e+01  1.307e+02   0.631 0.527875
SE_Radius              1.636e+01  4.600e+00   3.555 0.000378 ***
SE_Texture            -3.087e+00  1.114e+00  -2.770 0.005609 **
SE_Concavity           2.190e+01  1.504e+01   1.456 0.145458
SE_ConcavePoints       2.076e+02  1.340e+02   1.549 0.121429
SE_FractalDimension   -1.208e+03  4.091e+02  -2.952 0.003158 **
Worst_Texture          5.316e-01  1.128e-01   4.713 2.44e-06 ***
Worst_Smoothness       9.442e+01  2.978e+01   3.170 0.001524 **
Worst_Compactness      1.113e+01  4.724e+00   2.356 0.018479 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  77.142  on 557  degrees of freedom
AIC: 101.14

Number of Fisher Scoring iterations: 10

         Actual
Predicted Benign Malignant
        B    352         9
        M      5       203
```

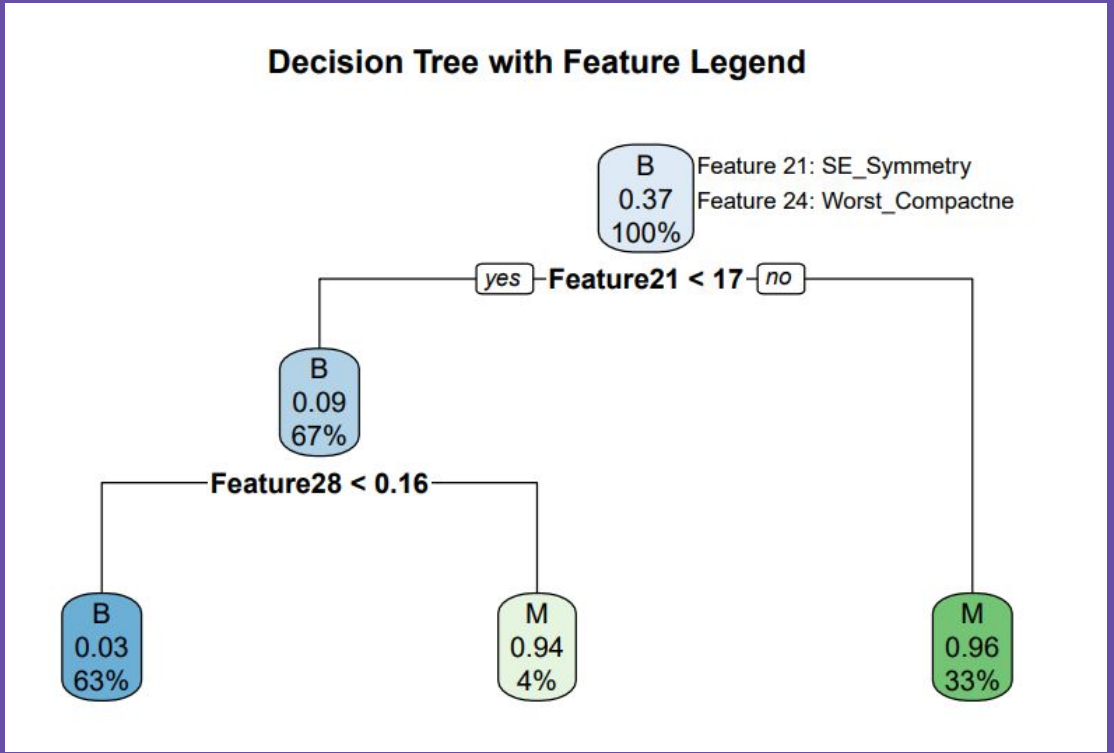
Data Splitting and Model Selection

- Standard 70/30 split
- 1) Decision Tree Model
- 2) Random Forest Model
- 3) K-Nearest Neighbors (KNN) Model with Optimized k using Cross-Validation
- 4) Support Vector Machine with Best Kernel Identification and Hyperparameter Fine Tuning

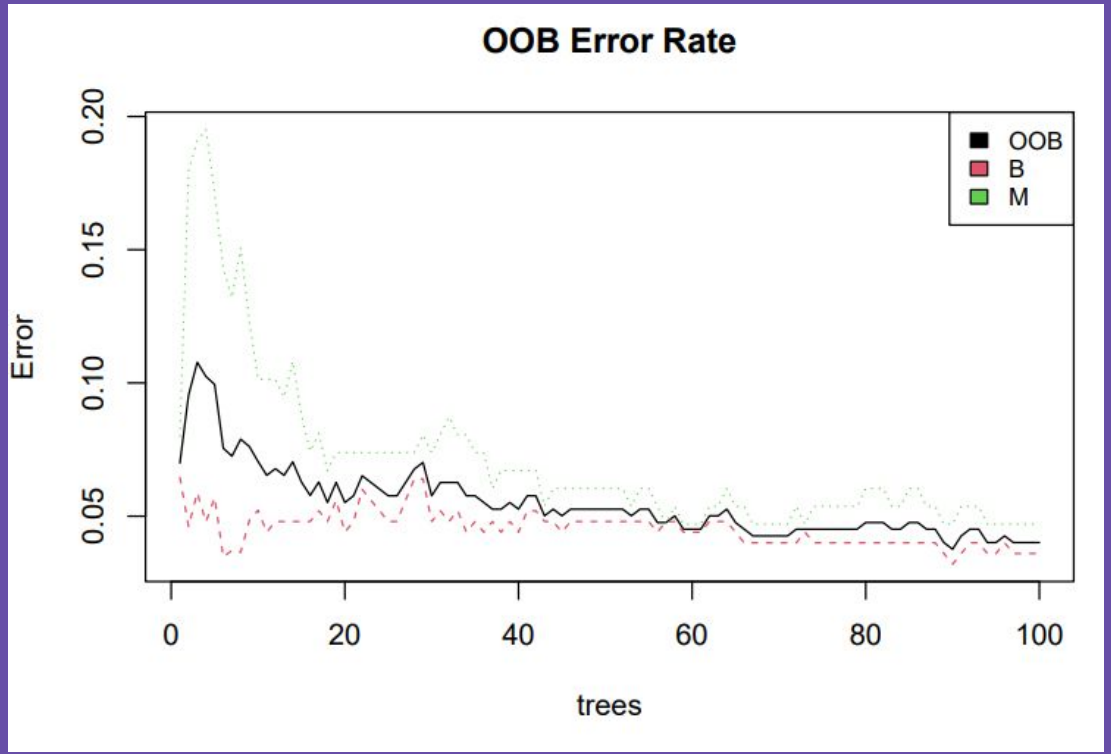
Based off of our initial exploratory data analysis and feature relationship introduction, we decided to select these models because of the complexity of the relationship between the features in the data.

Decision Tree

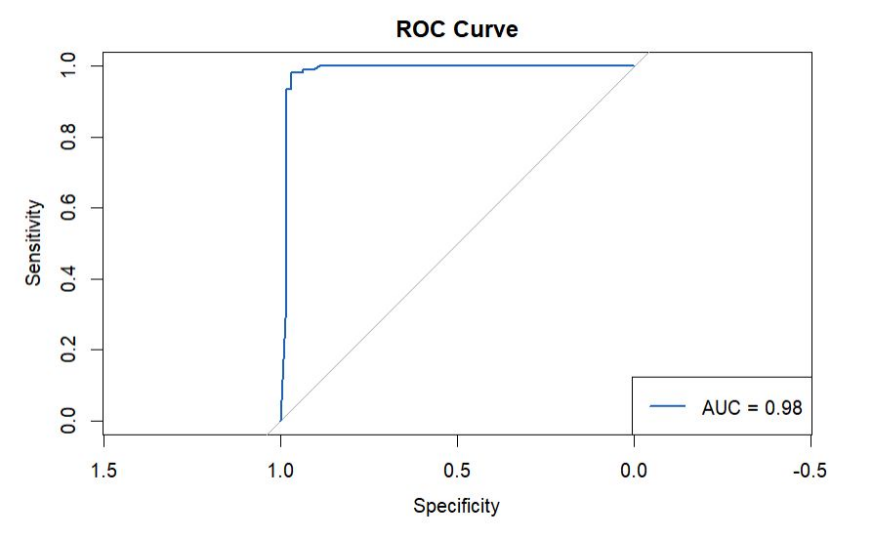
The decision tree is simple, easy to read, and achieves a strong classification performance with high accuracy, sensitivity, and specificity. Misclassifications are balanced between False Positives and False Negatives, with 7 instances each, which indicates good performance in handling both classes.

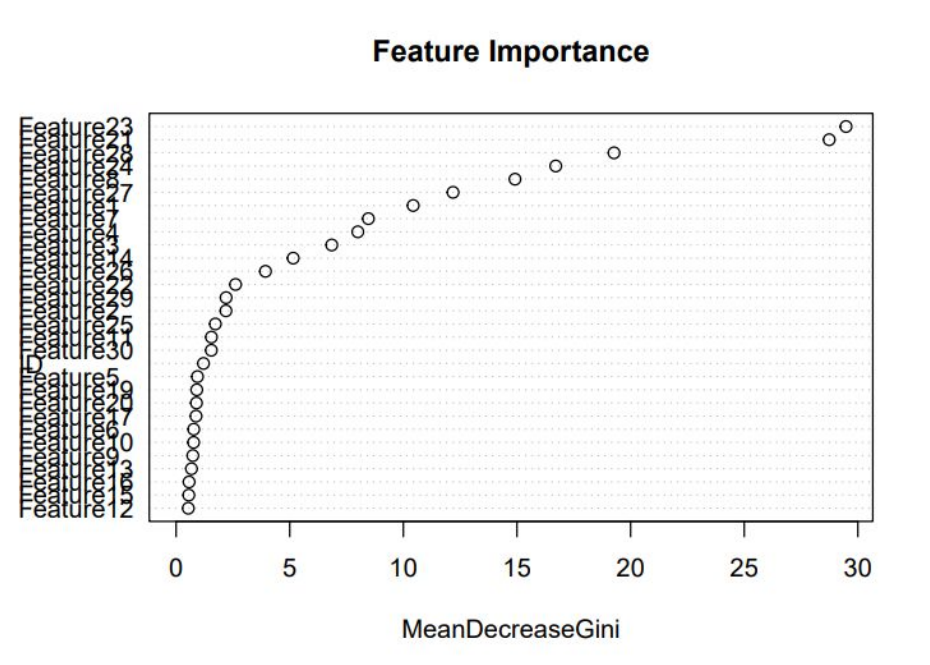


Random Forest



Random Forest



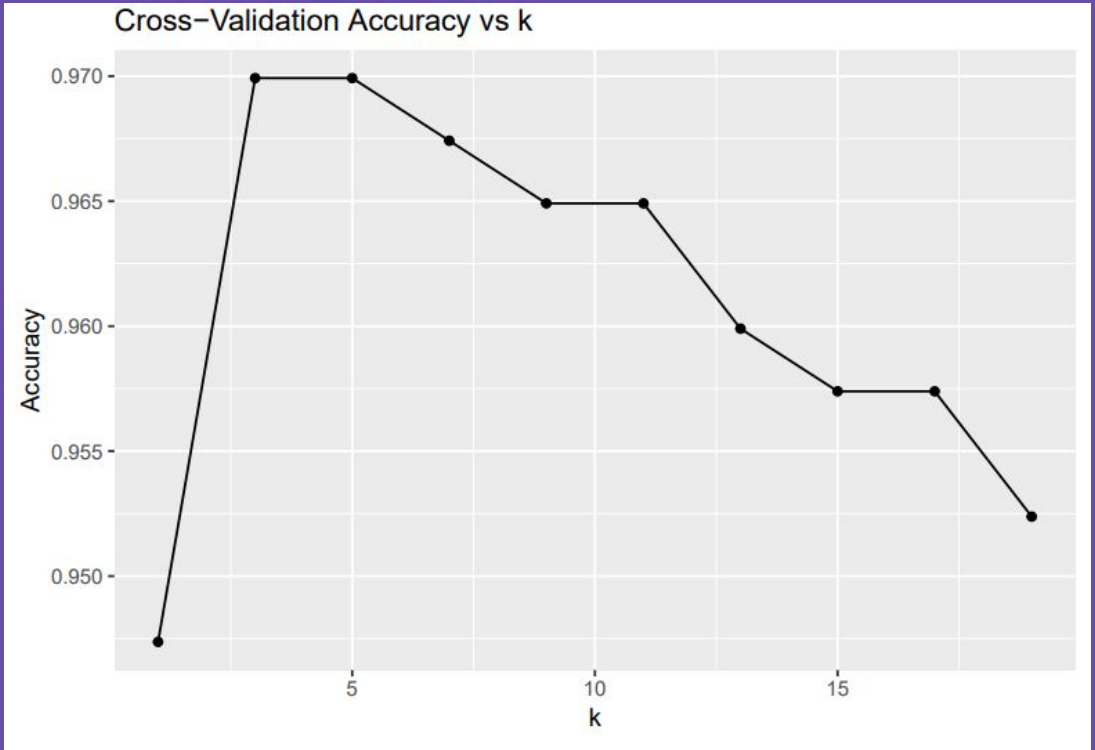


Random Forest

The Random Forest model shows strong classification performance, with a high AUC (0.98) and low OOB error rates. Increasing the number of trees beyond 40 does not significantly improve the model, suggesting that 40-50 trees may be sufficient for this classification problem. Another point to take note of is that the model performs slightly better at classifying the benign class compared to the malignant class, which is common in imbalanced datasets because of the difference in amount of data classified for each label.

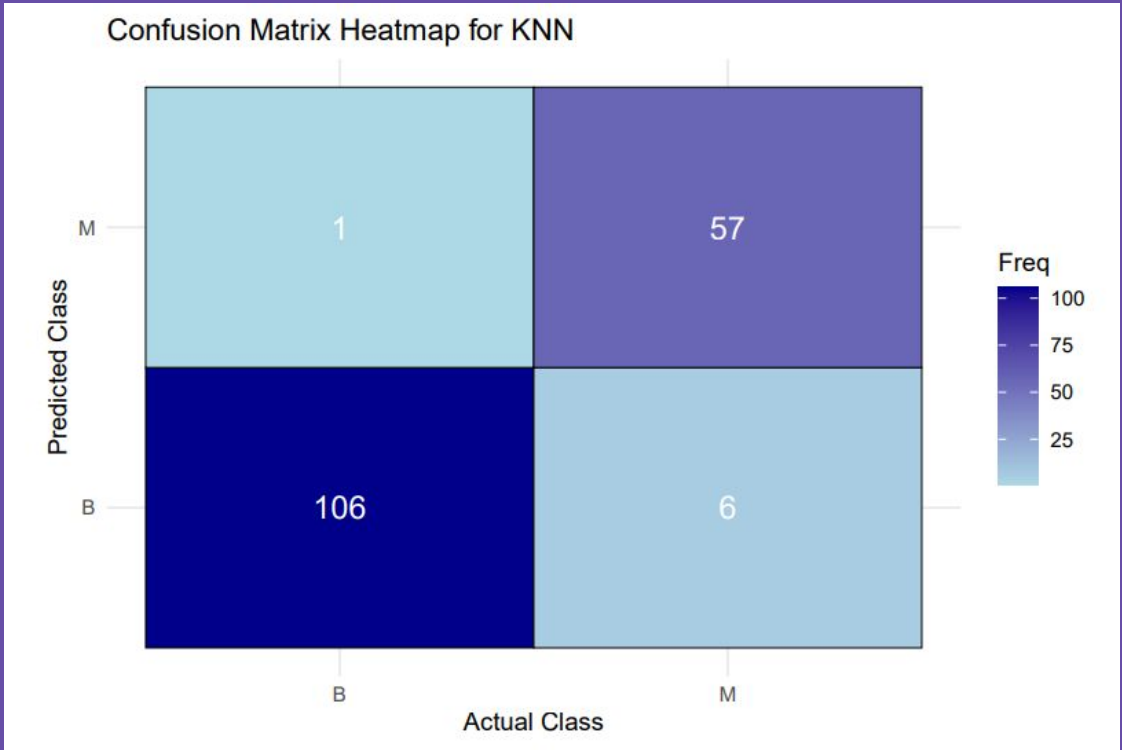
As for feature importance, see here that the Random Forest model relies heavily on a few key features (**SE_Symmetry**, **Worst_Radius**, **Worst_Texture**) for classification. This suggests that these are the most significant variables. The model also exhibits diminishing returns in predictive power for less important features, which indicates that feature selection or dimensionality reduction can further optimize the model.

KNN



Based on the visualization of the accuracy of k values 0 - 20, we can conclude that the optimal value for k is 5. Now, we implement the KNN Model using this optimized k value.

KNN

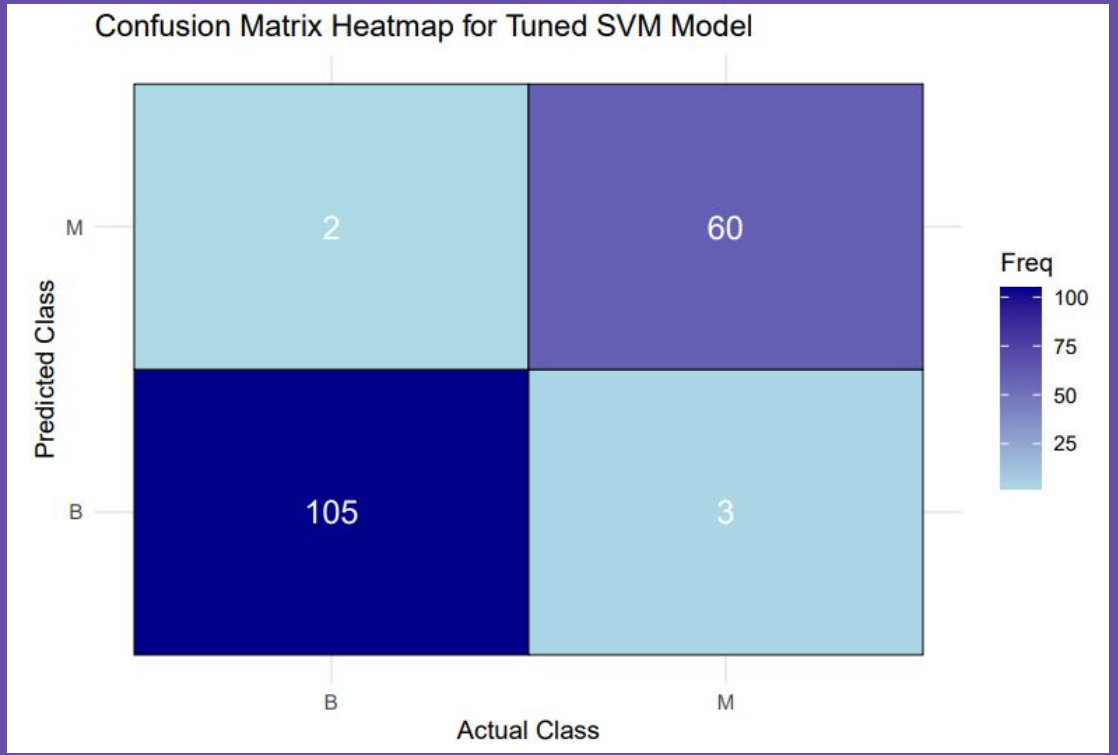


The model performs well overall, with a high number of correct classifications for both benign and malignant cases (only 1 benign case was misclassified as malignant). It struggles slightly with sensitivity, as 6 malignant cases were misclassified as benign. In summary, The model has strong precision for both classes, but a slight imbalance in sensitivity for malignant cases.

Support Vector Machine

with Best Kernel Identification and Hyperparameter Fine Tuning (only on best kernel for optimized time complexity)

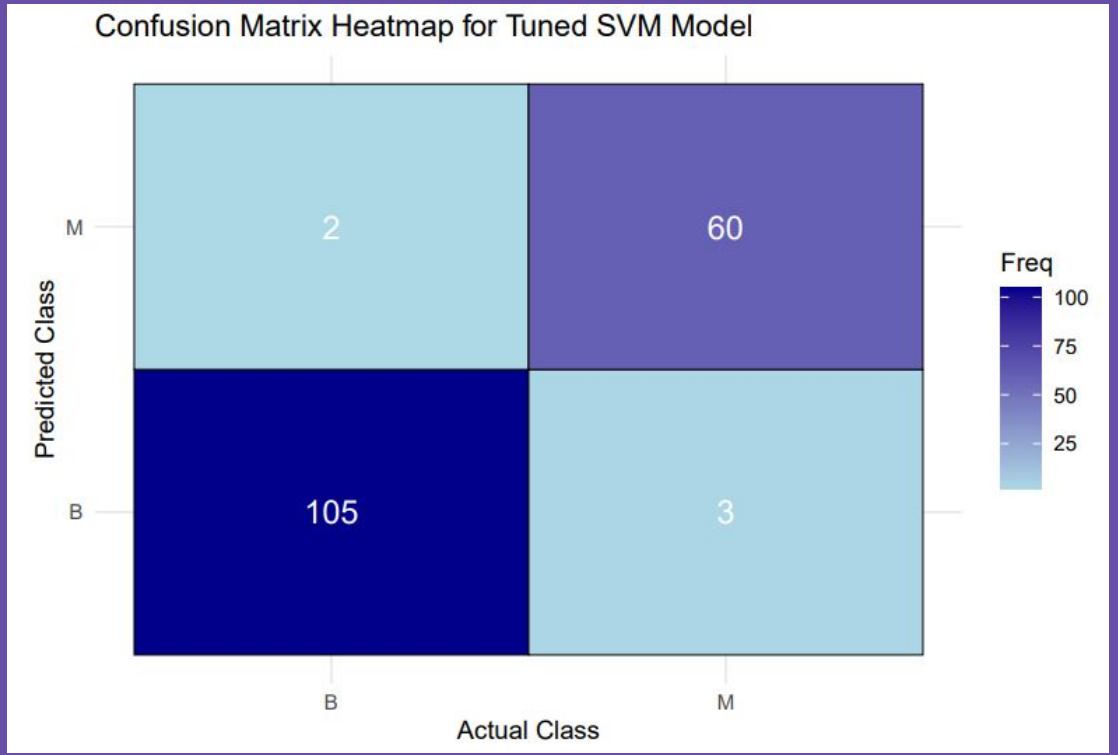
Best Kernel: linear
Best Parameters for the linear Kernel: cost=0.5 and gamma=0.25
Tuned Model Accuracy: 0.9705882



Support Vector Machine

with Best Kernel Identification and Hyperparameter Fine Tuning (only on best kernel for optimized time complexity)

The SVM model has high true positives (60) and true negatives (105). The number of false positives (2) is low, indicating the model rarely misclassifies benign cases as malignant. In terms of model weakness, it misclassifies 3 malignant cases as benign (false negatives), but overall, the model is accurate with strong true positive and true negative rates.



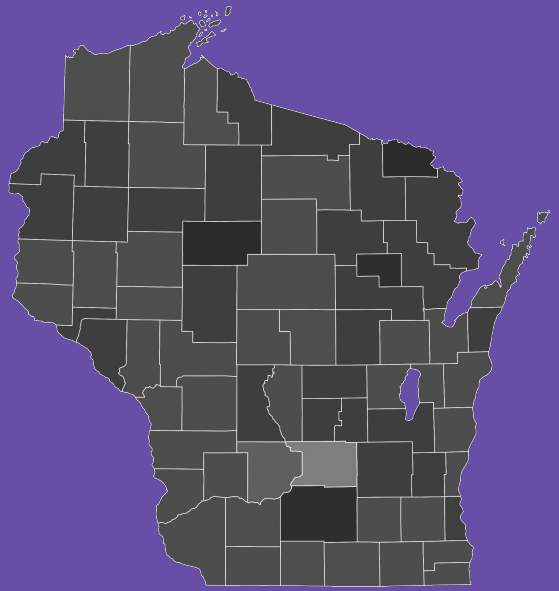
Model Comparison and Conclusion

Decision Tree Accuracy: 0.9176471
Random Forest Accuracy: 0.9705882
KNN Accuracy: 0.9588235
Tuned SVM Linear Kernel Accuracy: 0.9705882

In conclusion, both Random Forest and SVM emerged as the best-performing models, achieving the highest accuracy and balanced classification performance. While Random Forest is ideal for feature importance analysis and ensemble robustness, SVM is a strong contender for clean, linear separable datasets.
Decision Tree and KNN are valuable for interpretability and simplicity, respectively, but they fall short in terms of overall accuracy and sensitivity. Future analysis could be fruitful if we explore deeper feature engineering, balancing techniques, and hybrid models to enhance classification performance even further.

THE BEST MODELS

RF & SVM WITH ~97% ACCURACY



Questions?

Comments?