University of
South Australia

Kavan Mayankkumar Patel

ID: 110386407

Predictive Analytics

Assignment 3

Table of Contents

Introduction and Recap

Overview and Objective:

Assignment 1 laid the groundwork by delving into the complexities of healthcare fraud, a critical issue causing financial strains in the healthcare sector with losses in the billions. This foundational work set the stage for Assignment 2, which aimed to address this pressing issue through the development of a predictive decision tree model. Assignment 2's primary objective was to leverage the insights gained from the initial exploration and feature engineering efforts of Assignment 1, focusing on building a robust model capable of accurately predicting instances of healthcare fraud, thereby aiding in the reduction of its significant financial and societal impacts.

Methodology and Data Analysis:

In Assignment 2, a detailed evaluation of data led to the identification of pivotal features such as Provider IDs, PotentialFraud markers, and various metrics related to claims and reimbursements. These features were critically analyzed to discern patterns indicative of fraudulent activities. The dataset was meticulously partitioned into training (80%) and testing (20%) sets to create an optimal balance for model training. This was followed by the construction of four distinct decision tree models, each varying in depth, complexity, and splitting criteria. These models were then empirically tested to determine their efficiency in fraud detection, with visual aids like histograms and boxplots used to interpret the distribution of data points and identify outliers or anomalies.

Outcomes and Conclusion:

The outcomes of these models highlighted the nuanced relationship between model complexity and accuracy. Models with moderate complexity (Models 2 and 3) proved to be most effective, demonstrating the critical role of fine-tuning in achieving optimal performance. The assignment concluded with an emphasis on the importance of machine learning in fraud detection within the healthcare sector. It underscored the need for continuous evolution and adaptation of defence mechanisms against sophisticated fraud threats, acknowledging that a well-calibrated model could be crucial for early detection and prevention of fraud, thus safeguarding the integrity and efficiency of the healthcare system.

Data exploration and Feature Selection

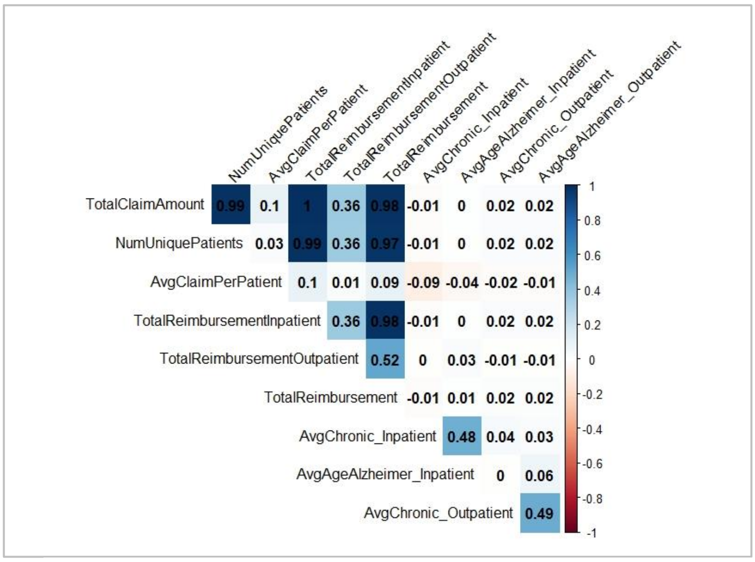


Fig 1: Heatmap

The heatmap I presented in Assignment 2 was a crucial tool in understanding the relationships between different numerical features in the healthcare fraud dataset which is also shown in figure 1 above. This visualization technique uses colour gradients to represent the correlation coefficients between pairs of variables. In our heatmap, the colour spectrum ranged from red (indicating negative correlations) to blue (indicating positive correlations). Such a visual representation was instrumental in quickly identifying patterns and relationships that might not be immediately apparent from a tabular dataset."

Insights Gained from the Heatmap:

"The heatmap revealed several noteworthy correlations. For instance, a strong positive correlation of 0.988 between TotalClaimAmount and NumUniquePatients indicated a direct and almost linear relationship between these variables. However, the perfect correlation of 1.000 between TotalClaimAmount and TotalReimbursementInpatient suggested redundancy, indicating these variables essentially convey the same information. Additionally, moderate correlations, such as between AvgChronic_Inpatient and AvgAgeAlzheimer_Inpatient (0.48), provided insights into the interplay between chronic conditions and age-related ailments in inpatient settings. In contrast, weak correlations near zero highlighted pairs of variables with negligible relationships, underlining the diversity in the dataset. These insights are particularly useful in fraud detection, as deviations from these patterns could signal fraudulent activities."

Alternative Representations Considered:

"While the heatmap was effective in illustrating correlations, I also considered other methods of representation. Scatter plots, for instance, could have offered a more granular view of the relationships between two variables at a time, potentially highlighting outliers, or non-linear relationships. Another approach could have been the use of Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, which would help visualize the most significant relationships in a two or three-dimensional scatter plot. However, the heatmap was chosen for its ability to simultaneously display multiple correlations in an easily interpretable format, making it highly suitable for our broad initial analysis."

Feature Selection:

In both assignments, the feature selection process was driven by the goal of effectively detecting healthcare fraud. Features were chosen based on their relevance to this objective, statistical significance, and data completeness. For example, 'TotalClaimAmount', 'NumUniquePatients', and 'AvgClaimPerPatient' were included due to their direct relation to healthcare claims. The 'Provider' and 'PotentialFraud' variables were essential for identifying fraudulent activities. The selection process involved a careful examination of each feature's potential to contribute meaningful insights into fraudulent behaviors.

Outlier detection was a critical step in preparing the data, especially given the nature of healthcare fraud, where anomalies can be indicative of fraudulent activities. Techniques like IQR (Interquartile Range) and Z-scores were employed to identify and assess outliers. For instance, extreme values in 'TotalClaimAmount' might indicate irregular billing practices. In some cases, outliers were retained because they represented genuine fraud cases, while in others, they were adjusted or removed to prevent skewing the analysis.

Variable scaling was performed to standardize the data, particularly important for models sensitive to the scale of data, such as distance-based algorithms. Features like 'TotalClaimAmount' and 'TotalReimbursement' were on different scales compared to other variables like 'AvgChronic_Inpatient'. We used standardization (Z-score normalization) to bring all variables to a common scale, enhancing the model's ability to learn from the data effectively. To refine the feature selection further, algorithmic methods were applied. Principal Component Analysis (PCA) was considered for dimensionality reduction.

Building Classification Models

In the realm of machine learning, classification models are vital tools for categorizing data, especially in sectors like healthcare, finance, and marketing. These models learn from labelled training data to predict or classify unseen data, a process crucial for tasks such as detecting healthcare fraud. Data partitioning, essential for model training and evaluation, will be executed with a 70:30 training-to-testing ratio in our study. Utilizing the robust capabilities of R for data manipulation and analysis, we will explore various classification algorithms, each with its unique strengths and assumptions. This approach not only aims to build effective models but also to deepen our understanding of the underlying patterns in healthcare data, particularly focusing on fraud detection.

1. KNN Model:

In the realm of healthcare fraud detection, the K-Nearest Neighbours (KNN) algorithm plays a crucial role by evaluating the similarity between new data points and existing cases in the training dataset. This non-parametric method classifies each instance based on the features that characterize fraudulent activities, such as claim amounts, patient metrics, etc. It calculates the distance between a new case's features and those of known cases, typically using the Euclidean distance metric. The simplicity and ease of implementation are notable advantages of the KNN algorithm. However, its effectiveness significantly hinges on the choice of k, the number of neighbours considered, and the distance metric used. The behaviour of accuracy for different of k can be seen in the figure 2 below.

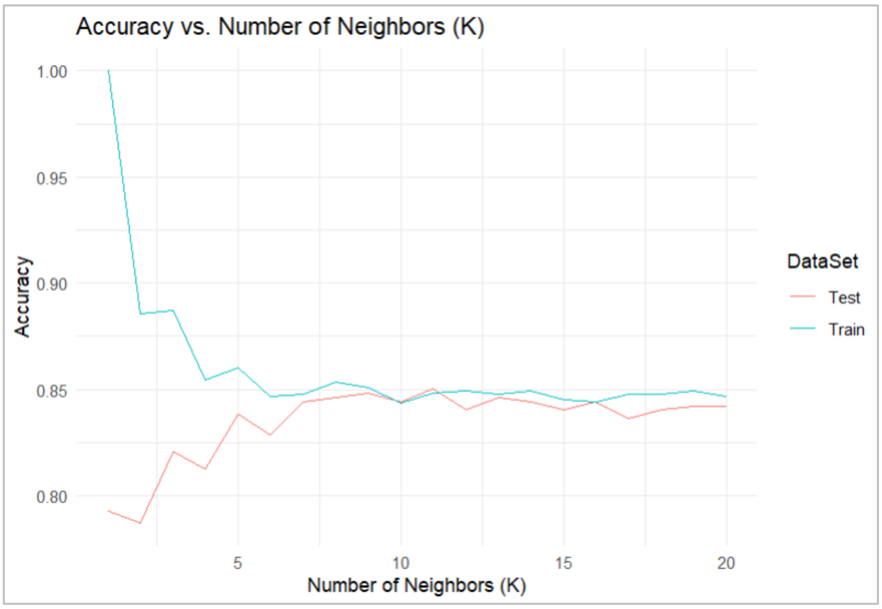
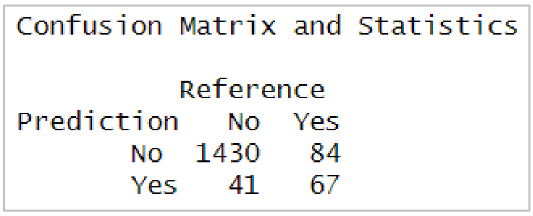


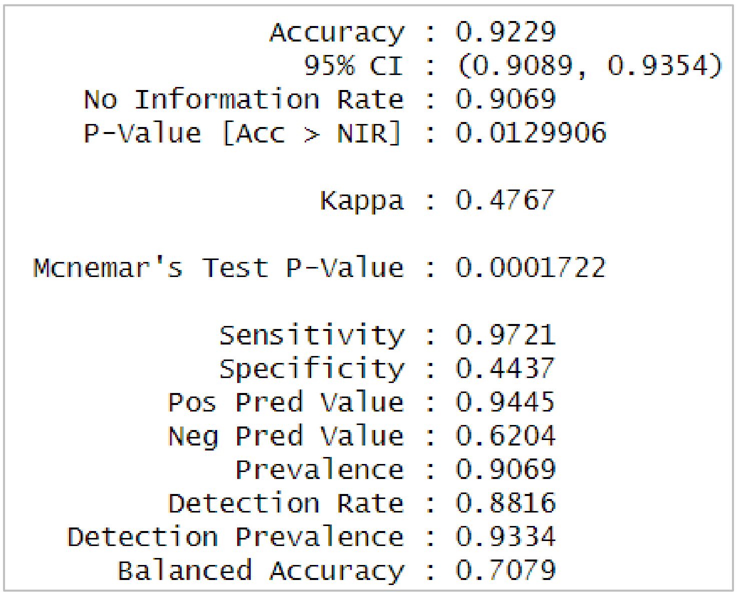
Fig 2: Accuracy vs k

The performance of this model, as assessed on a 70:30 training-to-testing data split, yielded an accuracy of approximately 92.29%, with a sensitivity of 97.21% and a specificity of 44.37%. These results were obtained with the default Euclidean distance metric used by R's ‘knn’ function. Notably, the model demonstrated a high positive predictive value (Pos Pred Value) of 94.45%, alongside a negative predictive value (Neg Pred Value) of 62.04%. The Kappa statistic stood at 0.4767, indicating moderate agreement between the predicted and actual labels.

Accuracy	**0.9229**
Precision	**0.444**
Recall	**0.620**
Kappa	**0.415**
F1	**0.519**
95% CI	**(0.9089, 0.9354)**



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1430   84
       Yes   41   67
```



```
               Accuracy : 0.9229
                 95% CI : (0.9089, 0.9354)
    No Information Rate : 0.9069
    P-Value [Acc > NIR] : 0.0129906

                  Kappa : 0.4767

 Mcnemar's Test P-Value : 0.0001722

            Sensitivity : 0.9721
            Specificity : 0.4437
         Pos Pred Value : 0.9445
         Neg Pred Value : 0.6204
             Prevalence : 0.9069
         Detection Rate : 0.8816
   Detection Prevalence : 0.9334
      Balanced Accuracy : 0.7079
```

2. SVM model:

The Support Vector Machine (SVM) stands as a cornerstone in machine learning for its adept handling of high-dimensional data, making it exceptionally well-suited for binary classification challenges such as healthcare fraud detection. By constructing an optimal hyperplane, SVM efficiently segregates data into two distinct categories, in this case, fraudulent and non-fraudulent healthcare claims. The strength of SVM lies in its kernel functions, which empower it to unravel non-linear relationships within the data. In our analysis, we evaluated the performance of various kernel types and focused on refining the radial basis function (RBF) kernel by tuning its gamma parameter.

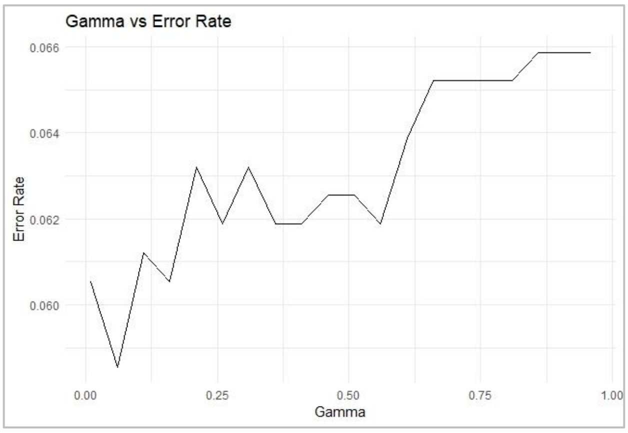
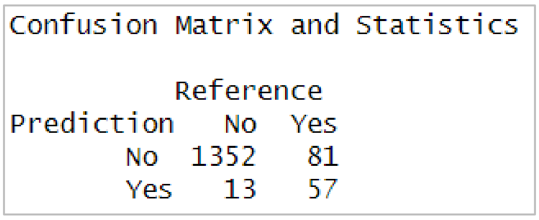


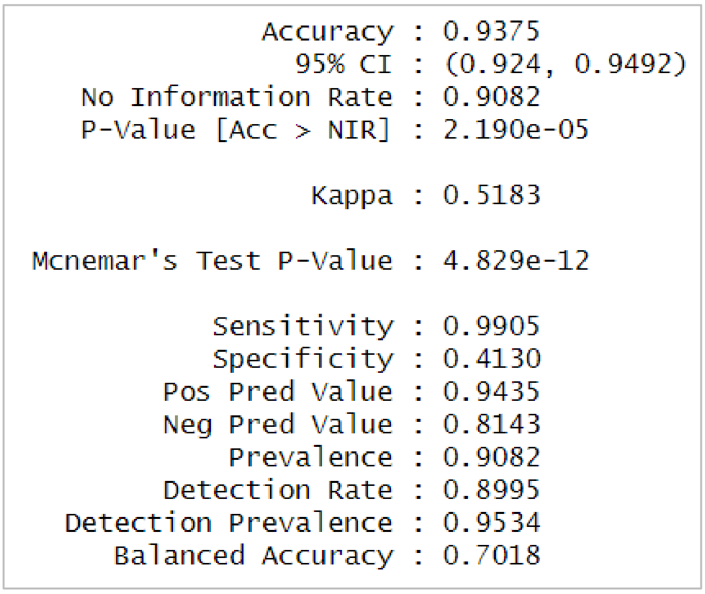
Fig 3: Gamma vs Error rate

Referring to Figure 3, which illustrates the performance of the SVM with the RBF kernel, we observed that the error rate systematically varied with gamma. Through this analysis, we identified that a gamma value around 0.06 yielded the lowest error rate, indicative of the model's heightened performance. With our SVM model, we achieved an accuracy of 93.75%, complemented by a Kappa statistic of 0.5183, suggesting a moderate agreement beyond chance. This level of accuracy, coupled with the model's sensitivity of 99.05% and a specificity of 41.30%, underscores its capability to reliably flag fraudulent cases. However, the relatively lower specificity indicates room for improvement, perhaps by exploring other kernel settings or incorporating additional features.

Accuracy	**0.9375**
Precision	**0.413**
Recall	**0.814**
Kappa	**0.5183**
F1	**0.549**
95% CI	**(0.924, 0.9492)**



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1352   81
       Yes   13   57
```



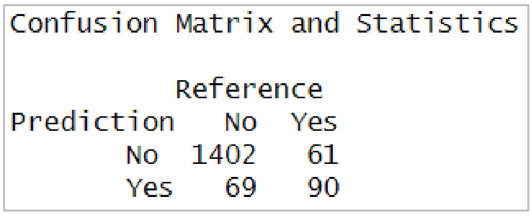
```
               Accuracy : 0.9375
                 95% CI : (0.924, 0.9492)
    No Information Rate : 0.9082
    P-Value [Acc > NIR] : 2.190e-05

                  Kappa : 0.5183

 Mcnemar's Test P-Value : 4.829e-12

            Sensitivity : 0.9905
            Specificity : 0.4130
         Pos Pred Value : 0.9435
         Neg Pred Value : 0.8143
             Prevalence : 0.9082
         Detection Rate : 0.8995
   Detection Prevalence : 0.9534
      Balanced Accuracy : 0.7018
```

As we continue to refine our model, the precision of approximately 41.3%, recall of 81.4%, and an F1 score of about 54.9% provide a balanced view of the model's current predictive power. These metrics, along with our graphed analysis, serve as a guidepost in our ongoing efforts to enhance the SVM's ability to combat healthcare fraud, ensuring that resources are allocated efficiently and ethically within the healthcare system.

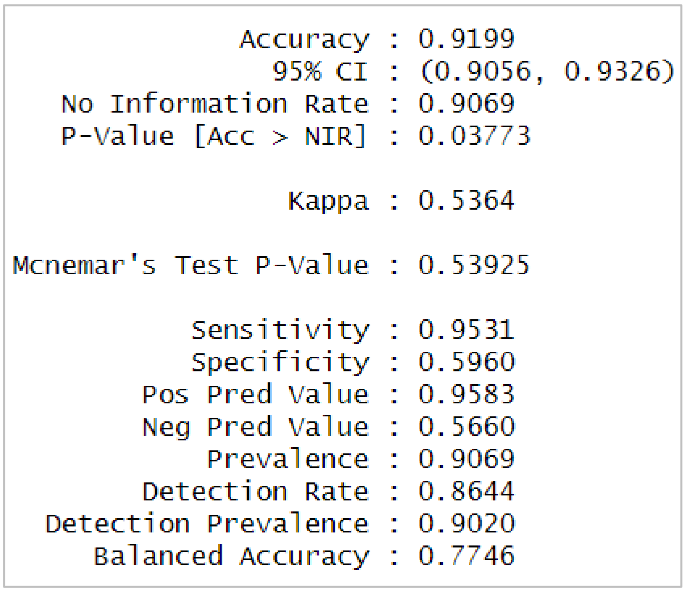
3. Naïve Bayes:

In the intricate domain of healthcare fraud detection, the Naïve Bayes classifier operates under the principles of Bayes' theorem, which evaluates the likelihood of an event based on prior knowledge of conditions that might be related to the event. This probabilistic model presumes that the features contributing to healthcare fraud are independent of one another, an assumption that simplifies the computation but can sometimes diverge from the complex interdependencies found in real-world data. Despite its simplifying assumptions, Naïve Bayes has demonstrated commendable performance in our dataset.

Accuracy	0.9199
Precision	**0.596**
Recall	**0.566**
Kappa	**0.5364**
F1	**0.581**
95% CI	**(0.9056, 0.9326)**



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1402   61
       Yes   69   90
```



```
               Accuracy : 0.9199
                 95% CI : (0.9056, 0.9326)
    No Information Rate : 0.9069
    P-Value [Acc > NIR] : 0.03773

                  Kappa : 0.5364

 Mcnemar's Test P-Value : 0.53925

            Sensitivity : 0.9531
            Specificity : 0.5960
         Pos Pred Value : 0.9583
         Neg Pred Value : 0.5660
             Prevalence : 0.9069
         Detection Rate : 0.8644
   Detection Prevalence : 0.9020
      Balanced Accuracy : 0.7746
```

The default parameters from the e1071 library were employed, without additional tuning, to maintain model simplicity and interpretability. According to Table above, the model achieved an accuracy of approximately 91.99%, with a Kappa value of 0.5364, indicating moderate agreement beyond chance. The model's specificity, at nearly 59.60%, and a sensitivity of 95.31% reflect its ability to identify fraudulent claims effectively while maintaining a balance between true positive and true negative rates. The positive predictive value of 95.83% and a balanced accuracy of 77.46% further illustrate the model's capability in this setting, proving that even with its foundational assumptions, Naïve Bayes remains a robust and valuable tool in the arsenal against healthcare fraud.

4. Decision tree

In our decision tree analysis for detecting healthcare fraud, we constructed a model with a controlled complexity, capping the tree's depth at four levels, as illustrated in Figure 4. This decision ensures a balance between model accuracy and interpretability. The tree's structure reveals a hierarchy of decisions, starting from the root and branching out to encompass various features at each subsequent level. By limiting the depth to four, the model avoids overfitting, resulting in a more generalized and robust classifier.

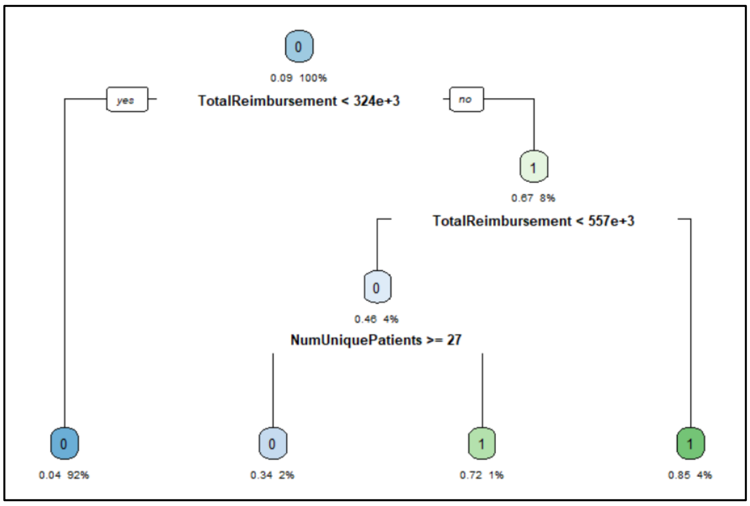
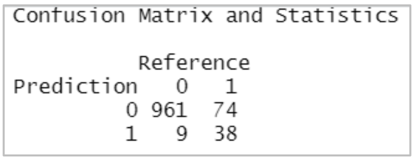


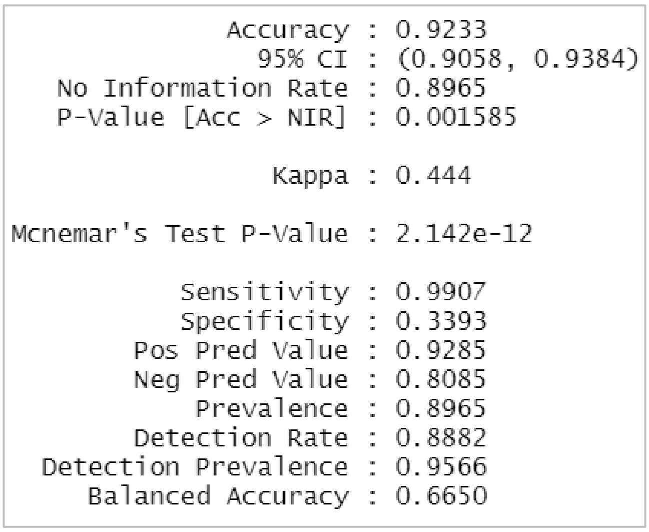
Fig 4: Decision Tree

The construction parameters for the tree maximum depth of 4, a minimum split threshold of 20, and a complexity parameter (cp) of 0.01 were pivotal in shaping its performance and structure. These parameters helped in focusing on the most significant features without overcomplicating the model, as depicted in Figure 8. Each node in the tree represents a critical decision point, with the leaf nodes providing the final classification.

Accuracy	**0.9233**
Precision	**0.3393**
Recall	**0.8085**
Kappa	**0.444**
F1	**0.5023**
95% CI	**(0.9058, 0.9384)**



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 961  74
         1   9  38
```



```
               Accuracy : 0.9233
                 95% CI : (0.9058, 0.9384)
    No Information Rate : 0.8965
    P-Value [Acc > NIR] : 0.001585

                  Kappa : 0.444

 Mcnemar's Test P-Value : 2.142e-12

            Sensitivity : 0.9907
            Specificity : 0.3393
         Pos Pred Value : 0.9285
         Neg Pred Value : 0.8085
             Prevalence : 0.8965
         Detection Rate : 0.8882
   Detection Prevalence : 0.9566
      Balanced Accuracy : 0.6650
```

Key performance metrics of this decision tree model, as established through our analysis, are as follows: an accuracy of 92.33%, precision of approximately 33.93%, and a recall rate of 80.85%. The model achieves a Kappa statistic of 0.444, indicating moderate agreement, and an F1 score of 50.23%, which balances precision and recall. The confidence interval for accuracy is (90.58%, 93.84%), underpinning the reliability of the model.

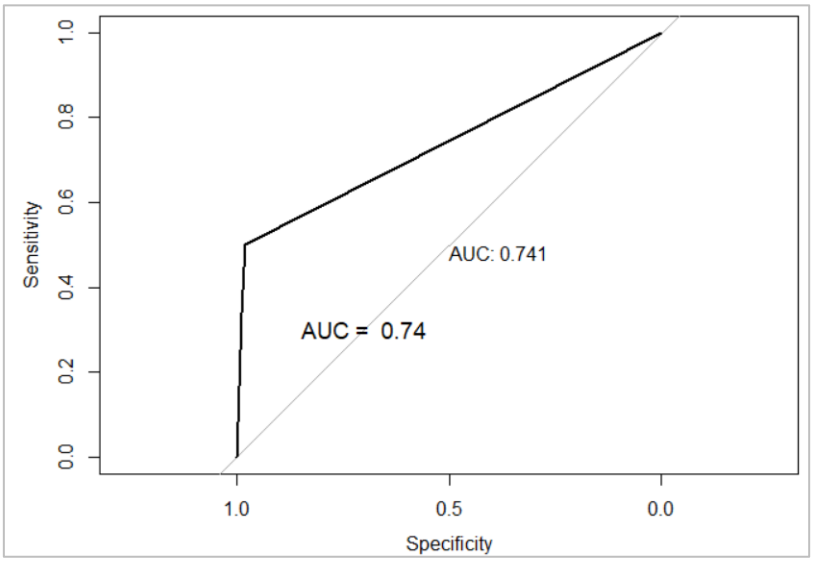


Fig 5: ROC curve

Further insight into the model's performance is gleaned from the ROC curve analysis in figure 5. Initially, the sensitivity of the model increases sharply, signifying a high true positive rate up to a sensitivity of 0.5. Beyond this point, the curve approaches the 45-degree line in the ROC space, indicating a decrease in test accuracy with increasing sensitivity. The Area Under the Curve (AUC) for this model is 0.74, reflecting its overall effectiveness in distinguishing between fraudulent and non-fraudulent cases up to a certain threshold of sensitivity.

Model Comparison:

In our comparative analysis of four distinct machine learning models- KNN, SVM, Naïve Bayes, and Decision Tree applied to healthcare fraud detection, each demonstrated unique strengths and weaknesses across key metrics like accuracy, precision, recall, kappa, and F1 score. The KNN model, prized for its simplicity and interpretability, achieved an accuracy of approximately 92.29%, with a moderate kappa value of 0.415, suggesting reasonable agreement between predictions and actual labels. Despite its impressive performance, the potential for overfitting is indicated in the sensitivity analysis.

The SVM model, leveraging a radial basis function kernel, excelled with an accuracy of about 93.75% and a kappa statistic of 0.5183. This showcases its robustness in complex data scenarios, though it demands a deeper understanding for effective parameter tuning. The Naïve Bayes model, with its probabilistic approach, recorded an accuracy of 91.99% and the highest kappa of 0.5364 among the models. Its balanced accuracy and practical utility stood out, despite the simplifying assumption of feature independence, which may not always align with real-world data complexities.

Meanwhile, the Decision Tree model, noted for its structured and interpretable decision-making process, achieved an accuracy of 92.33% and a kappa of 0.444. This model's appeal lies in its clear hierarchical structure and controlled complexity, despite having a lower precision and F1 score compared to the other models.

In summary, while the KNN and SVM models demonstrated higher accuracy and kappa values, indicative of strong predictive performance, the Naïve Bayes and Decision Tree models offered a balance between accuracy and interpretability. This comparative analysis highlights the necessity of considering a range of performance metrics and factors such as model complexity, interpretability, and dataset specificity when selecting the most suitable model for healthcare fraud detection.

Conclusion

In our analysis of KNN, SVM, Naïve Bayes, and Decision Tree models for healthcare fraud detection, each demonstrated unique strengths in accuracy, precision, and recall, with SVM leading in accuracy (93.75%) and Naïve Bayes showing the highest kappa value (0.5364). The KNN model balanced simplicity with performance, while the Decision Tree offered clear interpretability. This comparative study highlights the importance of considering not just raw performance metrics, but also factors like model complexity and interpretability, in choosing the most suitable model for specific healthcare fraud detection tasks. The findings emphasize that effective model selection hinges on aligning model strengths with the unique demands and characteristics of the healthcare data.