Faculty of Engineering and Technology

Department of Electrical and Computer Engineering

ENCS5341

MACHINE LEARNING AND DATA SCIENCE

Report Assignment #3

"Machine learning for Risk Stratification of Thyroid Cancer"

Prepared by:

Dunia Jaser	1201345
Raneem Daqa	1202093

Instructor: Dr. Yazan Abu Farha.

Section: 1

20 January, 2024

Birzeit

Introduction

The aim of this assignment was to employ machine learning methods in order to predict the probability of relapse in those diagnosed with well-differentiated thyroid cancer. To achieve this, we explored the application of various machine learning models, including k- nearest neighbour (KNN), Support Vector Machines with a Soft margin, and Decision tree. These models were chosen for their potential to capture complex patterns in the data and make accurate predictions regarding cancer relapse.

Data Set

The dataset includes data from 383 patients who were followed for at least 10 years out of a 15-year period. Sixteen clinicopathologic features were analyzed to predict the potential for recurrence. The dataset contains information about age, gender, lifestyle habits (like smoking), thyroid function, physical exam results, pathology details, cancer risk, TNM staging (Tumor, Node, Metastasis), clinical stage, treatment response, and whether or not the cancer came back. No missing values exist for any feature, and the majority of the features consist of categorical data.

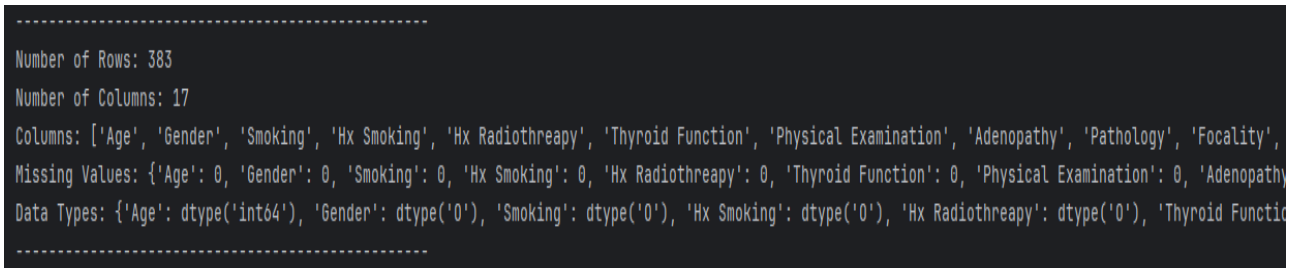

Figure 1: Data set summary

Descriptive Statistics

Numerical Feature (Age)

- Count: 383 patients
- Mean Age: Approximately 40.87 years
- Standard Deviation: 15.13 years
- Minimum Age: 15 years
- 25th Percentile: 29 years
- Median Age (50th Percentile): 37 years
- 75th Percentile: 51 years
- Maximum Age: 82 years

```
------------------------------------------------
Descriptive Statistics for Numerical Columns:
              Age
count  383.000000
mean    40.866841
std     15.134494
min     15.000000
25%     29.000000
50%     37.000000
75%     51.000000
max     82.000000
------------------------------------------------
```

Figure 2: Descriptive Statistics for Age feature

Categorical Features

- Gender: Predominantly female (312 out of 383)
- Smoking History: Majority are non-smokers (334 out of 383)
- Radiotherapy History: Most patients did not have a history of radiotherapy (355 out of 383)
- Thyroid Function: Most are euthyroid (332 out of 383)
- Physical Examination: Most common finding is 'Multinodular goiter' (140 occurrences)
- Adenopathy: Most patients do not have adenopathy (277 out of 383)
- Pathology: 'Papillary' is the most common (287 out of 383)
- Focality: 'Uni-Focal' is more common than 'Multi-Focal' (247 out of 383)
- Risk Level: Most are categorized as low risk (249 out of 383)
- T Stage: 'T2' is the most frequent (151 out of 383)
- N Stage: 'N0' indicates no regional lymph node metastasis (268 out of 383)
- M Stage: 'M0' indicates no distant metastasis (365 out of 383)
- Stage: Most are in Stage I (333 out of 383)
- Response to Treatment: 'Excellent' response is the most common (208 out of 383)
- Recurrence: Majority did not experience recurrence (275 out of 383)

```
------------------------------------------------
Descriptive Statistics for Categorical Columns:
       Gender Smoking Hx Smoking Hx Radiothreapy  ...    M Stage   Response Recurred
count     383        383                    383  ...  383   383        383       383
unique      2          2                      2  ...    2     5          4         2
top         F         No                     No  ...   M0     I  Excellent        No
freq      312        334                    376  ...  365   333        208       275
```

Figure 3: Descriptive Statistics for Categorical Features

Data Visualizations

Proportion of Risk Levels for Different Pathologies:

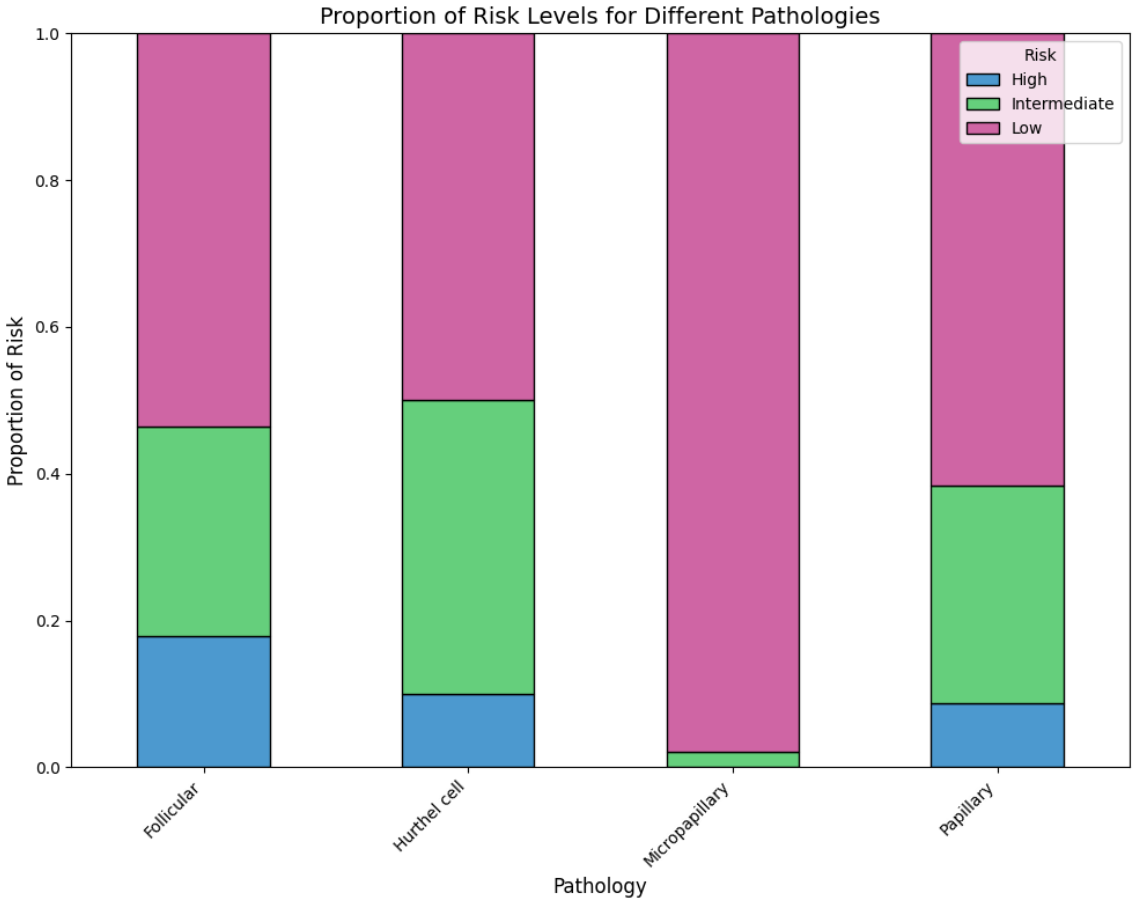

Figure 4: Stacked bar chart of the distribution of risk levels

The stacked bar chart visualizes the risk levels for four different pathologies: Follicular, Hurtle Cell, Micropapillary, and Papillary. It demonstrates that for Follicular and Hurtle Cell pathologies, there is a significant proportion of low-risk cases, with a smaller intermediate risk, and minimal high-risk cases. In contrast, Micropapillary pathology predominantly falls into the low-risk category, with negligible intermediate and high-risk proportions. The Papillary pathology almost entirely consists of low-risk cases, with an insignificant amount of

intermediate risk and no high-risk cases evident. This distribution indicates that the majority of cases in these pathologies are considered low risk.

Combined Age Distribution by Gender:

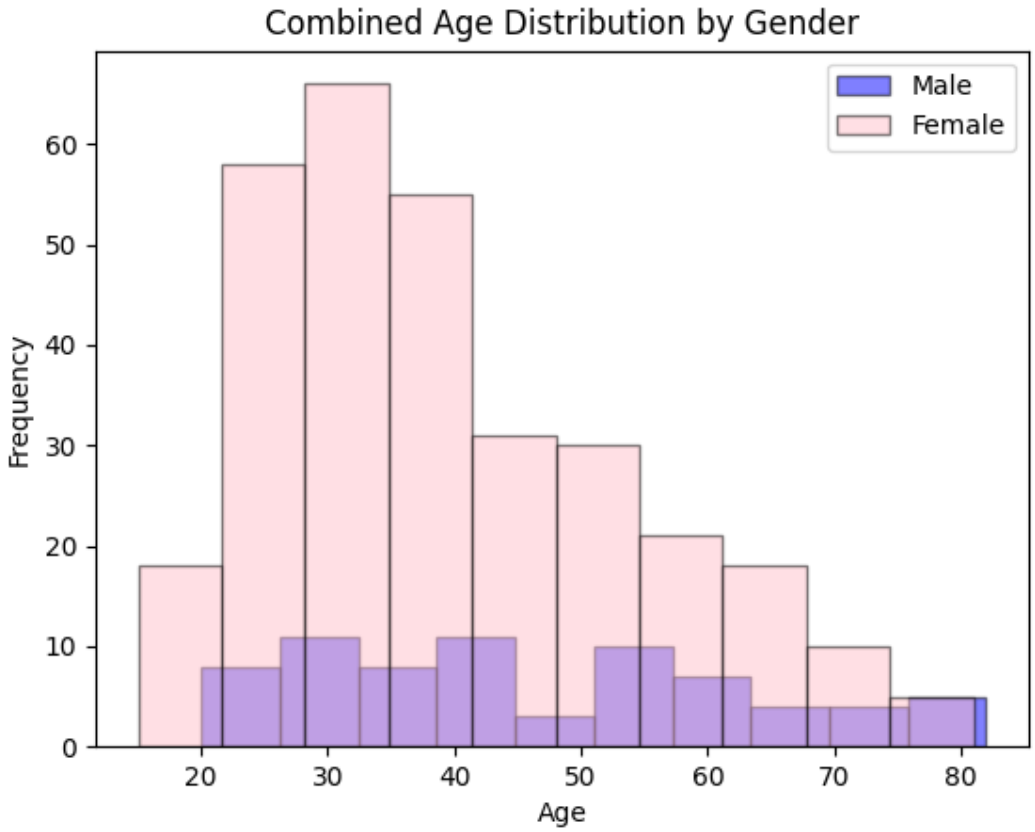

Figure 5: Combined Age Distribution by Gender

The presented histogram illustrates the distribution of ages, categorized by gender. The male frequencies are depicted in blue, while the female frequencies are represented in pink. Interestingly, both distributions seem to follow a similar pattern, with a notable peak in the 40-50 age range indicating that individuals within this age bracket make up the largest group in the sample for both genders. As the age range progresses beyond 50, the frequency for both genders decreases, although it can be noted that the male frequency declines at a steadier pace compared to the female frequency, which showcases a more abrupt decrease. Additionally, it is worth noting that the frequency of females in the 20-30 age range is higher compared to males, while for other age ranges, the frequencies are relatively similar or slightly higher for females.

Box Plot of Age by Stage:

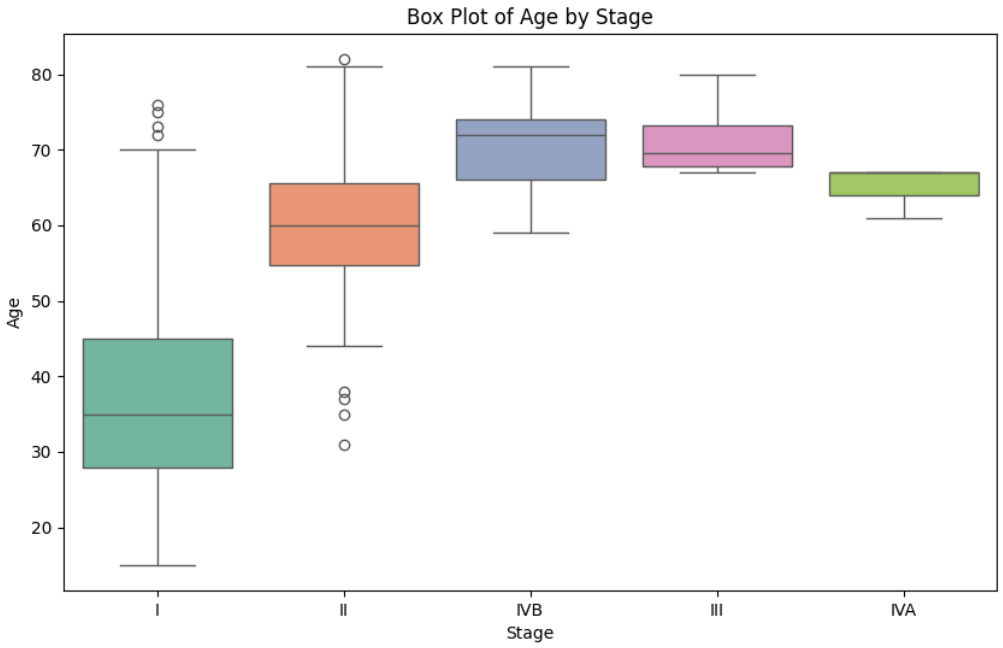

Figure 6: The age distribution across different stages of thyroid pathology.

The box plot illustrates the age distributions for different stages, possibly of a condition or process, with median ages varying across the stages. Stage I has a broad age range with a median around 45 and outliers suggesting ages from early 20s to over 70. Stage II shows a higher median age near 60, with a spread from the mid-40s to mid-70s and some older outliers. Stage IVB presents a narrower range around a median of 55, while Stage III has a median close to 50, with a similar spread. Stage IVA indicates the narrowest age distribution with its median also around 50. The plot suggests a general trend of higher median ages in the initial stages, with a slight decrease in median ages at later stages.

Experiments, Results and Analysis

Since we have known input-output pairs, the choice was made to employ a supervised machine learning approach, which relies on training models with labeled data. Since we have a small dataset for thyroid cancer, Decision Trees, SVMs, and KNN prove effective due to their ability to handle limited data efficiently. Decision Trees offer easy interpretability, SVMs skillfully avoid overfitting, and KNN's simplicity allows for capturing essential patterns, making them well-suited for precise and reliable medical predictions.

For model evaluation, we implement:

1. K- Nearest Neighbour classifier (KNN)

The KNN classifier's performance is assessed for various values of k (1, 3, 5, 7, 9, 11, 12, 15, 18, 20, 25), with accuracies plotted to determine the best k value:

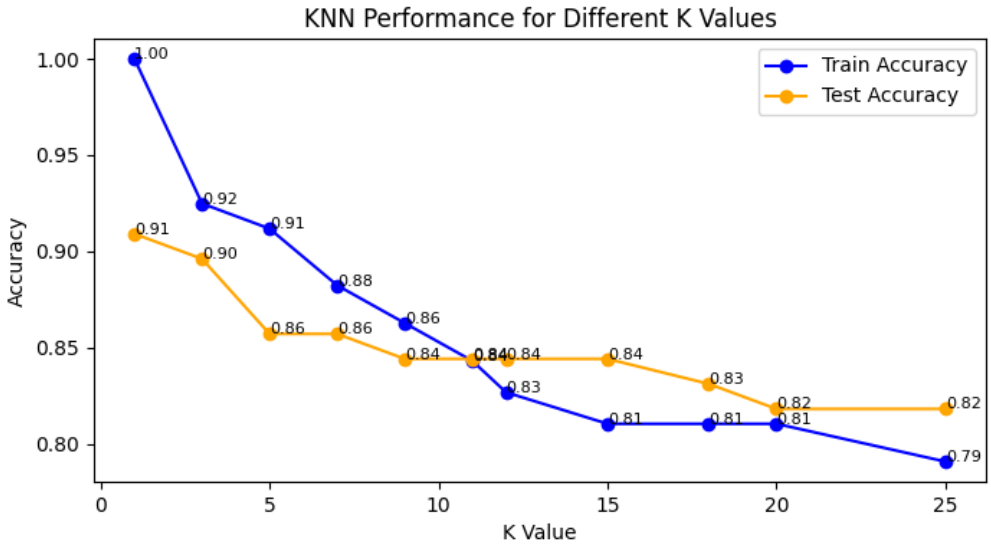

Figure 7: Performance of The KNN classifier

The figure above shows how the K-Nearest Neighbours (KNN) classifier's accuracy changes with different K values. The relationship between the K value and model accuracy reveals crucial insights. As K increases, the training accuracy decreases, suggesting reduced overfitting and better generalization. Test accuracy initially dips, then stabilizes, and finally declines with larger K values, aiding in identifying the optimal K. The model achieves the best balance between training and testing performance at K=11, where both accuracies are around 0.84. This indicates K=11 as the optimal choice for this KNN model. In contrast, a K value of 1 shows perfect training accuracy but lower test accuracy, indicating overfitting, while K=3 shows some improvement but is not optimal. The analysis highlights the importance of tuning K in KNN to achieve a balance between learning and generalizing.

2. Support Vector Machine (SVM)

In a linear Support Vector Machine (SVM) analysis, various regularization strengths, represented by different C values (0.1, 1, 10, 100), are systematically evaluated. For each C

value, the model is trained and tested, and the resulting training and testing accuracies are computed.

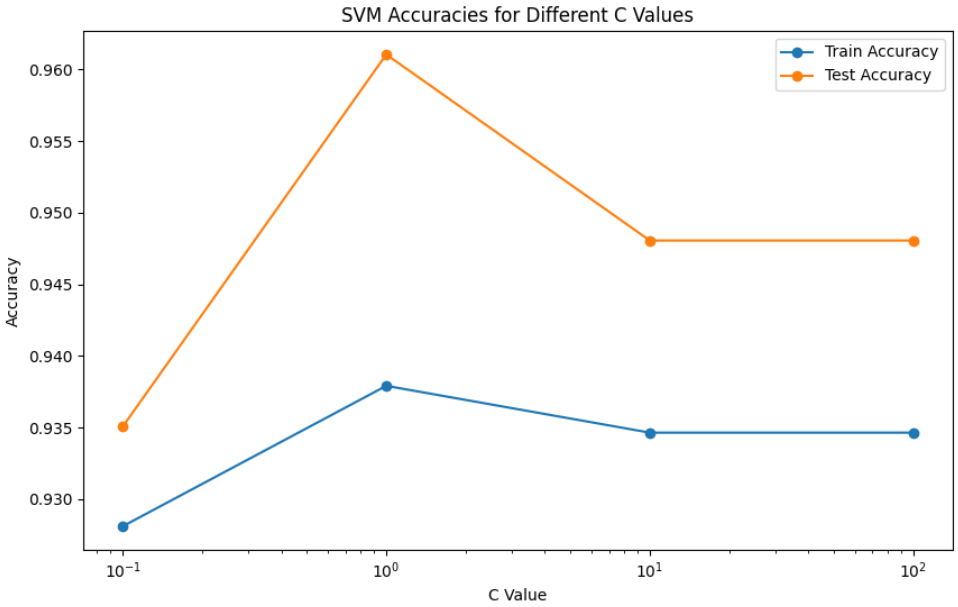

Figure 8: SVM Accuracies for Different C Values

The optimal performance on the test data is attained at C=1, which suggests that this level of regularization effectively prevents the model from fitting excessively to the training data while preserving a sufficient level of complexity to capture the underlying patterns. Increasing C beyond this point results in overfitting (very large C), as evidenced by a reduction in test accuracy, indicating that the model is too finely tuned to the training data and loses its predictive power on unseen data. The decrease in training accuracy with higher C values is atypical and may point to an over-penalization of the model's complexity, leading to a reduction in the ability to correctly classify the training samples. Thus, the value of C=1 emerges as the most appropriate in this context, offering a harmonious balance between model complexity and generalization. To prove the earlier discussion on the SVM's performance across various C values, we refer to the numerical data provided by classification reports. These reports detail the precision, recall, and F1 scores for the model at different levels of regularization.

For C=0.1, the model has good precision and recall but is slightly less confident in predicting class '1'. When we increase C to 1, both precision and recall for class '1' improve, and the model reaches its highest accuracy at 96%. This is evidence that the model with C=1 is the most effective, balancing between being too lenient and too strict.

However, as we continue to increase C to 10 and then to 100, we don't see any further improvements in recall for class '1'; it remains at 0.79, despite precision being perfect at 1.00. This suggests that while the model is highly confident in its predictions (high precision), it's not capturing all the instances of class '1' (no improvement in recall), likely due to overfitting. The accuracy slightly drops to 0.95, reinforcing the point that a higher C value does not necessarily lead to better generalization to new data. These reports substantiate the idea that a C value of 1 is optimal for this dataset, as it achieves the highest accuracy and a balanced precision-recall trade-off.

Classification report for C=0.1:

	precision	recall	f1-score	support
0	0.93	0.98	0.96	58
1	0.94	0.79	0.86	19
accuracy			0.94	77
macro avg	0.94	0.89	0.91	77
weighted avg	0.94	0.94	0.93	77

Classification report for C=1:

	precision	recall	f1-score	support
0	0.95	1.00	0.97	58
1	1.00	0.84	0.91	19
accuracy			0.96	77
macro avg	0.98	0.92	0.94	77
weighted avg	0.96	0.96	0.96	77

Classification report for C=10:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	58
1	1.00	0.79	0.88	19
accuracy			0.95	77
macro avg	0.97	0.89	0.92	77
weighted avg	0.95	0.95	0.95	77

Classification report for C=100:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	58
1	1.00	0.79	0.88	19
accuracy			0.95	77
macro avg	0.97	0.89	0.92	77
weighted avg	0.95	0.95	0.95	77

Figure 9: Classification reports

3. Decision Tree classifier

In a decision tree model analysis, different tree depths (2, 4, 6, 8, 10) are methodically tested to identify the depth that best balances model fit and predictive performance.

The accuracy plot in Figure 10 shows that as the depth of the tree increases, the training accuracy continues to rise, indicating that the model fits the training data increasingly well. However, the test accuracy peaks at a depth of 4 and then starts to decline, which suggests that deeper trees are starting to overfit the training data—meaning they are capturing noise or patterns that do not generalize to unseen data.

```
--------------------------------------------------
The best test accuracy is 0.97 at max depth 4.
--------------------------------------------------
```

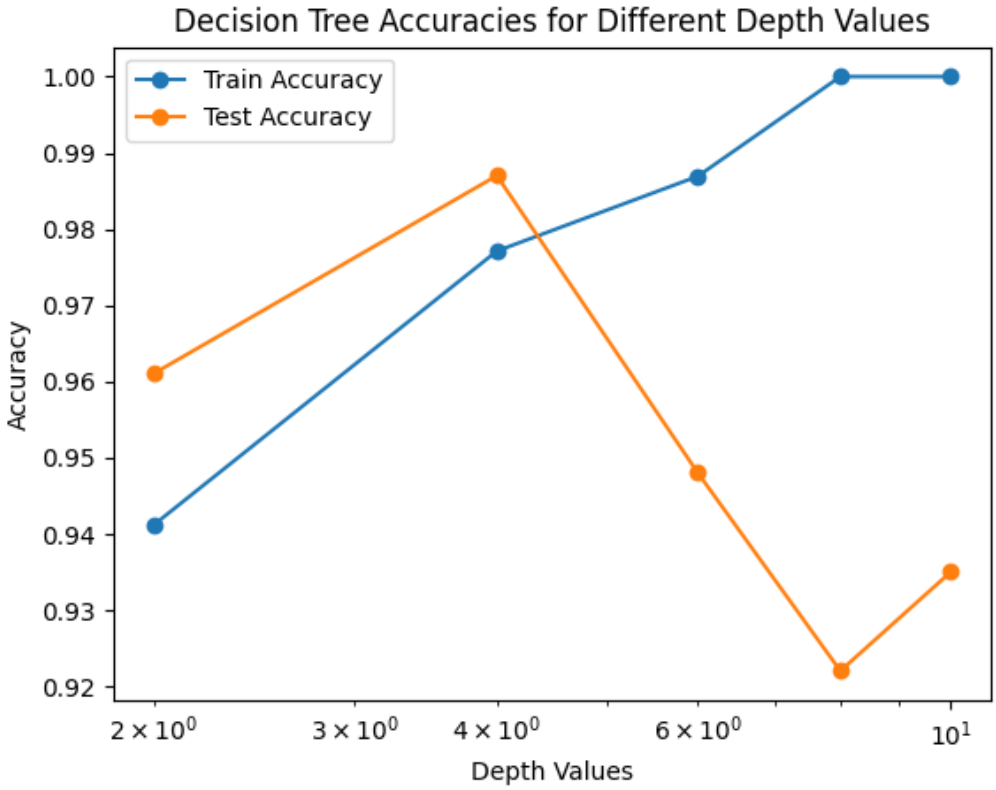

Figure 10: Decision Tree Accuracies for test and train data

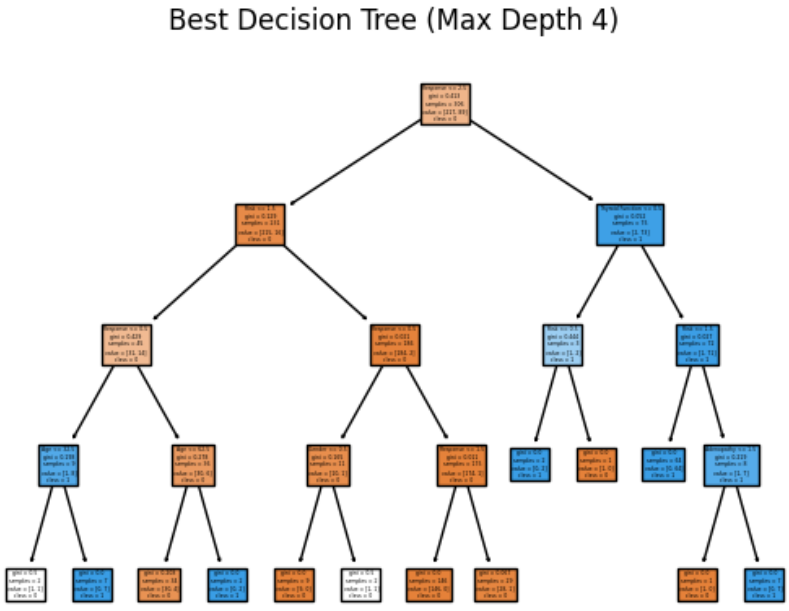

Figure 11: The best decision tree

The detailed output of the best decision tree, with a maximum depth of 4, showcases the decision-making process of the model. The initial split is made on the 'Response' feature: samples with a 'Response' value above 2.5 follow one branch, while the rest follow another. Subsequent splits further refine the classification, with 'Risk', 'Age', 'Gender', Thyroid Function', and 'Adenopathy' being key features used to make decisions at various nodes. Notably, 'Age' is used twice as a decision criterion, with different thresholds indicating its significance in predicting the target classes. The counts at the leaves indicate the number of samples that have followed the path to that point, with some leaves showing a clear preference for one class over the other. This decision tree structure aids in interpreting the model by showing which features and thresholds are most influential in predicting the outcome.

```
Best Decision Tree (Max Depth 4) Nodes:
Node 0: feature=Response (index 15), threshold=2.5, count=306
  Node 1: feature=Risk (index 10), threshold=1.5, count=231
    Node 2: feature=Response (index 15), threshold=0.5, count=45
      Node 3: feature=Age (index 0), threshold=32.5, count=9
        Leaf 4: class=0 count=2
        Leaf 5: class=1 count=7
      Node 6: feature=Age (index 0), threshold=62.5, count=36
        Leaf 7: class=0 count=34
        Leaf 8: class=1 count=2
    Node 9: feature=Response (index 15), threshold=0.5, count=186
      Node 10: feature=Gender (index 1), threshold=0.5, count=11
        Leaf 11: class=0 count=9
        Leaf 12: class=0 count=2
      Node 13: feature=Response (index 15), threshold=1.5, count=175
        Leaf 14: class=0 count=146
        Leaf 15: class=0 count=29
  Node 16: feature=Thyroid Function (index 5), threshold=0.5, count=75
    Node 17: feature=Risk (index 10), threshold=0.5, count=3
      Leaf 18: class=1 count=2
      Leaf 19: class=0 count=1
    Node 20: feature=Risk (index 10), threshold=1.5, count=72
      Leaf 21: class=1 count=64
      Node 22: feature=Adenopathy (index 7), threshold=1.5, count=8
        Leaf 23: class=0 count=1
        Leaf 24: class=1 count=7
```

Figure 12: The detailed decision tree

Conclusions and Discussion

When the full dataset is utilized for training, the comparative performance of machine learning models in stratifying recurrence risk in well-differentiated thyroid cancer is noteworthy. The Decision Tree model, optimized at a maximum depth of 4, stands out as the most accurate, achieving an impressive 97% test accuracy. Closely following is the SVM model with a C value of 1, demonstrating strong generalization with a 96% test accuracy. The KNN model, tuned to a K value of 11, shows effectiveness as well but with a lower accuracy of 84%. These models play a crucial role in healthcare by efficiently predicting recurrence risks, thereby aiding in customizing treatment plans and setting appropriate follow-up intervals for patients. Their high accuracy rates underscore their potential in enhancing patient-specific treatment strategies and optimizing healthcare resources in managing thyroid cancer.

However, it's crucial to acknowledge the limitations inherent in each model, such as potential overfitting in the Decision Tree and sensitivity to irrelevant features in KNN. Moreover, the reliance on accuracy as the sole metric might not fully encapsulate the models' clinical efficacy, warranting further exploration into other evaluation metrics like sensitivity and specificity. The generalizability of these results to other datasets and patient populations remains a vital area for future research, alongside ethical considerations in deploying such models in clinical settings. Despite these limitations, the study underscores the growing relevance of machine learning in improving thyroid cancer management.