**Question 1**

(a) Class is binary, only two categories: no-recurrence-events and recurrence-events.

(b) Age: Ordered categories

Menopause: Categorical

Tumour-size: Ordered categories

Inv-nodes: Ordered categories

Node-caps: Binary

Deg-malig: Ordered categories

Breast: Binary

Breast quad: Categoric

Irradiat: Binary

(c) A normal distribution. Normal(51.1, 10.1).

(d)

**Question 2**

(a) Accuracy of majority class classifier: 65.9794%

ZeroR predicts class value: no-recurrence-events, therefore, ZeroR predicts everything as no-recurrence-events.

(b) 71.134%

(c) With numbers 4, 5, 6 and 7, obtained an accuracy of 73.1959%

(d) Obtained identical results, KNN ranks the nearest points and assigns a ranking. In this scenario, the rankings are equivalent for the two methods.

(e) 68.0412% accuracy

The J48 classifier operates by recursively splitting a decision tree. The attribute chosen for the highest accuracy is the one with the maximum information content. At each tree node, the C4.5 algorithm selects the attribute that best divides the data. The criterion for this division is known as normalised information gain, derived from the difference in entropy. The attribute with the highest normalised information gain is then chosen for splitting. The process continues using a divide-and-conquer algorithm, ultimately constructing the decision tree based on a greedy approach.

However, it is important to note that J48 often exhibits overfitting, resulting in slightly lower accuracy compared to other models. Upon cross fold evaluation, there is an improvement in accuracy by over 6%, suggesting a reduction in the observed overfitting tendency.

(f) 70.1031% accuracy, better than the J48.

Highest weighting: 0.6353 \* (normalized) **deg-malig**=3

This is logical since a higher degree of malignancy indicates a greater likelihood of the patient having cancer. J48 also utilised the variable "deg\_malig" to split the classes.

**Question 3**

False positives rate: 0.515 - Bayes

Precision: 0.757 – Bayes

These outcomes, particularly for cancer predictions, are highly unfavourable. Ideally, we aim for a precision close to 1 and a false positive (FP) rate close to 0. A high FP rate suggests that individuals with cancer are incorrectly informed that they do not have it, leading to a lack of treatment for the disease. A FP rate of 0.515, even in the best case, is unacceptably high.

**Question 4**

(a) All categories in each attribute have the same counts. This could indicate a deliberate design choice for this specific dataset, or it may suggest the presence of fabricated data.

(b) SVM classifier accuracy: 93.3673%

The attribute "**Buying**" exhibits almost zero weighting, indicating that it does not effectively capture the data. A similar pattern is observed for the "**Maintenance**" attribute. In contrast, "**Doors**" carries a higher weighting, providing a better representation of the data. "**Persons**" does not seem to contribute significantly to capturing the data. "**Lug\_Boot**" captures the data well overall, but for medium-sized boots, the information is less discernible. Notably, it suggests that larger boots are more likely to be acceptable ("acc"), while smaller boots are likely to be unacceptable ("unacc").

Among the parameters, "**Safety**" stands out as the most effective in capturing the data. It exhibits a very high absolute weighting for both values, indicating its substantial impact on the dataset.

**Question 5**

(a) The “persons” and the “safety” attribute.

(b) J48 accuracy: 94.2177% and SVM accuracy: 94.3878%

SVM provides a clearer interpretation as it assigns weights to each attribute, indicating their contribution to capturing the data. A negative weight suggests an "unacc" classification, while a positive weight indicates an "acc" classification. The magnitude of the weight reflects the extent of the contribution.

On the other hand, J48 generates a complex decision tree with 79 leaves and a size of 108, making it challenging to read and interpret. The splits occur between safety, persons, buying, maintenance, and lug\_boot. Each branch of the tree presents various possibilities with corresponding numbers in the lowest categories, contributing to its overall complexity and making it more difficult to comprehend due to its size.