DM Models 1

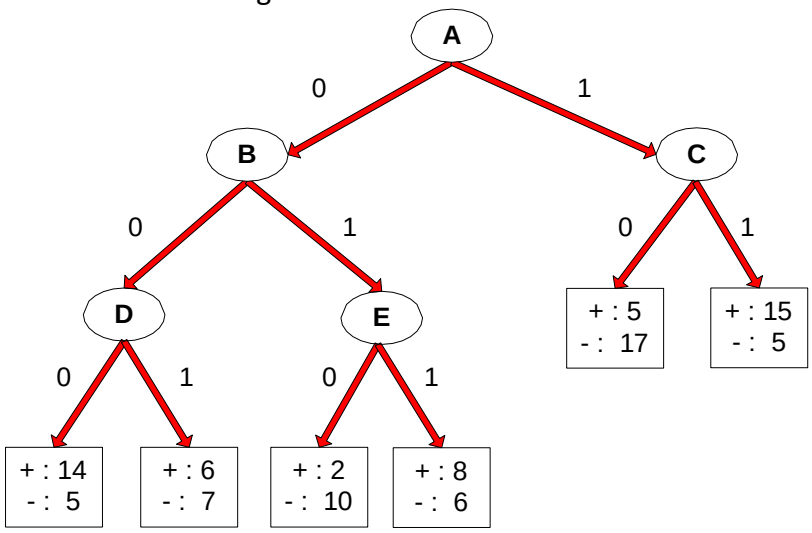
Task 1 (20 points) For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1) Preprocess your Titanic training data;
2) (5 points) Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;
3) (5 points) Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;
4) (5 points) Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;
5) (5 points) Which algorithm is better, Decision Tree or Random Forest? What are your observations and conclusions from the algorithm comparison and analysis?

Solution for 1- 4

Random Forest outperforms Decision Tree, as it avoids overfitting and gives better accuracy by combining multiple decision trees. However, the trade-off is that it may be less interpretable. The choice of model should be based on how well the specific use case balances between model performance and interpretability.

Task 2 (15 points) Understanding Training Error and Testing
Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



(a) (10 points) What is the training error rate for the tree? Explain how you get the answer?

We have leaf nodes with the count for both + and - classes. To calculate the training error rate for the decision tree, we need to sum up the training records that are misclassified

and divide by the total number of training records.

$$Training\ error\ rate = \frac{Misclassified\ training\ records}{Total\ number\ of\ training\ records}$$

Here, the misclassified training records is the sum of – records in a + class and + records in a - class.

Leaf nodes (Left to Right)	No. of records in '+' class	No. of records in '-' class	Class of Leaf node	Misclassified training records
Leaf 1	14	5	+	5
Leaf 2	6	7	-	6
Leaf 3	2	10	-	2
Leaf 4	8	6	+	6
Leaf 5	5	17	-	5
Leaf 6	15	5	+	5

$$\begin{aligned}\text{Misclassified training records} &= (-\text{ in } '+' \text{ class}) + (+\text{ in } '-' \text{ class})\\ &= (5+6+5)+(6+2+5)\\ &= 29\end{aligned}$$

$$\begin{aligned}\text{Total no. of records} &= (\text{No. of records in } '+' \text{ class}) + (\text{No. of records in } '-' \text{ class})\\ &= (14+6+2+8+5+15)+(5+7+10+6+117+5)\\ &= 50+50\\ &= 100\end{aligned}$$

$$\text{Training error rate} = \frac{\text{Misclassified training records}}{\text{Total no. of records}} = \frac{29}{100} = 0.29$$

(b) (5 points) Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

- Starting from the root node (A), we move to the left child of A (B) since A=0.
- At node B, we move to the right child of B (E) since B=1.
- At node E, we move to the left child of E since E=1.
- So the class for the test instance is '-' [Leaf node = (+ : 2, - : 10)].
- Reason: The decision tree follows the path based on the values of the attributes (A, B, C, D, E) in the test instance and arrives at the class label based on the leaf node it reaches. In this case, it reaches the leaf node with the class - since the majority is -.

Task 3 (20 points) Understand Splitting Process
Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Q1: (5 points) What is the overall gini before splitting?

$$Number\ of\ instances\ with\ class\ '+' = 4$$
$$Number\ of\ instances\ with\ class\ '-' = 6$$
$$Total\ number\ of\ instances = 10$$
$$Overall\ Gini\ before\ split = 1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right] = \frac{10^2 - 4^2 - 6^2}{10^2} = \frac{100 - 16 - 36}{100} = \frac{48}{100} = 0.48$$

Q2: (5 points) What is the gain in gini after splitting on A?

We have two nodes after splitting on A, (A = F) and (A = T).

For A = F,

$$Number\ of\ instances\ with\ class\ '+' = 0$$
$$Number\ of\ instances\ with\ class\ '-' = 3$$
$$Total\ number\ of\ instances = 3$$
$$Gini_{A=F} = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = \frac{3^2 - 0^2 - 3^2}{3^2} = \frac{9 - 0 - 9}{9} = \frac{0}{9} = 0$$

For A = T,

$$Number\ of\ instances\ with\ class\ '+' = 4$$
$$Number\ of\ instances\ with\ class\ '-' = 3$$
$$Total\ number\ of\ instances = 7$$
$$Gini_{A=T} = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right] = \frac{7^2 - 4^2 - 3^2}{7^2} = \frac{49 - 16 - 9}{49} = \frac{24}{49} = 0.4898$$
$$Gini\ after\ split = \frac{3}{10} Gini_{A=F} + \frac{7}{10} Gini_{A=T} = \frac{3}{10} * 0 + \frac{7}{10} * \frac{24}{49} = 0 + \frac{12}{35} = \frac{12}{35} = 0.3428$$
$$Gain\ in\ Gini = Gini\ before\ split - Gini\ after\ split = \frac{48}{100} - \frac{12}{35}$$
$$= \frac{(48 * 35) - (12 * 100)}{100 * 35} = \frac{1680 - 1200}{3500} = \frac{480}{3500}$$
$$= 0.1371$$

Q3: (5 points) What is the gain in gini after splitting on B?

We have two nodes after splitting on B, (B = F) and (B = T).

For B = F,

$$Number\ of\ instances\ with\ class\ '+' = 1$$
$$Number\ of\ instances\ with\ class\ '-' = 5$$
$$Total\ number\ of\ instances = 6$$
$$Gini_{B=F} = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right] = \frac{6^2 - 1^2 - 5}{6^2} = \frac{36 - 1 - 25}{36} = \frac{10}{36} = 0.2777$$

For B = T,

$$Number\ of\ instances\ with\ class\ '+' = 3$$
$$Number\ of\ instances\ with\ class\ '-' = 1$$
$$Total\ number\ of\ instances = 4$$
$$Gini_{B=T} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = \frac{4^2 - 3^2 - 1^2}{4^2} = \frac{16 - 9 - 1}{16} = \frac{6}{16} = 0.375$$
$$Gini\ after\ split = \frac{6}{10} Gini_{B=F} + \frac{4}{10} Gini_{B=T} = \frac{6}{10} * \frac{10}{36} + \frac{4}{10} * \frac{6}{16} = \frac{1}{6} + \frac{3}{20}$$
$$= \frac{(1 * 20) + (3 * 6)}{6 * 20} = \frac{20 + 18}{120} = \frac{38}{120} = 0.3166$$
$$Gain\ in\ Gini = Gini\ before\ split - Gini\ after\ split = \frac{48}{100} - \frac{38}{120}$$
$$= \frac{(48 * 120) - (38 * 100)}{100 * 120} = \frac{5760 - 3800}{12000} = \frac{1960}{12000}$$
$$= 0.1633$$

Q4: (5 points) Which attribute would the decision tree choose?

The attribute with highest gain in Gini with be chosen by decision tree algorithm. After comparing both the gain in Gini after split using A and B attributes, we can see that the gain in Gini after split using attribute B is higher when compared to the gain in Gini after split using attribute A.

Therefore, attribute B will be chosen by the decision tree.

Task 4: (10 points) Please answer and explain.

Q1: (5 points) Are decision trees a linear classifier? Why?

No, decision trees are not linear classifiers. The input features are linear combinations in linear classifiers whereas decision tree makes decision by classifying and regressing the data using true or false and 1 or 0 based on certain thresholds which are determined during implementation. In

decision tree the data is partitioned into two or more subsets at particular node depending on threshold value. The decision tree's decision boundary is a series of these splits, which can be highly nonlinear and complex, allowing it to model and capture intricate relationships within the data.

Q2: (5 points) Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

During implementation, Gini index is more preferred over misclassification because it is more robust to imbalanced data, Gini index is less sensitive to class imbalances because it accounts for the class probabilities within a node. It won't heavily favour the majority class. Moreover, Ginni index results in decision trees that generalize better compared to unseen data as it considers the distribution of classes within node. Additionally, Ginni index tends to produce more balanced tree, which precludes risk of over fitting misclassification on other hand lead to more complex tree and overfits data.

Task 5: (10 points) What are the weaknesses of bagging? What is the difference between bagging and random forests, and why such difference can overcome the weaknesses of bagging?

Weakness of Bagging:

1) **Lack of Diversity:** Bagging relies on creating multiple models by training on bootstrap samples, which are randomly drawn datasets with replacement that are a subset of the training data. However, these subsets can have significant overlap between them and these may lead to models that are highly correlated.
2) **Unfit for highly biased data:** It is not efficient to imbalance in dataset.
3) **High Bias:** Bagging have considerable high bias then single model as it fits multiple models on bootstrapped subsets of data.

Difference between bagging and random forest:

	Bagging	**Random Forest**
Feature Randomness	Uses bootstrapped samples of the data to train each decision tree. However, it does not involve feature randomness, meaning that all features are considered at each split in decision trees.	Uses feature randomness by selecting a random subset of features for consideration at each split in each decision tree. This feature selection decorrelates the trees and reduces the risk of overfitting.
Overfitting Control	Reduces overfitting by averaging predictions across the decision	Reduces both variance and overfitting by introducing feature

	trees but doesn't necessarily address the issue of feature overfitting.	randomness and the de-correlated decision trees.

This is how Random Forests overcomes the weakness of Bagging.

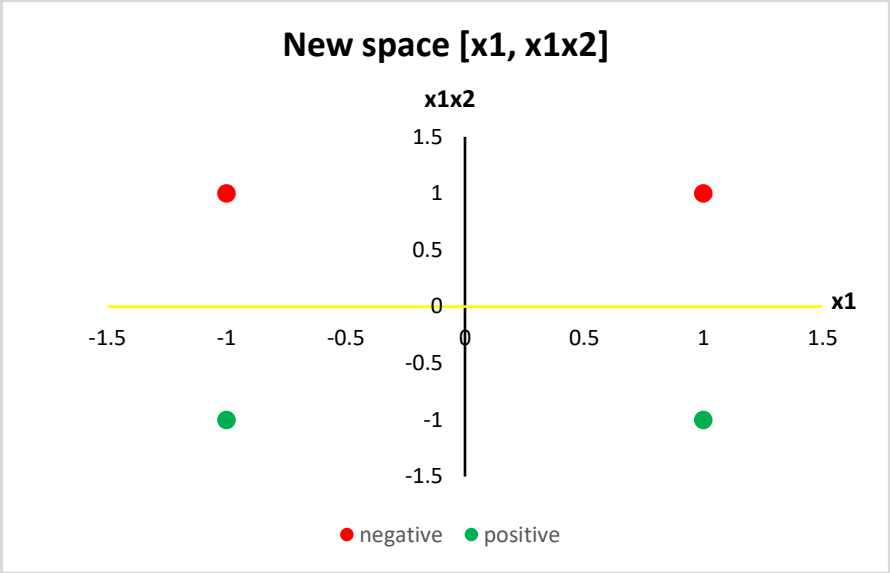
1) **De-correlated Decision Trees:** By introducing randomness in feature selection, Random Forests de-correlate the decision trees more effectively than traditional bagging. This helps in reducing the risk of overfitting and improves the generalization performance.
2) **Enhanced Diversity:** The use of feature randomness adds an additional layer of diversity to the decision trees in Random Forest. This helps capture different aspects of the data and makes the ensemble more robust.
3) **Better Generalization:** The combination of de-correlated decision trees and enhanced diversity often leads to Random Forests having better generalization performance compared to traditional bagging.

Task 6: (20 points) Construct a support vector machine that computes the kernel function. Use four values of +1 and -1 for both inputs and outputs:
[−1, −1] (negative)
[−1, +1] (positive)
[+1, −1] (positive)
[+1, +1] (negative).
Map the input [x1, x2] into a space consisting of x1 and x1x2. Draw the four input points in this space, and the maximal margin separator. What is the margin? 【To be consistent with our lecture notes, margin is defined as the distance from the middle way/hyperplane to either support vectors. 】

Mapping input points in $[x_1, x_2]$ space into a new space $[x_1, x_1x_2]$

[-1, -1] maps to [-1, +1]. Class = negative
[-1, +1] maps to [-1, +1]. Class = positive
[+1, -1] maps to [-1, +1]. Class = positive
[+1, +1] maps to [-1, +1]. Class = negative

From this new mapping, we can see that the negative samples have $x_1x_2 = +1$ and the positive samples have $x_1x_2 = -1$.



The yellow line in the above plot is the maximal margin separator and is given by the equation $x_1 x_2 = 0$.

Distance between a point (x_1, y_1) and a line $Ax + By + C = 0$ is $d = \frac{Ax_1 + By_1 + C}{\sqrt{A^2 + B^2}}$.

In the new space $[x_1, x_1 x_2]$, $y = x_1 x_2$. Therefore, equation of the maximal margin separator line is $y = 0$.

$$Distance\ between\ the\ point\ (1,1)\ and\ line\ (y = 0) = \frac{0 * 1 + 1 * 1 + 0}{\sqrt{0^2 + 1^2}} = 1$$

Therefore, margin = 1

Task 7: (10 points) Recall that the equation of the circle in the 2-dimensional plane is (x1 − a)^2 + (x2 − b)^2 − r2 = 0. Please expand out the formula and show that every circular region is linearly separable from the rest of the plane in the feature space (x1, x2, x1^2, x2^2).

The equation of the circle in 2-dimensional space is given by:

$$(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$$

Expanding the equation, we get:

$$x_1{}^2 - 2ax_1 + a^2 + x_2{}^2 - 2bx_2 + b^2 - r^2 = 0$$
$$\implies -2ax_1 - 2bx_2 + x_1{}^2 + x_2{}^2 + (a^2 + b^2 - r^2) = 0$$

If we consider the new feature space (x_1, x_2, x_1^2, x_2^2), we can write the above equation as:

$$w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_4^2 + c = 0$$

where $w_1 = -2a, w_2 = -2b, w_3 = 1, w_4 = 1, c = a^2 + b^2 - r^2$.

For any point inside the circle, $w_1x_1 + w_2x_2 + x_3 + x_4 + c < 0$.
For any point outside the circle, $w_1x_1 + w_2x_2 + x_3 + x_4 + c > 0$.

We can draw any linear boundary $w_1x_1 + w_2x_2 + x_3 + x_4 + c = 0$ which linearly separates the points inside the circle from the points outside the circle. This shows that every circular region is linearly separable from the rest of the plane in the feature space (x_1, x_2, x_1^2, x_2^2).

Task 8: (10 points) Recall that the equation of an ellipse in the 2-dimensional plane is c(x1 − a)^2 + d(x2 − b)^2 − 1 = 0. Please show that an SVM using the polynomial kernel of degree 2, K(u, v) = (1 + u · v)^2, is equivalent to a linear SVM in the feature space (1, x1, x2, x1^2, x2^2, x1x2) and hence that SVMs with this kernel can separate any elliptic region from the rest of the plane.

To show that an SVM using the polynomial kernel of degree 2, $K(u, v) = (1 + u.v)^2$ is equivalent to a linear SVM in the feature space $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$, we need to find a transformation φ such that $K(u, v) = \varphi(\mathrm{u}).\varphi(\mathrm{v})$.

Let $u = [u_1, u_2]$ and $v = [v_1, v_2]$.

We can define φ(u) and φ(v) as:
$\varphi(\mathrm{u}) = [1\ \sqrt{2}u_1\ \sqrt{2}u_2\ u_1^2\ u_2^2\ \sqrt{2}u_1u_2]$
$\varphi(\mathrm{v}) = [1\ \sqrt{2}v_1\ \sqrt{2}v_2\ v_1^2\ v_2^2\ \sqrt{2}v_1v_2]$.

$$K(u, v) = (1 + u.v)^2 = (1 + [u_1\ u_2].[v_1\ v_2])^2 = (1 + u_1v_1 + u_2v_2)^2$$

$$= 1 + u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_1 + 2u_2v_2 + 2u_1v_1u_2v_2$$

Rearranging the terms in the above equation, we get

$$K(u, v) = 1 + \left(\sqrt{2}u_1\right)\left(\sqrt{2}v_1\right) + \left(\sqrt{2}u_2\right)\left(\sqrt{2}v_2\right) + (u_1^2)(v_1^2) + (u_2^2)(v_2^2) + \left(\sqrt{2}u_1u_2\right)\left(\sqrt{2}v_1v_2\right)$$

$$= \left[1\ \sqrt{2}u_1\ \sqrt{2}u_2\ u_1^2\ u_2^2\ \sqrt{2}u_1u_2\right].[1\ \sqrt{2}v_1\ \sqrt{2}v_2\ v_1^2\ v_2^2\ \sqrt{2}v_1v_2]$$

$$= \varphi(\mathrm{u}).\varphi(\mathrm{v})$$

Hence, an SVM using the polynomial kernel of degree 2, $K(u, v) = (1 + u.v)^2$ is equivalent to a linear SVM in the feature space $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$,

The equation of an ellipse in 2-dimensional space is given by:

$$c(x_1 - a)^2 + d(x_2 - b)^2 - 1 = 0$$

Expanding the equation, we get:

$$cx_1{}^2 - 2cax_1 + ca^2 + dx_2{}^2 - 2dbx_2 + db^2 - 1 = 0$$
$$\Longrightarrow (ca^2 + db^2 - 1) - 2cax_1 - 2dbx_2 + x_1{}^2 + x_2{}^2 = 0$$

If we consider the feature space $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$, we can write the above equation as:

$$w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 = 0$$

where $w_o = ca^2 + db^2 - 1, w_1 = -2ca, w_2 = -2db, w_3 = 1, w_4 = 1, w_5 = 0.$

For any point inside the ellipse, $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 < 0.$
For any point outside the ellipse, $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 > 0.$

We can draw any linear boundary $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 = 0$ which linearly separates the points inside the ellipse from the points outside the ellipse. This shows that SVMs with this kernel can separate any elliptic region from the rest of the plane.