Predictive Model Plan And Layout

1.Model Logic (Generated with GenAI) :

The below givel model algorithm was generated by Google's Gemini 2.5 Flash

```
function BuildDecisionTree(dataset, features, target):
if dataset is empty or all target values are the same:
return a Leaf Node with the majority class (or base class if empty)

Find the Best Split:
Initialize best_feature = None, best_threshold/category = None, max_info_gain = -infinity

For each feature in features:
For each possible split point (threshold for numerical, category for categorical):
Calculate Gini Impurity (or Entropy) of the parent node.
Split dataset into left_child_data and right_child_data based on the split point.
Calculate Gini Impurity (or Entropy) for left_child_data and right_child_data.
Calculate Information Gain for this split.

If Information Gain > max_info_gain:
max_info_gain = Information Gain
best_feature = current feature
best_split_point = current split point

if max_info_gain is close to zero (no good split found) or max_depth_reached or
min_samples_criteria_met:
return a Leaf Node with the majority class of the dataset

Create an Internal Node:
node.feature = best_feature
node.split_point = best_split_point
node.left_child = BuildDecisionTree(left_child_data, features, target)
node.right_child = BuildDecisionTree(right_child_data, features, target)

return node


function Predict(tree_node, new_data_point):
if tree_node is a Leaf Node:
return tree_node.class_label

If new_data_point[tree_node.feature] <= tree_node.split_point:  // For numerical features
return Predict(tree_node.left_child, new_data_point)
Else:
return Predict(tree_node.right_child, new_data_point)
```

2.Justification for Model Choice

- The working of this model is simple and easily explainable to the stakeholders.

- The model is really easy to implement as it doesn't require feature scaling and works well with both numerical and categorical vaeriables.

- with respect to this usecase, the model working consists of refining the data with respect to a series of questions (binary or conditional) which more or less similar to the financial assessment of an individual.

- The model automatically selects the most impactful features for predictive analysis, which is crucial for building a robust model and the model is also suitable with small datasets such as Geldium's which poses a chance of overfitting incase of other models.

3.Evaluation Strategy

• **Adopted performance metrics :**

- Accuracy, recall and confusion matrix and AUC_ROC curve visualization.

• **Interpretation of performance metrics :**

1. confusion matrix consists of 4 cells TP,TN,FP,FN. High TP and TN indicates that your model is excelling and vice versa

2 .High Accuracy indicates a high percentage of correct predictons and vice versa

3 . High recall denotes your model is good at catching actual positive cases and low recall implies the model is missing many actual positive cases

4. For AUC_ROC curve if the area under the curve if closer to 1 than model has better discriminative power, if it is equal to 0.5 then models performs with random chance and if it is less than that then the model is worse the random

• Bias Reduction Strategies :

- Stratifying the dataset with respect to both classes to reduce partiallity and using ensemble methods at the end to generalize the result as much as possible

• Ethical Considerations :

- Checking if any of the column contains customer's sensitive information and dispose of them accordingly and making sure the model performs equally across different key groups through fairness metrics