Exercise 1 Report

Submission Date: 22/12/2024

Student Name: Adir Serruya

Student ID: 316472455

Introduction

In this exercise, I investigate and enhance decision tree models. Specifically, I address the limitation of binary routing in standard decision trees by introducing soft splits. Additionally, I evaluate the performance of these models on multiple datasets using a rigorous evaluation framework.

Programming Task 1: Soft Splits in Decision Trees

Code Implementation

For the soft split implementation, as requested, only the inference functionality was modified.

1. Soft Split Function: I began by implementing a soft_split function. This function evaluates a sample at a split node and assigns a probability of routing the sample to the "opposite" branch instead of the branch dictated by the splitting condition. This probability is controlled by the alpha parameter:
    - With probability α , the sample is routed in the opposite direction.
    - With probability $1-\alpha$, it is routed in the expected direction based on the splitting condition.
2. predict_sample_proba Implementation: I then implemented the predict_sample_proba function. This function simulates the soft split inference for a single sample nnn times (where nnn is the number of simulations). During each simulation, the sample is routed probabilistically through the tree, producing a probability vector. These simulations are averaged to generate a final probability vector for the sample.
3. Overriding predict_proba: Finally, I overrode the predict_proba method of the sklearn decision tree. The new implementation uses predict_sample_proba for all samples in the dataset and returns the averaged probability vectors across nnn simulations. This ensures that the final predictions reflect the uncertainty introduced by the soft splits.

The training process remains unchanged, preserving the original behavior of the sklearn decision tree during the fit process.

1.2 Datasets Used - Classification

For evaluating the classifier, I used the following datasets, each of which has more than 1,000 samples and a variety of features and target classes:

1. **Obesity Dataset:**
    - **Description:** A classification dataset used to predict obesity levels based on health-related behaviors.
    - **Samples:** Over 2,000 samples.
    - **Features:** 16 Total, Gender, Agem Height, Weight, Family history etc.
    - **Target Classes:** Multi-class problem with different obesity categories.
2. **Adult Income Dataset:**
    - **Description:** Predicts whether an individual's income exceeds $50K/year based on demographic and economic attributes.
    - **Samples:** Approximately 48,000 samples.
    - **Features:** 14 Total, Includes age, education, work hours, and more.
    - **Target Classes:** Binary classification (above or below $50K income).
3. **Wine Quality Dataset:**
    - **Description:** Predicts the quality of wine based on chemical properties.
    - **Samples:** Over 1600 samples.
    - **Features:** 11 Total, Includes acidity, alcohol content, and other chemical measures.
    - **Target Classes:** Multi-class classification (wine quality scores).
4. **Bank Marketing Dataset:**
    - **Description:** Predicts whether a customer will subscribe to a term deposit based on marketing campaign data.
    - **Samples:** Over 45,000 samples.
    - **Features:** 17 Total, Includes contact duration, previous campaign outcomes, and more.
    - **Target Classes:** Binary classification (subscribed or not).
5. **Student Success Dataset:**
    - **Description:** Predicts student success in education based on demographic and academic performance.
    - **Samples:** Over 4000 samples.
    - **Features:** Includes parental education level, study time, and previous grades.
    - **Target Classes:** Multi-class classification (Enrolled, Graduate, Dropout).

Datasets - Regression

1. **Garments Worker Productivity Dataset**:
   - **Description**: Predicts the productivity of garment workers based on various operational and environmental factors.
   - **Samples**: 1,197 samples.
   - **Features**: 14 total, including production efficiency, work hours, and departmental indicators.
   - **Target Variable**: Worker productivity.
2. **Air Quality Dataset**:
   - **Description**: Predicts air quality metrics based on atmospheric and environmental factors.
   - **Samples**: 9,471 samples.
   - **Features**: 16 total, including CO, NOx levels, temperature, and humidity.
   - **Target Variable**: Air quality index or pollutant levels.
3. **Bike Sharing Dataset**:
   - **Description**: Predicts the number of bike rentals based on weather conditions, time, and seasonality.
   - **Samples**: 17,379 samples.
   - **Features**: 16 total, including temperature, humidity, and wind speed.
   - **Target Variable**: Number of bike rentals.
4. **Apartments for Rent Dataset**:
   - **Description**: Predicts the rental price of apartments based on various property features and location information.
   - **Samples**: 10,000 samples.
   - **Features**: 21 total, including apartment size, number of rooms, and location data.
   - **Target Variable**: Apartment rental price.
5. **Energy Efficiency Dataset**:
   - **Description**: Predicts energy consumption based on household environmental and operational factors.
   - **Samples**: 19,735 samples.
   - **Features**: 28 total, including temperature, humidity, and equipment usage data.
   - **Target Variable**: Energy consumption or efficiency metrics.

1.3 Exploratory Data Analysis – Obesity Dataset

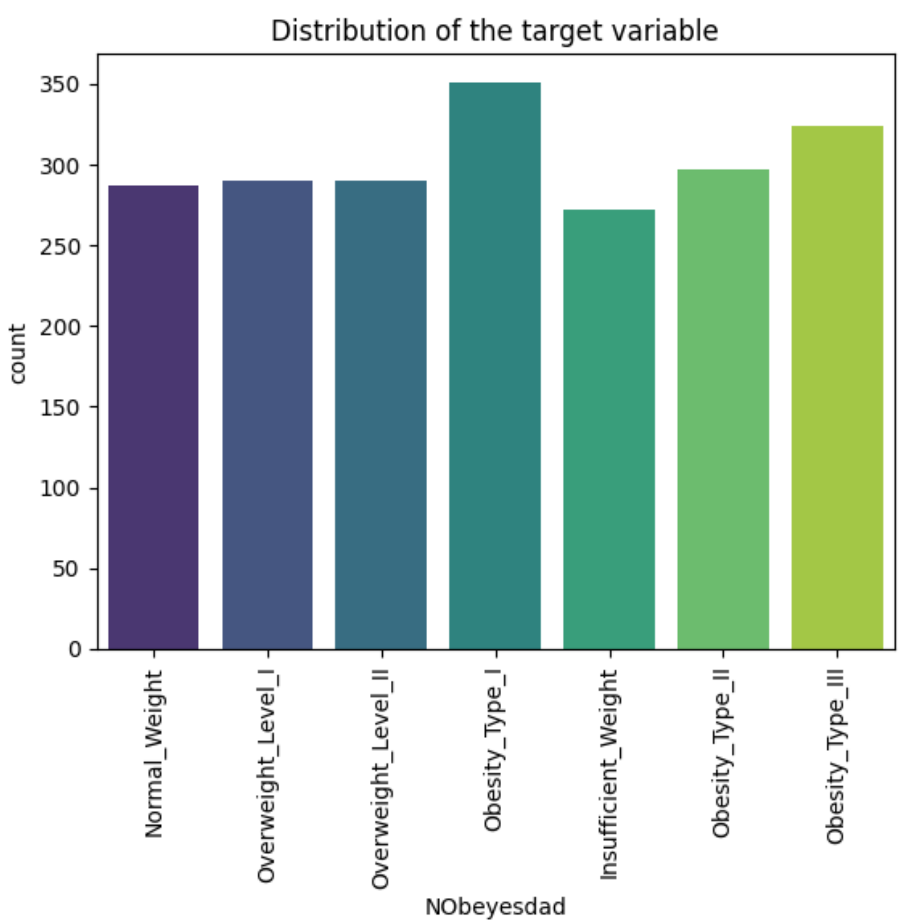
Description Statistics of the continuous variables:

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
mean	24.31	1.70	86.59	2.42	2.69	2.01	1.01	0.66
std	6.35	0.09	26.19	0.53	0.78	0.61	0.85	0.61
min	14.00	1.45	39.00	1.00	1.00	1.00	0.00	0.00
25%	19.95	1.63	65.47	2.00	2.66	1.58	0.12	0.00
50%	22.78	1.70	83.00	2.39	3.00	2.00	1.00	0.63
75%	26.00	1.77	107.43	3.00	3.00	2.48	1.67	1.00
max	61.00	1.98	173.00	3.00	4.00	3.00	3.00	2.00

Description Statistics of the categorical variables:

	Gender	family_history_with_overweight	FAVC	CAEC	SMOKE	SCC	CALC	MTRANS	NObeyesdad
unique	2	2	2	4	2	2	4	5	7
top	Male	yes	yes	Sometimes	no	no	Sometimes	Public_Transportation	Obesity_Type_I
freq	1068	1726	1866	1765	2067	2015	1401	1580	351

Distribution of Target Variable:



Contingency Tables between Categoricals + Chi Square Test for Independence:

family_history_with_overweight	no	yes
NObeyesdad		
Insufficient_Weight	146	126
Normal_Weight	132	155
Obesity_Type_I	7	344
Obesity_Type_II	1	296
Obesity_Type_III	0	324
Overweight_Level_I	81	209
Overweight_Level_II	18	272

```
Chi-Square Test Results:
Chi-Square Statistic: 621.9794354
p-value: 0.0000000
Degrees of Freedom: 6
Result: Significant association between the variables.
```

FAVC	no	yes
NObeyesdad		
Insufficient_Weight	51	221
Normal_Weight	79	208
Obesity_Type_I	11	340
Obesity_Type_II	7	290
Obesity_Type_III	1	323
Overweight_Level_I	22	268
Overweight_Level_II	74	216

```
Chi-Square Test Results:
Chi-Square Statistic: 233.3413036
p-value: 0.0000000
Degrees of Freedom: 6
Result: Significant association between the variables.
```

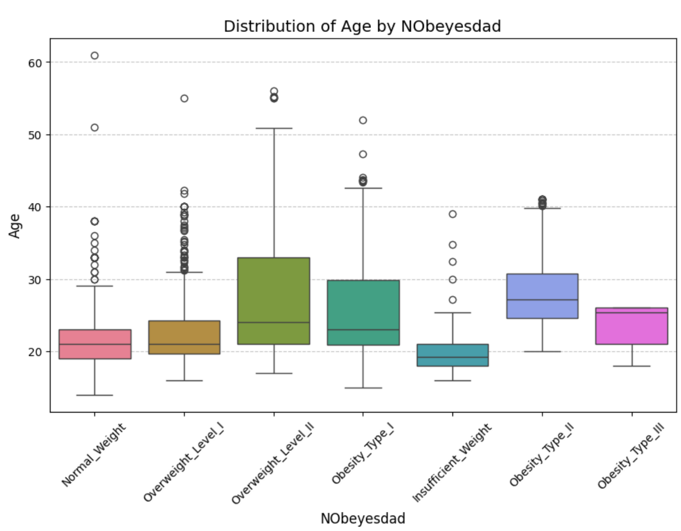
SMOKE	no	yes
NObeyesdad		
Insufficient_Weight	271	1
Normal_Weight	274	13
Obesity_Type_I	345	6
Obesity_Type_II	282	15
Obesity_Type_III	323	1
Overweight_Level_I	287	3
Overweight_Level_II	285	5

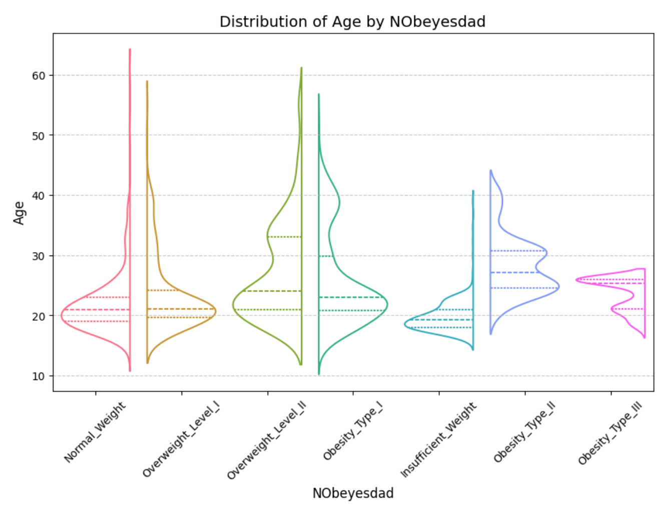
```
Chi-Square Test Results:
Chi-Square Statistic: 32.1378321
p-value: 0.0000154
Degrees of Freedom: 6
Result: Significant association between the variables.
```

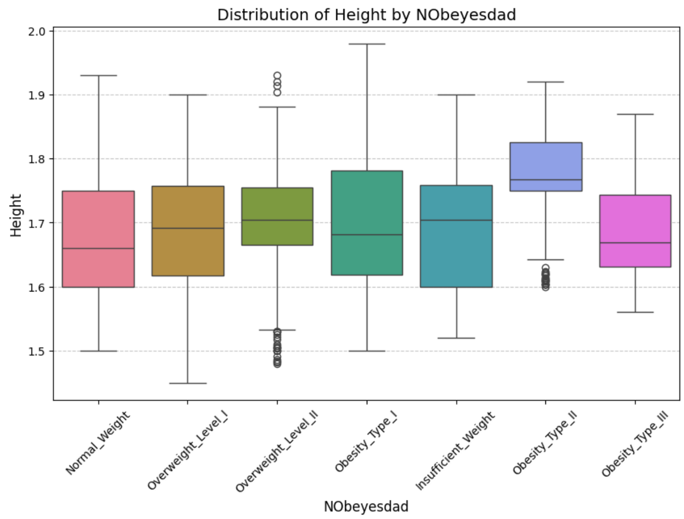
SCC	no	yes
NObeyesdad		
Insufficient_Weight	250	22
Normal_Weight	257	30
Obesity_Type_I	349	2
Obesity_Type_II	296	1
Obesity_Type_III	324	0
Overweight_Level_I	253	37
Overweight_Level_II	286	4

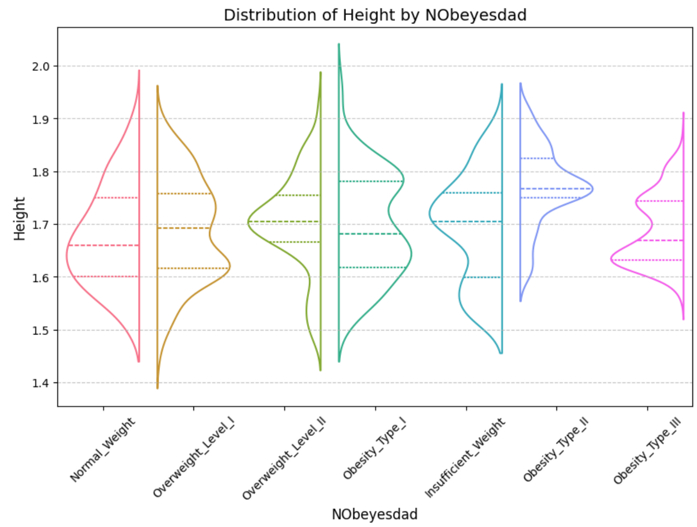
```
Chi-Square Test Results:
Chi-Square Statistic: 123.0238987
p-value: 0.0000000
Degrees of Freedom: 6
Result: Significant association between the variables.
```

Distributions of Continuous variables per category:

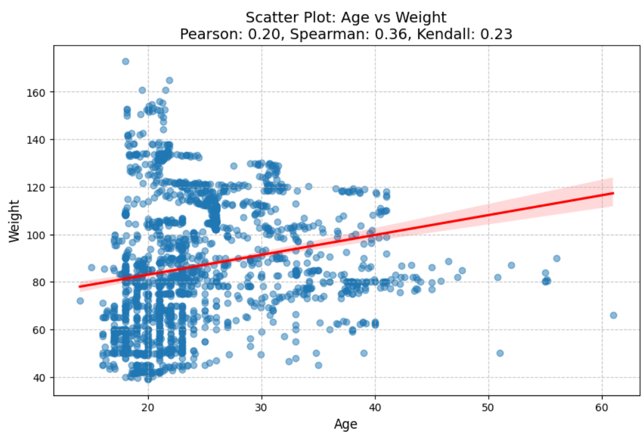






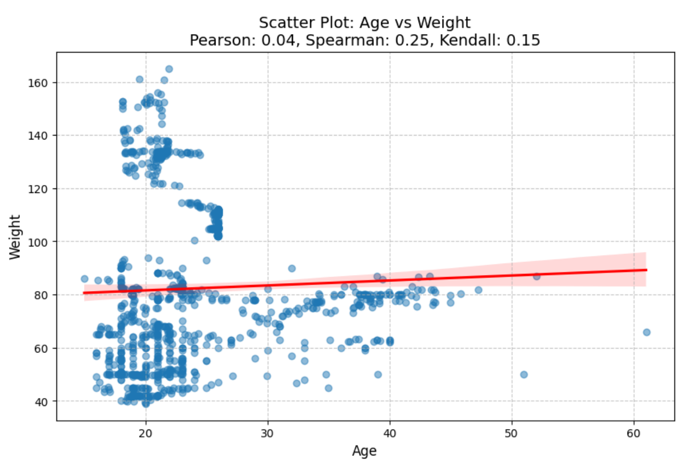


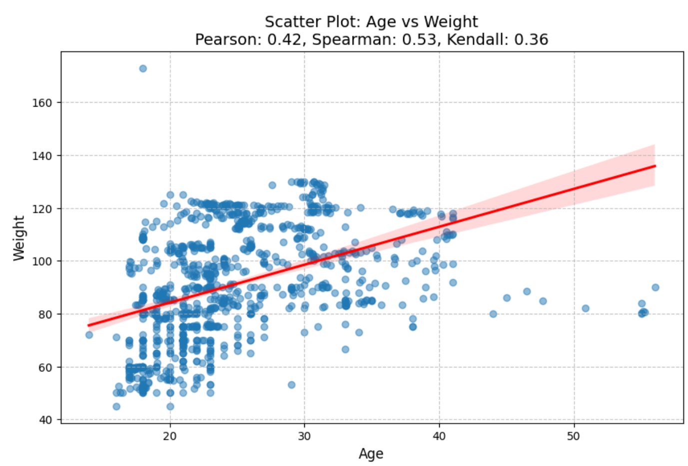
Correlations between continuous variables



The above didn't look very informative so I looked at it separately for females and males…

Females and Males Accordingly :





We can see that for males age is much more correlated with the weight, if the goal of the project was to get the best results possible I would probably try to create interaction features here.

Sensitivity Analysis for Hyperparameters – Task 1

Student Success Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.65	0.78
2	0.100000	50	Soft Split Decision Tree	0.63	0.77
0	0.100000	10	Soft Split Decision Tree	0.55	0.74
10	0.200000	100	Soft Split Decision Tree	0.43	0.74
8	0.200000	50	Soft Split Decision Tree	0.42	0.73
6	0.200000	10	Soft Split Decision Tree	0.41	0.65
12	0.300000	10	Soft Split Decision Tree	0.34	0.59
16	0.300000	100	Soft Split Decision Tree	0.34	0.69
14	0.300000	50	Soft Split Decision Tree	0.33	0.66
18	0.400000	10	Soft Split Decision Tree	0.31	0.53
22	0.400000	100	Soft Split Decision Tree	0.31	0.61
20	0.400000	50	Soft Split Decision Tree	0.31	0.58

bank Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.87	0.83
2	0.100000	50	Soft Split Decision Tree	0.87	0.81
0	0.100000	10	Soft Split Decision Tree	0.84	0.74
10	0.200000	100	Soft Split Decision Tree	0.84	0.77
8	0.200000	50	Soft Split Decision Tree	0.81	0.73
16	0.300000	100	Soft Split Decision Tree	0.79	0.66
22	0.400000	100	Soft Split Decision Tree	0.78	0.57
14	0.300000	50	Soft Split Decision Tree	0.76	0.63
6	0.200000	10	Soft Split Decision Tree	0.76	0.64
20	0.400000	50	Soft Split Decision Tree	0.74	0.55
12	0.300000	10	Soft Split Decision Tree	0.70	0.56
18	0.400000	10	Soft Split Decision Tree	0.69	0.52

adult_income Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.47	0.71
2	0.100000	50	Soft Split Decision Tree	0.46	0.69
10	0.200000	100	Soft Split Decision Tree	0.45	0.68
0	0.100000	10	Soft Split Decision Tree	0.44	0.64
8	0.200000	50	Soft Split Decision Tree	0.43	0.66
6	0.200000	10	Soft Split Decision Tree	0.39	0.60
16	0.300000	100	Soft Split Decision Tree	0.36	0.64
14	0.300000	50	Soft Split Decision Tree	0.35	0.61
12	0.300000	10	Soft Split Decision Tree	0.33	0.56
18	0.400000	10	Soft Split Decision Tree	0.29	0.53
20	0.400000	50	Soft Split Decision Tree	0.27	0.56
22	0.400000	100	Soft Split Decision Tree	0.26	0.58

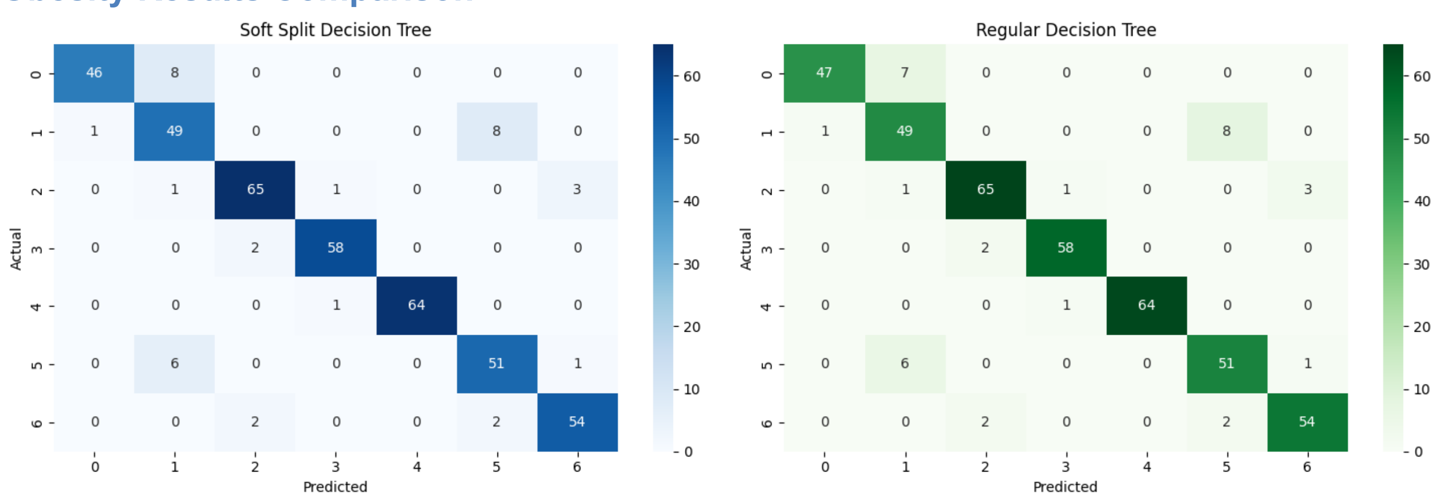
obesity Dataset

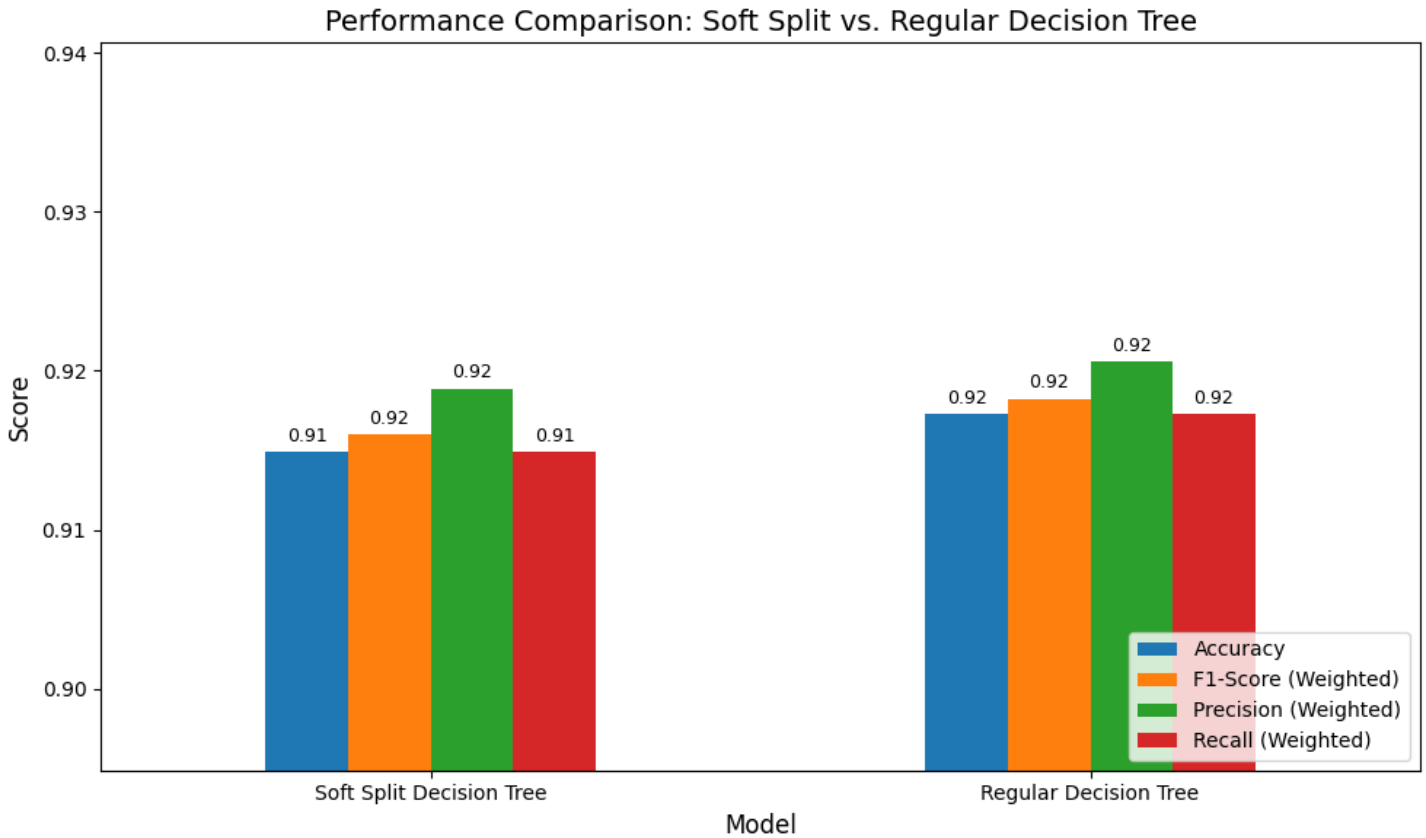
	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.92	0.98
2	0.100000	50	Soft Split Decision Tree	0.91	0.98
0	0.100000	10	Soft Split Decision Tree	0.84	0.96
10	0.200000	100	Soft Split Decision Tree	0.76	0.97
8	0.200000	50	Soft Split Decision Tree	0.74	0.96
6	0.200000	10	Soft Split Decision Tree	0.58	0.89
16	0.300000	100	Soft Split Decision Tree	0.52	0.91
14	0.300000	50	Soft Split Decision Tree	0.48	0.88
12	0.300000	10	Soft Split Decision Tree	0.38	0.77
22	0.400000	100	Soft Split Decision Tree	0.29	0.80
20	0.400000	50	Soft Split Decision Tree	0.28	0.74
18	0.400000	10	Soft Split Decision Tree	0.23	0.63

wine_quality Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
2	0.100000	50	Soft Split Decision Tree	0.59	0.71
4	0.100000	100	Soft Split Decision Tree	0.59	0.72
0	0.100000	10	Soft Split Decision Tree	0.55	0.67
10	0.200000	100	Soft Split Decision Tree	0.54	0.73
8	0.200000	50	Soft Split Decision Tree	0.53	0.68
6	0.200000	10	Soft Split Decision Tree	0.49	0.61
16	0.300000	100	Soft Split Decision Tree	0.48	0.67
14	0.300000	50	Soft Split Decision Tree	0.46	0.65
20	0.400000	50	Soft Split Decision Tree	0.44	0.58
12	0.300000	10	Soft Split Decision Tree	0.44	0.57
22	0.400000	100	Soft Split Decision Tree	0.43	0.59
18	0.400000	10	Soft Split Decision Tree	0.40	0.53

Obesity Results Comparison





Results Table – Task 1

- Note – for each dataset, I showed here the best result we got (per alpha / n_simulations).

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Obesity	Original DT	-	-	0.92	0.99
Obesity	Soft DT	0.1	100	0.92	0.98

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Adult Income	Original DT			0.89	0.92
Adult Income	Soft DT	0.1	100	0.47	0.71

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Wine Quality	Original DT			0.6	0.61
Wine Quality	Soft DT	0.1	100	0.59	0.72

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Bank Campaign	Original DT			0.87	0.7
Bank Campaign	Soft DT	0.1	100	0.87	0.83

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Student Success	Original DT			0.94	0.95
Student Success	Soft DT	0.1	100	0.65	0.78

Programming Task 2: Regression with Soft Splits

2.1 Description of the Adaptation from Classification to Regression

The transition from the SoftSplitDecisionTreeClassifier to the SoftSplitDecisionTreeRegressor involves the following key adaptations:

1. Prediction Objective

- Classification: The predict_proba method returns class probabilities. Each leaf node contains class counts, which are normalized to compute probabilities for each class.
- Regression: The predict method returns a continuous numerical value. Each leaf node contains the average of the target values for the samples that fall into that leaf.

Change Made:

- Replaced the _predict_sample_proba method, which computed class probabilities, with _predict_sample, which computes the regression value (the mean of the target values in the leaf).

2. Leaf Node Representation

- Classification: Leaf nodes store the count of samples per class.
- Regression: Leaf nodes store the mean target value.

Change Made:

- In _predict_sample, the returned value is directly accessed as the first value of the leaf node (tree.value[node][0, 0]), representing the mean target value for regression.

3. Soft Split Logic

- The logic of soft splits remains identical for both classification and regression. A sample is probabilistically routed to the left or right child node based on the alpha parameter.

No Changes Needed:

- The _soft_split method is reused without modifications.

4. Aggregation of Predictions

- Classification: Aggregates multiple simulations to compute the average class probabilities.
- Regression: Aggregates multiple simulations to compute the mean predicted value for regression.

Change Made:

- In the predict method, predictions from multiple simulations are averaged to return the final regression value.

2.2 Results and Analysis

worker_productivity Dataset

	Alpha	n_simulations	Model	Mean MSE	Mean RMSE	Mean MAE	Mean Max Error	Mean R2
10	0.200000	100	Soft Split Decision Tree	0.03096	0.17563	0.14330	0.45249	0.00529
8	0.200000	50	Soft Split Decision Tree	0.03170	0.17775	0.14435	0.46915	-0.01940
4	0.100000	100	Soft Split Decision Tree	0.03273	0.18024	0.13507	0.50325	-0.05300
2	0.100000	50	Soft Split Decision Tree	0.03325	0.18166	0.13545	0.53100	-0.06949
16	0.300000	100	Soft Split Decision Tree	0.03448	0.18514	0.15652	0.45282	-0.10821
14	0.300000	50	Soft Split Decision Tree	0.03463	0.18555	0.15614	0.46570	-0.11305
6	0.200000	10	Soft Split Decision Tree	0.03468	0.18584	0.14879	0.50413	-0.11650
0	0.100000	10	Soft Split Decision Tree	0.03531	0.18746	0.14052	0.56247	-0.13561
22	0.400000	100	Soft Split Decision Tree	0.03764	0.19326	0.16488	0.45160	-0.20794
12	0.300000	10	Soft Split Decision Tree	0.03816	0.19483	0.16163	0.49693	-0.22573
20	0.400000	50	Soft Split Decision Tree	0.03896	0.19664	0.16740	0.46027	-0.25098
18	0.400000	10	Soft Split Decision Tree	0.04262	0.20581	0.17056	0.50433	-0.37214

air_quality Dataset

	Alpha	n_simulations	Model	Mean MSE	Mean RMSE	Mean MAE	Mean Max Error	Mean R2
4	0.100000	100	Soft Split Decision Tree	11292.37677	105.90131	76.42274	1018.50400	0.40478
2	0.100000	50	Soft Split Decision Tree	11530.88374	107.00077	76.48858	970.89600	0.39219
0	0.100000	10	Soft Split Decision Tree	13975.64402	117.77602	76.71353	1051.39000	0.26349
10	0.200000	100	Soft Split Decision Tree	19586.17294	139.35078	118.98461	963.38200	-0.03279
8	0.200000	50	Soft Split Decision Tree	20033.62359	140.96075	118.64004	985.26800	-0.05657
6	0.200000	10	Soft Split Decision Tree	23807.30643	153.63685	118.98299	1028.02000	-0.25498
16	0.300000	100	Soft Split Decision Tree	27941.41817	166.51688	150.40832	969.46500	-0.47358
14	0.300000	50	Soft Split Decision Tree	28453.11933	168.05393	149.86390	977.37800	-0.50013
12	0.300000	10	Soft Split Decision Tree	33928.43295	183.46640	151.00458	1018.78000	-0.78893
22	0.400000	100	Soft Split Decision Tree	35670.52802	187.99833	173.67822	1020.30800	-0.88048
20	0.400000	50	Soft Split Decision Tree	36339.29497	189.80575	173.32093	1023.26800	-0.91605
18	0.400000	10	Soft Split Decision Tree	43047.56307	206.61863	175.28831	1013.22000	-1.27201

apartment_rent Dataset

	Alpha	n_simulations	Model	Mean MSE	Mean RMSE	Mean MAE	Mean Max Error	Mean R2
4	0.100000	100	Soft Split Decision Tree	13257115.01821	3288.31378	2733.54914	21993.68200	-12.26612
2	0.100000	50	Soft Split Decision Tree	14102401.49645	3400.85004	2738.85201	22722.91600	-13.12915
0	0.100000	10	Soft Split Decision Tree	20926720.57977	4164.50147	2776.08702	28395.32000	-20.19945
10	0.200000	100	Soft Split Decision Tree	36617684.27035	5426.92760	4924.81188	22595.59700	-36.03380
8	0.200000	50	Soft Split Decision Tree	37844331.65649	5541.38197	4913.63726	24195.66000	-37.42466
6	0.200000	10	Soft Split Decision Tree	50275450.55970	6423.67507	4956.82226	33302.19000	-50.25432
16	0.300000	100	Soft Split Decision Tree	61579587.49157	7088.88131	6723.24600	22619.13900	-61.82504
14	0.300000	50	Soft Split Decision Tree	63775274.30977	7229.58623	6744.48877	24508.35200	-64.22950
12	0.300000	10	Soft Split Decision Tree	79204464.08894	8123.96120	6746.36126	33936.82000	-80.47150
22	0.400000	100	Soft Split Decision Tree	85541858.71366	8480.78673	8240.83855	23146.82600	-87.51520
20	0.400000	50	Soft Split Decision Tree	86825456.90823	8564.97688	8202.39934	25045.54800	-89.04647
18	0.400000	10	Soft Split Decision Tree	105444272.99311	9491.93053	8223.31850	34174.27000	-109.03844

bike_sharing Dataset

	Alpha	n_simulations	Model	Mean MSE	Mean RMSE	Mean MAE	Mean Max Error	Mean R2
4	0.100000	100	Soft Split Decision Tree	1039.97462	32.24099	26.20372	190.80700	0.96863
2	0.100000	50	Soft Split Decision Tree	1103.71219	33.21276	26.54223	201.25400	0.96672
0	0.100000	10	Soft Split Decision Tree	1574.41636	39.66791	29.20591	272.16000	0.95250
10	0.200000	100	Soft Split Decision Tree	4959.47435	70.40635	59.66797	313.45700	0.85044
8	0.200000	50	Soft Split Decision Tree	5123.18049	71.55706	60.15528	351.75400	0.84554
6	0.200000	10	Soft Split Decision Tree	6531.38366	80.79113	64.11630	399.23000	0.80299
16	0.300000	100	Soft Split Decision Tree	13991.07520	118.25735	102.69790	438.48200	0.57811
14	0.300000	50	Soft Split Decision Tree	14316.56723	119.61763	103.21873	446.29400	0.56829
12	0.300000	10	Soft Split Decision Tree	16953.14388	130.17793	108.12801	511.67000	0.48887
22	0.400000	100	Soft Split Decision Tree	31229.88434	176.68038	155.77188	529.69300	0.05824
20	0.400000	50	Soft Split Decision Tree	31740.86336	178.11309	156.26867	548.17600	0.04291
18	0.400000	10	Soft Split Decision Tree	35763.05341	189.04700	161.03943	598.04000	-0.07809

energy Dataset

	Alpha	n_simulations	Model	Mean MSE	Mean RMSE	Mean MAE	Mean Max Error	Mean R2
4	0.100000	100	Soft Split Decision Tree	5.61343	2.36918	1.91937	7.23963	0.97331
2	0.100000	50	Soft Split Decision Tree	5.94971	2.43910	1.94325	8.68224	0.97171
0	0.100000	10	Soft Split Decision Tree	8.64637	2.94011	2.14403	14.44509	0.95889
10	0.200000	100	Soft Split Decision Tree	24.51028	4.95068	4.07054	12.82724	0.88346
8	0.200000	50	Soft Split Decision Tree	25.30226	5.03000	4.10270	14.13459	0.87969
6	0.200000	10	Soft Split Decision Tree	31.48418	5.61091	4.38365	21.19700	0.85028
16	0.300000	100	Soft Split Decision Tree	61.59845	7.84824	6.50678	18.18629	0.70713
14	0.300000	50	Soft Split Decision Tree	62.67511	7.91666	6.54524	19.47964	0.70199
12	0.300000	10	Soft Split Decision Tree	71.93152	8.48054	6.86026	26.59504	0.65801
22	0.400000	100	Soft Split Decision Tree	121.73839	11.03323	9.28390	23.31209	0.42118
20	0.400000	50	Soft Split Decision Tree	123.49644	11.11253	9.34658	25.06100	0.41283
18	0.400000	10	Soft Split Decision Tree	137.12957	11.70982	9.74165	32.44695	0.34796

Results Table – Task 2

Dataset	Method	Alpha	N_simulations	MSE	RMSE	MAE	Max Error	R2
Air Quality	Original DT			8917	94	19.51	1235	0.53
Air Quality	Soft DT	0.1	100	11292	105.9	76	1018	0.4

Dataset	Method	Alpha	N_simulations	MSE	RMSE	MAE	Max Error	R2
Apartment	Original DT			740781	682	37.15	21697	0.55
Apartment	Soft DT	0.1	100	13,255,115	3288	2733	21993	-12.27

Dataset	Method	Alpha	N_simulations	MSE	RMSE	MAE	Max Error	R2
Bike Sharing	Original DT			33.62	5.78	2.67	99.7	0.99
Bike Sharing	Soft DT	0.1	100	1039	32.24	26.2	190.8	0.97

Dataset	Method	Alpha	N_simulations	MSE	RMSE	MAE	Max Error	R2
Energy	Original DT			0.0008	0.0091	0.0065	0.06186	0.99
Energy	Soft DT	0.1	100	5.61	2.369	1.9	7.23	0.973

Dataset	Method	Alpha	N_simulations	MSE	RMSE	MAE	Max Error	R2
Worker Productivity	Original DT			0.03	0.175	0.143	0.70	0.0017
Worker Productivity	Soft DT	0.2	100	0.03	0.175	0.143	0.45	0.0053

Programming Task 3: Weighted Prediction

Proposed Method: Improved Soft Splits Using Distance from Uniform Distribution

Description

In this alternative method, I propose weighting the decision tree leaves during prediction based on their distance from a uniform class distribution. The key idea is to adjust the randomness of routing decisions within the tree (via soft splits) by incorporating a measure of uncertainty—specifically, how far the class distribution in a node is from being uniform.

Nodes with a high distance from a uniform distribution are considered more certain (i.e., dominated by a specific class), while nodes closer to uniformity indicate greater uncertainty. This additional information is used to dynamically adjust the split probabilities, promoting smoother decision boundaries and reducing overfitting.

Theoretical Justification

1. Soft Splits and Uncertainty:
   - Soft splits introduce stochasticity in routing decisions, which helps to avoid overfitting by reducing deterministic biases in individual splits.
   - Incorporating the distance from uniformity ensures that splits are guided by the reliability of the class distribution in the node.
2. Distance from Uniformity:
   - A uniform distribution indicates maximal uncertainty in class assignments.
   - By penalizing nodes with higher uncertainty (closer to uniformity), I encourage more confident predictions in later stages of the tree.
3. KL Divergence as a Measure:
   - KL divergence quantifies how much a node's class distribution diverges from a uniform distribution: DKL(P||U)=∑i=1nP(i)log(P(i)U(i))D_{\text{KL}}(P || U) = \sum_{i=1}^{n} P(i) \log\left(\frac{P(i)}{U(i)}\right)DKL(P||U)=i=1∑nP(i)log(U(i)P(i)) where PPP is the observed class probability, and UUU is the uniform distribution.
4. Regularization via Adjusted Alpha:
   - The split probability α\alphaα is adjusted based on the KL divergence: $\alpha\ adjusted = \frac{\alpha}{\alpha + D_{KL}}$
   - This ensures that nodes with uncertain distributions are less likely to make confident split decisions.

Implementation Overview

1. Calculate Distance from Uniformity:
   - Compute the KL divergence between the node’s class distribution and a uniform distribution.
2. Adjust Alpha Dynamically:
   - Modify the split probability α\alphaα based on the calculated KL divergence.
3. Soft Split Decision:
   - Use the adjusted α\alphaα to probabilistically route samples during inference.
4. Simulation for Robust Predictions:
   - Perform multiple routing simulations for each sample and average the predicted probabilities to smooth the results.

Steps I Took

1. **Enhanced Split Logic:**
   - The _soft_split function was updated to incorporate the distance from uniformity via KL divergence.
   - This function adjusts the probability of going left or right based on the node’s uncertainty.
2. **Dynamic Weighting:**
   - The _distance_from_uniform function calculates KL divergence to inform the split probability adjustments.
   - Nodes with high KL divergence (low uncertainty) are weighted more heavily in routing decisions.
3. **Stochastic Predictions:**
   - The predict_proba method runs multiple simulations for each sample.
   - The averaged probabilities from all simulations ensure robustness against noise and overfitting.

Advantages

- Reduced Overfitting:
    - By dynamically penalizing uncertain splits, this approach discourages overconfident routing in noisy or ambiguous regions of the tree.
- Improved Generalization:
    - Promotes smoother decision boundaries by incorporating uncertainty into routing.
- Robustness:
    - Averaging over multiple simulations ensures stability in predictions.

Disadvantages

- Increased Computational Cost:
    - Multiple simulations during inference increase computational complexity compared to standard decision trees.
- Sensitivity to Hyperparameters:
    - The effectiveness of the method depends on the choice of α and the number of simulations

3.2 Results and Comparison

wine_quality_improved Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.60	0.71
2	0.100000	50	Soft Split Decision Tree	0.59	0.72
0	0.100000	10	Soft Split Decision Tree	0.59	0.65
10	0.200000	100	Soft Split Decision Tree	0.59	0.72
8	0.200000	50	Soft Split Decision Tree	0.59	0.72
16	0.300000	100	Soft Split Decision Tree	0.57	0.74
6	0.200000	10	Soft Split Decision Tree	0.55	0.65
14	0.300000	50	Soft Split Decision Tree	0.54	0.72
22	0.400000	100	Soft Split Decision Tree	0.52	0.72
20	0.400000	50	Soft Split Decision Tree	0.52	0.69
12	0.300000	10	Soft Split Decision Tree	0.50	0.63
18	0.400000	10	Soft Split Decision Tree	0.48	0.63

bank_improved Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.88	0.86
2	0.100000	50	Soft Split Decision Tree	0.88	0.84
10	0.200000	100	Soft Split Decision Tree	0.87	0.83
0	0.100000	10	Soft Split Decision Tree	0.86	0.78
8	0.200000	50	Soft Split Decision Tree	0.85	0.80
16	0.300000	100	Soft Split Decision Tree	0.84	0.76
22	0.400000	100	Soft Split Decision Tree	0.82	0.67
14	0.300000	50	Soft Split Decision Tree	0.82	0.72
6	0.200000	10	Soft Split Decision Tree	0.81	0.70
20	0.400000	50	Soft Split Decision Tree	0.79	0.64
12	0.300000	10	Soft Split Decision Tree	0.76	0.62
18	0.400000	10	Soft Split Decision Tree	0.72	0.56

obesity_improved Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.93	0.98
2	0.100000	50	Soft Split Decision Tree	0.93	0.98
10	0.200000	100	Soft Split Decision Tree	0.92	0.98
0	0.100000	10	Soft Split Decision Tree	0.91	0.97
8	0.200000	50	Soft Split Decision Tree	0.91	0.98
6	0.200000	10	Soft Split Decision Tree	0.80	0.96
16	0.300000	100	Soft Split Decision Tree	0.77	0.97
14	0.300000	50	Soft Split Decision Tree	0.75	0.96
12	0.300000	10	Soft Split Decision Tree	0.61	0.90
22	0.400000	100	Soft Split Decision Tree	0.54	0.93
20	0.400000	50	Soft Split Decision Tree	0.53	0.91
18	0.400000	10	Soft Split Decision Tree	0.43	0.80

student_success_improved Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
4	0.100000	100	Soft Split Decision Tree	0.68	0.80
2	0.100000	50	Soft Split Decision Tree	0.68	0.79
0	0.100000	10	Soft Split Decision Tree	0.64	0.77
10	0.200000	100	Soft Split Decision Tree	0.63	0.79
8	0.200000	50	Soft Split Decision Tree	0.61	0.77
6	0.200000	10	Soft Split Decision Tree	0.53	0.72
16	0.300000	100	Soft Split Decision Tree	0.50	0.75
14	0.300000	50	Soft Split Decision Tree	0.47	0.73
12	0.300000	10	Soft Split Decision Tree	0.43	0.66
18	0.400000	10	Soft Split Decision Tree	0.37	0.60
22	0.400000	100	Soft Split Decision Tree	0.37	0.70
20	0.400000	50	Soft Split Decision Tree	0.37	0.66

adult_income_improved Dataset

	Alpha	n_simulations	Model	Mean Accuracy	Mean AUC
10	0.200000	100	Soft Split Decision Tree	0.48	0.71
4	0.100000	100	Soft Split Decision Tree	0.47	0.71
2	0.100000	50	Soft Split Decision Tree	0.47	0.70
8	0.200000	50	Soft Split Decision Tree	0.47	0.69
0	0.100000	10	Soft Split Decision Tree	0.46	0.65
16	0.300000	100	Soft Split Decision Tree	0.45	0.68
14	0.300000	50	Soft Split Decision Tree	0.43	0.66
6	0.200000	10	Soft Split Decision Tree	0.43	0.62
12	0.300000	10	Soft Split Decision Tree	0.38	0.59
22	0.400000	100	Soft Split Decision Tree	0.35	0.64
20	0.400000	50	Soft Split Decision Tree	0.34	0.62
18	0.400000	10	Soft Split Decision Tree	0.33	0.56

Results Table – Task 3

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Obesity	Original DT	-	-	0.928	0.957
Obesity	Improved Soft DT	0.1	100	0.928	0.980

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Adult Income	Original DT			0.442	0.575
Adult Income	Improved Soft DT			0.483	0.707

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Wine Quality	Original DT			0.596	0.614
Wine Quality	Improved Soft DT			0.595	0.717

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Bank Campaign	Original DT			0.871	0.696
Bank Campaign	Improved Soft DT			0.875	0.855

Dataset	Method	Alpha	N_simulations	Accuracy	AUC
Student Success	Original DT			0.679	0.726
Student Success	Improved Soft DT			0.681	0.797