Data Mining Final Report

Student	ID
Joud Bakarman	443201057
Lamees Alghamdi	444201177
Maha Alruwais	444200749
Deem Aljarba	444200523
Norah Almadhi	444200890

Problem

Recently, the prevalence of mental health issues, particularly depression, has been increasing, becoming more common across diverse populations. This trend leads to numerous serious implications for individuals' personal, social, and professional lives, potentially resulting in long-term adverse outcomes if left unaddressed. In our project, we aim to study and analyze patient data related to depression, which will greatly assist in identifying possible factors and risks associated with this condition. By predicting the likelihood of depression, we can help many individuals seek timely interventions and take measures to improve their mental well-being.

Data Mining Task

In our project, we will employ two data mining tasks to help predict the likelihood of depression: classification and clustering.

For classification, we will train our model to predict whether a patient is depressed or not, based on various factors such as age, marital status, number of children, education level, income sources, and others. The class attribute for this task will be "depressed," where 1 indicates the patient is depressed, and 0 indicates they are not.

As for clustering, our model will create groups of patients who share similar characteristics without considering whether they are depressed or not. These groups will help identify patterns and similarities in the data, potentially leading to a deeper understanding of the factors influencing depression and uncovering new insights into mental health risks.

Data

The source: https://www.kaggle.com/datasets/diegobabativa/depression?resource=download

Number of object: 1429

Number of attribute: 23

Atributes description

Attribute	Data type	Possible values
Survey_id	Nominal	1 to 1429 (unique ID for each survey)
Ville_id	Nominal	1 to 292 (unique ID for each city)
sex	Binary (Symmetric)	0 (Female), 1 (Male)
Age	Numeric (Ratio)	17 to 91 (Age of the individual)
Married	Binary (Asymmetric)	0 (Not married), 1 (Married)
Number_children	Numeric (Ratio)	0 to 11 (Number of children)
education_level	Ordinal	1 to 19 (Education levels from basic to university)
total_members	Numeric (Ratio)	1 to 12 (Total number of household members)
gained_asset	Numeric (Ratio)	325,112 to 99,127,550 (Value of gained assets)
durable_asset	Numeric (Ratio)	162,556 to 99,615,600 (Value of durable assets)
save_asset	Numeric (Ratio)	0 to 79,366,717 (Value of saved assets)
living_expenses	Numeric (Ratio)	0 to 35,491,530 (Living expenses)
other_expenses	Numeric (Ratio)	0 to 10,400,000 (Other expenses)
incoming_salary	Binary (Asymmetric)	0 (No income), 1 (Income)
incoming_own_farm	Binary (Asymmetric)	0 (No income), 1 (Income from farm)
incoming_business	Binary (Asymmetric)	0 (No income), 1 (Income from business)
incoming_no_business	Binary (Asymmetric)	0 (No income), 1 (No income from business)
incoming_agricultural	Numeric (Ratio)	325,112 to 99,127,550 (Income from agriculture)
farm_expenses	Numeric (Ratio)	0 to 99,651,190 (Farm expenses)
labor_primary	Binary (Asymmetric)	0 (No), 1 (Yes)
lasting_investment	Numeric (Ratio)	0 to 99,616,000 (Value of lasting investments)
no_lasting_investmen	Numeric (Ratio)	0 to 69,219,765 (Value of no lasting investments)
depressed	Binary (Asymmetric)	0 (Not depressed), 1 (Depressed)

Missing values

```
Survey_id                0
Ville_id                 0
sex                      0
Age                      0
Married                  0
Number_children          0
education_level          0
total_members            0
gained_asset             0
durable_asset            0
save_asset               0
living_expenses          0
other_expenses           0
incoming_salary          0
incoming_own_farm        0
incoming_business        0
incoming_no_business     0
incoming_agricultural    0
farm_expenses            0
labor_primary            0
lasting_investment       0
no_lasting_investmen    10
depressed                0
dtype: int64
```

We have 4 missing value in only one attribute(no_lasting_investmen)

Statistical Measures

	Age	Number_children	total_members	gained_asset	durable_asset	save_asset	living_expenses	other_expenses	incoming_agricultural	farm_expenses	lasting_investment	no_lasting_investmen
count	550.000000	550.000000	550.000000	5.500000e+02	5.500000e+02	5.500000e+02	5.500000e+02	5.500000e+02	5.500000e+02	5.500000e+02	5.500000e+02	5.400000e+02
mean	35.470909	2.916364	5.030909	3.390004e+07	2.700609e+07	2.673890e+07	3.173706e+07	3.420073e+07	3.322826e+07	3.495804e+07	3.249485e+07	3.351874e+07
std	14.423786	1.922048	1.762013	1.999274e+07	1.784842e+07	1.659822e+07	2.140596e+07	2.258004e+07	1.964630e+07	2.043604e+07	2.119582e+07	2.070478e+07
min	17.000000	0.000000	1.000000	3.251120e+05	1.729660e+05	1.242339e+06	5.015480e+05	1.729660e+05	1.040999e+06	1.096608e+06	7.429200e+04	1.263120e+05
25%	25.000000	2.000000	4.000000	2.478189e+07	1.956678e+07	2.339998e+07	1.985906e+07	2.102017e+07	2.202113e+07	2.268844e+07	2.001914e+07	2.307214e+07
50%	31.000000	3.000000	5.000000	2.891220e+07	2.286194e+07	2.339998e+07	2.669228e+07	2.820307e+07	3.002882e+07	3.136343e+07	2.841172e+07	2.829271e+07
75%	42.000000	4.000000	6.000000	3.647977e+07	2.481582e+07	2.339998e+07	3.273141e+07	4.244073e+07	3.716900e+07	4.081695e+07	3.707504e+07	4.073354e+07
max	87.000000	11.000000	12.000000	9.844437e+07	9.961560e+07	9.960110e+07	9.609222e+07	9.929529e+07	9.978910e+07	9.828510e+07	9.944667e+07	9.675953e+07

Outliers

```
Outliers in 'Age':
[84 74 70 73 81 80 69 78 73 81 67 87 81 65 80 66 67 67 73 71 70 77 66 65
 65 72 72 86 82 76 73]

Outliers in 'Number_children':
[ 7  7  7  9  7  7  7  7  7  7  8  7 11  8  7  8  7  7  8  7  7  7  7  8
  7  8  7]

Outliers in 'total_members':
[ 1  9  9  1  9 12  9  9  1  9  9 10  1  9  1  1 10  1  1  1  1  1  1  9
  9  1 11 10  9  1  9 10 11 10 10  1 10  9  1 10 10 10]

Outliers in 'gained_asset':
[93596368 82606287 86736603 82606287 82606287 82606293 82606287 82606287
 86736603 82606287 82606287 92590485 82606287 82606287 82606287 83646683
 88885307 96092224 82606287 82606287 98444366 87443924 82606287 82606287
 82606287 92088379 83646683 81678391 75386055 82606287 81923103 84329857
 88460167 82606287 82606287 83646683 96143182 82606287 75386055 82606287
 82606287 86736603 87776993 74353119 82606287 94314049 83646683 82606287]

Outliers in 'durable_asset':
[83440079 86162689 87283768 75752698 91287605 68225479 72069168 69666862
 81678391 68545784 96092216 96092216 97853912 75592545 83279922 92088379
 97693758 96092216 80076847 80076851 81678391 96092216 63260712 80877617
 80076847 66463783 72869934 70467628 84881462 80076849 99455444 73190239
 99615601 75432396]

Outliers in 'save_asset':
[80076847 80076851 80076847 80076847 80076851 96092224 97693756 80076849
 96092224 96092224 64061478 96092224 80076851 80076847 83295555 79841393
 94983925 99601105 95600319 80076847 80076847 80076847 80076847 80076851
 86483002 80076847 80076847 72069163 96092224 80076847 96092224 90948227
 80076847]
```

```
Outliers in 'living_expenses':
[93422985 84080696 93422991 93422991 80076849 93422991 84080696 87417221
 74738398 90353384 90753761 94223757 88191299 85415304 93422995 86082611
 80076849 96092218 76339927 93422991 78742236 83413382 93422991 91020679
 77407622 75405703 88084536 76073008 93422991 80076849 93422991 76740313
 80076849 93422985 93422991 80076849 89419144 77007241 80877619 93422994
 80076847]

Outliers in 'other_expenses':
[88084536 81838539 85522079 93689919 95739883 90486839 95131302 97693758
 94170372 88404846 88645073 98814831 96092224 87283764 88084536 94490685
 79996773 90006378 96092218 89686069 81037773 83279924 80076847 93369606
 83279924 94490681 83440079 96092218 85041618 89686069 80076847 80076847
 96892986 92889147 84881458 89686069 89686069 79756546 98974991 96092224
 88084534 81678391 99295292 83279924]

Outliers in 'incoming_agricultural':
[86749923 74738393 93422985 96092224 74738393 80076849 74738393 93422991
 94624138 82746086 93422991 82479153 79609737 90353384 99789095 94223757
 85415304 77607818 80076847 72869935 80076847 98761454 74738393 88084536
 77407622 86749926 80076847 76873775 84080696 80076847 82479153 85415304
 93422991 85415304 80076847 93422991 80076847 78742236]

Outliers in 'farm_expenses':
[77808008 79187107 91821451 87584057 88796329 88084537 78964669 88974275
 80076847 97426834 80076847 77852494 96092218 88974275 91643506 98285103
 88529406 78297365 95647341 82857299 84525566 84525573 84525561 80076851
 88662863 80521727 97382345 93422991 86749923 80877619 82879543]

Outliers in 'lasting_investment':
[87443916 89686073 79597247 75828381 88982855 87777887 98875537 94307068
 82606287 76953865 84687622 87916122 97350049 96092216 96092216 98519951
 89370978 89686073 76705368 96092224 86779846 80728137 78290515 94195801
 82682971 81443433 99446667 80582733 98555206 92843073 80076847 87115356
 79402557 95670868 90039685 92661629 85998456 83630414 75428271]
```

Boxplot

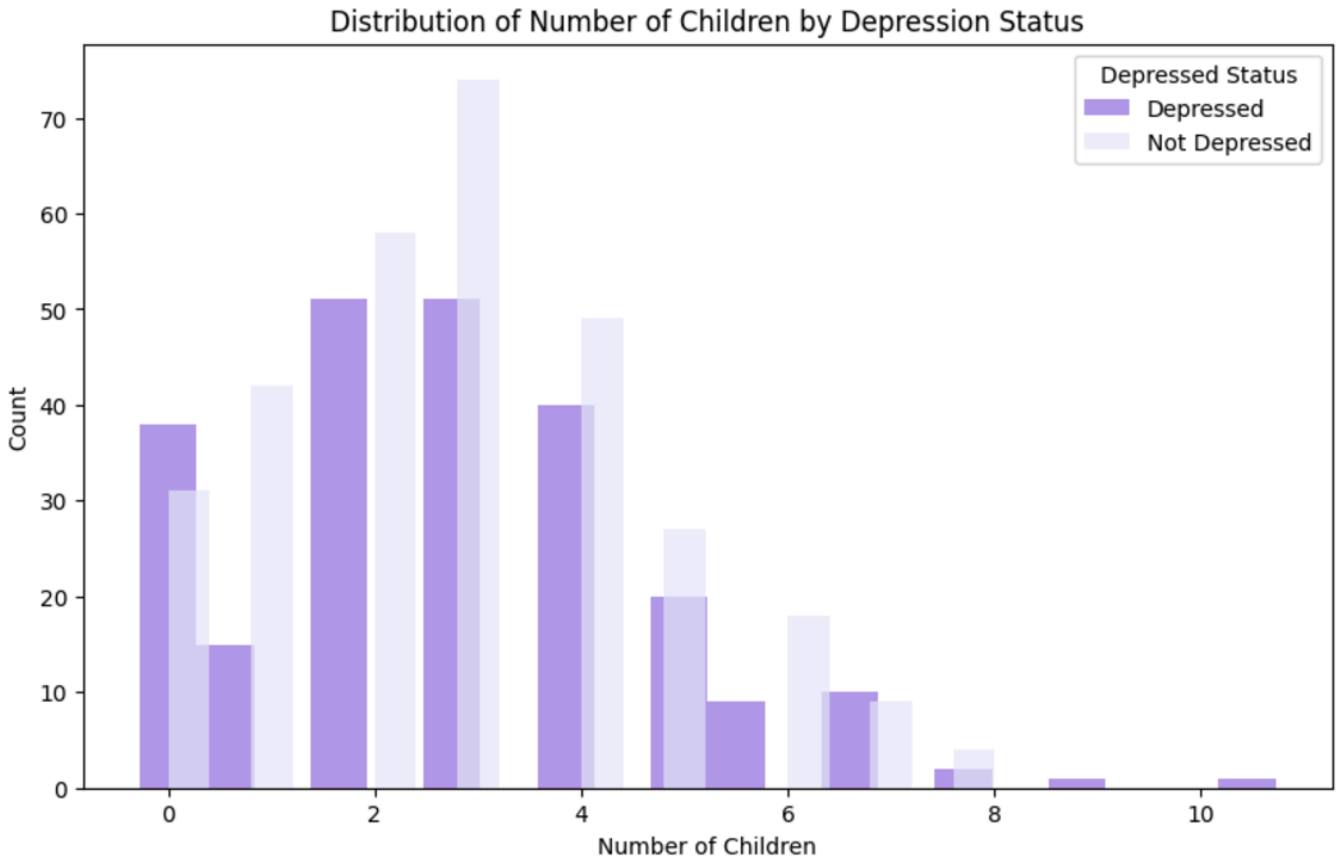
Graph

Description

Using a boxplot We observed the distribution of education levels for two groups: "Not Depressed" (label 0) and "Depressed" (label 1). The y-axis represents education level, and the x-axis shows the class labels. The median education level is slightly higher in the "Depressed" group, though both groups have a similar spread. There are more outliers in the "Not Depressed" group at the lower end of the education level

Plotting Methods

Graph

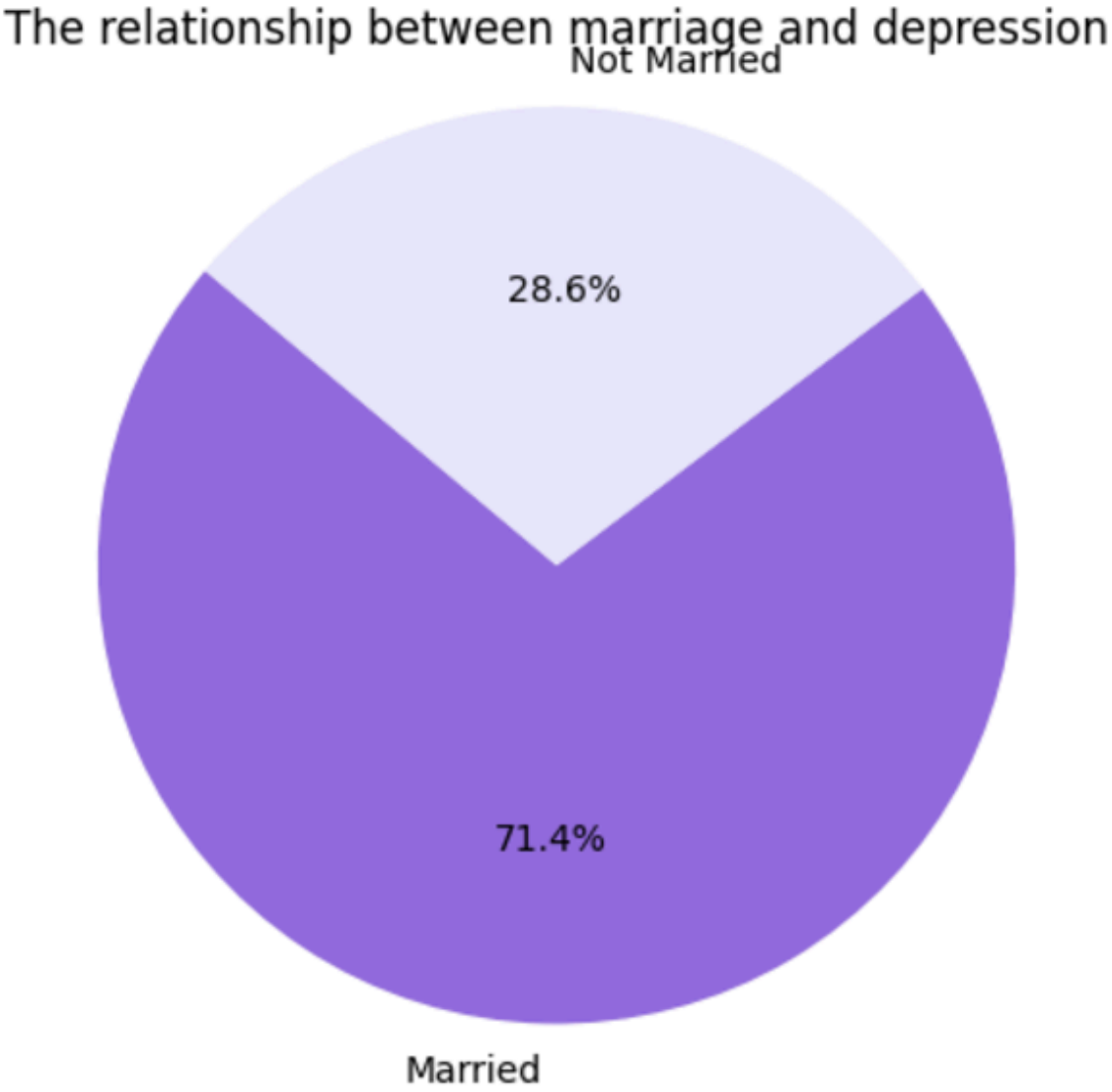


Description

The bar chart shows the relationship between the number of children and depression status. It highlights that people with no children or two children appear most often in both the "Depressed" and "Not Depressed" groups. However, as the number of children increases beyond two, more individuals fall into the "Not Depressed" category. More analysis is needed to understand this relationship fully.

The number of children might be an important feature to consider during model training, as it shows a possible relationship with depression. It could be beneficial to treat this variable appropriately in the model.

Graph

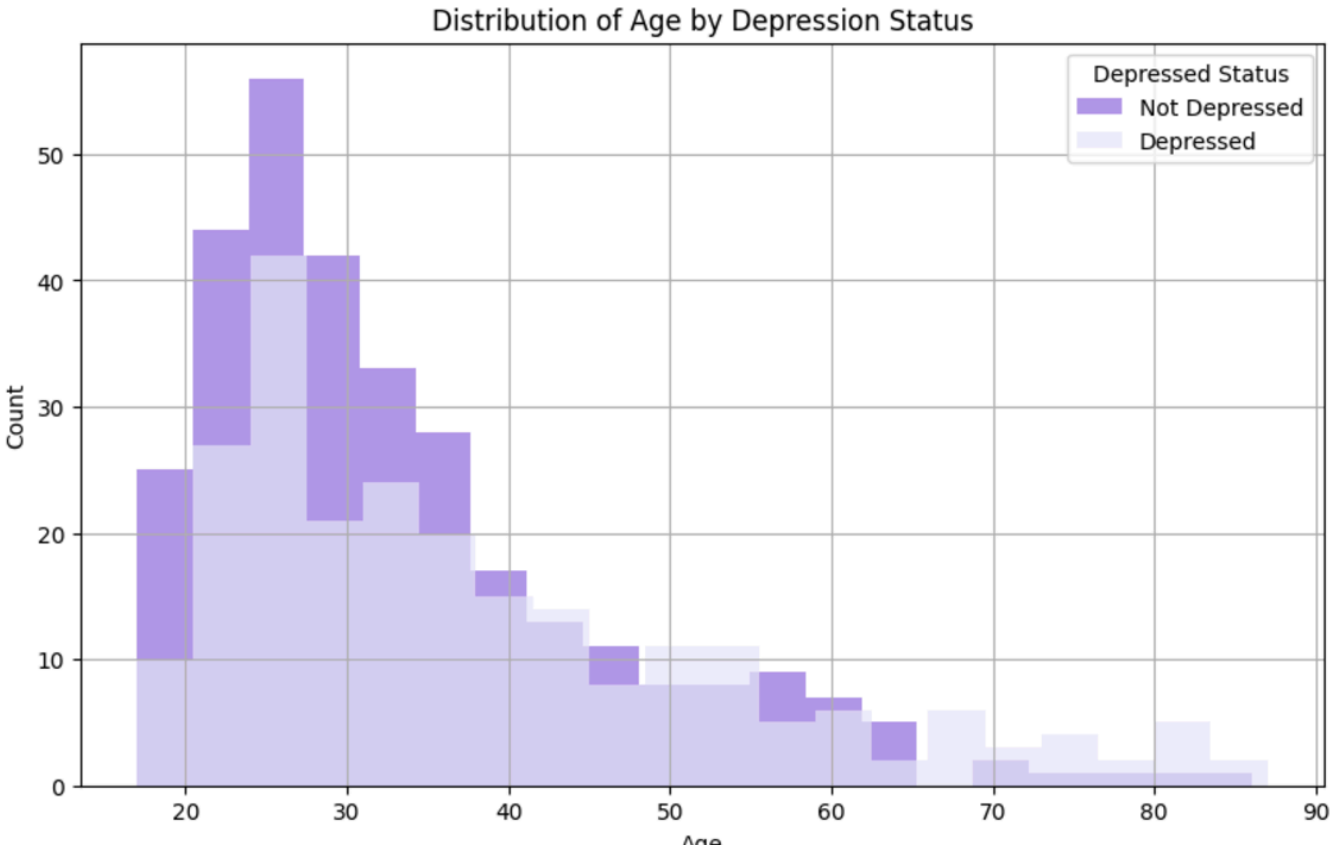


Description

The chart shows a high percentage of depressed individuals being married, which suggests that marital status could have a significant relationship with depression. This information may indicate that marital status could be an important factor when studying depression in this dataset.

Imbalance in marital status: The chart shows that the majority of depressed individuals are married, which could indicate a potential imbalance in the data. This could suggest that the marital status feature may have a strong influence on depression and should be carefully considered during data preparation.

graph



Description

The bar chart shows the distribution of age by depression status. In this chart, the "Not Depressed" group is more prevalent across all age ranges, with the highest concentrations appearing in individuals in their 20s and early 30s. In contrast, the "Depressed" group has fewer individuals across the entire age spectrum, though it tends to follow a similar pattern to the "Not Depressed" group, with a higher count in the 20s age range. Further analysis is required to understand the underlying factors contributing to the distribution of depression status across different ages

Since age plays a role in depression status (with a concentration in the 20s for both groups), this feature might need to be treated with special attention in the modeling phase, especially in algorithms where age can significantly influence outcomes.

Data Pre-processing

Data Sampling

We sampled from the non-depressed group to match the number of depressed cases. This step helped prevent biases in our analysis, ensuring that both groups were equally represented for fair and accurate comparisons. Now we have 550 objects in the sample instead of 1429 in the original dataset.

	Survey_id	Ville_id	sex	Age	Married	Number_children	education_level	total_members	gained_asset	durable_asset	...	i
0	747	57	1	23	1	3	8	5	28912201	22861940	...	
1	849	130	0	34	0	1	9	3	41303144	21925041	...	
2	540	52	1	84	0	0	1	5	28912201	22861940	...	
3	603	100	1	56	1	0	12	2	93596368	21140288	...	
4	1001	207	1	40	0	0	7	5	28912201	22861940	...	
...	...	...	...	...	...	...	...	...	...	...	...	
545	979	120	0	28	1	0	9	5	28912201	22861940	...	
546	1012	65	1	36	0	4	10	6	65663017	22861940	...	
547	661	60	1	35	1	3	10	5	10325787	53411261	...	
548	321	49	1	48	1	4	9	5	28912201	22861940	...	
549	1368	85	1	39	1	7	7	5	28912201	22861940	...	

550 rows × 23 columns

Data Cleaning

Checking for missing values:

We detected 10 Missing values in the numerical column 'no_lasting_investmen', and they were filled using the column's mean value .This method preserves data integrity and ensures that valuable information isn't lost.

```
Missing values in each column:
Survey_id                0
Ville_id                 0
sex                      0
Age                      0
Married                  0
Number_children          0
education_level          0
total_members            0
gained_asset             0
durable_asset            0
save_asset               0
living_expenses          0
other_expenses           0
incoming_salary          0
incoming_own_farm        0
incoming_business        0
incoming_no_business     0
incoming_agricultural    0
farm_expenses            0
labor_primary            0
lasting_investment       0
no_lasting_investmen    10
depressed                0
dtype: int64
```

Handling the missing values using the mean:

```
Missing values per column:
Survey_id                0
Ville_id                 0
sex                      0
Age                      0
Married                  0
Number_children          0
education_level          0
total_members            0
gained_asset             0
durable_asset            0
save_asset               0
living_expenses          0
other_expenses           0
incoming_salary          0
incoming_own_farm        0
incoming_business        0
incoming_no_business     0
incoming_agricultural    0
farm_expenses            0
labor_primary            0
lasting_investment       0
no_lasting_investmen     0
depressed                0
dtype: int64
```

Detecting Outliers:

We calculated the Z-scores for numerical columns to detect outliers. A threshold (set at 2) was applied, and values with Z-scores exceeding this threshold were identified as potential outliers.

Outliers in each column:

```
Outliers in 'Age':
[84 74 70 73 81 80 69 78 73 81 67 87 81 65 80 66 67 67 73 71 70 77 66 65
 65 72 72 86 82 76 73]

Outliers in 'Number_children':
[ 7  7  7  9  7  7  7  7  7  7  8  7 11  8  7  8  7  7  8  7  7  7  7  8
  7  8  7]

Outliers in 'total_members':
[ 1  9  9  1  9 12  9  9  1  9  9 10  1  9  1  1 10  1  1  1  1  1  1  9
  9  1 11 10  9  1  9 10 11 10 10  1 10  9  1 10 10 10]

Outliers in 'gained_asset':
[93596368 82606287 86736603 82606287 82606287 82606293 82606287 82606287
 86736603 82606287 82606287 92590485 82606287 82606287 82606287 83646683
 88885307 96092224 82606287 82606287 98444366 87443924 82606287 82606287
 82606287 92088379 83646683 81678391 75386055 82606287 81923103 84329857
 88460167 82606287 82606287 83646683 96143182 82606287 75386055 82606287
 82606287 86736603 87776993 74353119 82606287 94314049 83646683 82606287]

Outliers in 'durable_asset':
[83440079 86162689 87283768 75752698 91287605 68225479 72069168 69666862
 81678391 68545784 96092216 96092216 97853912 75592545 83279922 92088379
 97693758 96092216 80076847 80076851 81678391 96092216 63260712 80877617
 80076847 66463783 72869934 70467628 84881462 80076849 99455444 73190239
 99615601 75432396]

Outliers in 'save_asset':
[80076847 80076851 80076847 80076847 80076851 96092224 97693756 80076849
 96092224 96092224 64061478 96092224 80076851 80076847 83295555 79841393
 94983925 99601105 95600319 80076847 80076847 80076847 80076847 80076851
 86483002 80076847 80076847 72069163 96092224 80076847 96092224 90948227
 80076847]

Outliers in 'living_expenses':
[93422985 84080696 93422991 93422991 80076849 93422991 84080696 87417221
 74738398 90353384 90753761 94223757 88191299 85415304 93422995 86082611
 80076849 96092218 76339927 93422991 78742236 83413382 93422991 91020679
 77407622 75405703 88084536 76073008 93422991 80076849 93422991 76740313
 80076849 93422985 93422991 80076849 89419144 77007241 80877619 93422994
 80076847]

Outliers in 'other_expenses':
[88084536 81838539 85522079 93689919 95739883 90486839 95131302 97693758
 94170372 88404846 88645073 98814831 96092224 87283764 88084536 94490685
 79996773 90006378 96092218 89686069 81037773 83279924 80076847 93369606
 83279924 94490681 83440079 96092218 85041618 89686069 80076847 80076847
 96892986 92889147 84881458 89686069 89686069 79756546 98974991 96092224
 88084534 81678391 99295292 83279924]

Outliers in 'incoming_agricultural':
[86749923 74738393 93422985 96092224 74738393 80076849 74738393 93422991
 94624138 82746086 93422991 82479153 79609737 90353384 99789095 94223757
 85415304 77607818 80076847 72869935 80076847 98761454 74738393 88084536
 77407622 86749926 80076847 76873775 84080696 80076847 82479153 85415304
 93422991 85415304 80076847 93422991 80076847 78742236]

Outliers in 'farm_expenses':
[77808008 79187107 91821451 87584057 88796329 88084537 78964669 88974275
 80076847 97426834 80076847 77852494 96092218 88974275 91643506 98285103
 88529406 78297365 95647341 82857299 84525566 84525573 84525561 80076851
 88662863 80521727 97382345 93422991 86749923 80877619 82879543]

Outliers in 'lasting_investment':
[87443916 89686073 79597247 75828381 88982855 87777887 98875537 94307068
 82606287 76953865 84687622 87916122 97350049 96092216 96092216 98519951
 89370978 89686073 76705368 96092224 86779846 80728137 78290515 94195801
 82682971 81443433 99446667 80582733 98555206 92843073 80076847 87115356
 79402557 95670868 90039685 92661629 85998456 83630414 75428271]

Outliers in 'no_lasting_investmen':
[]
```

To handle this, for each numeric column, the row with the largest deviation from the mean was removed iteratively, which is shown by the number of rows (526 rows after removing the outliers).

By handling outliers, we reduced the risk of skewed results and improved the robustness of our statistical analyses and models.

```
DataFrame after removing outliers from each column:
     Survey_id  Ville_id  sex  Age  Married  Number_children  education_level  \
0          603       100    1   56        1                0               12
1         1001       207    1   40        0                0                7
2          840       102    1   43        1                4                4
3          309        25    1   51        1                2               12
4          533       101    1   20        1                2                9
..         ...       ...  ...  ...      ...              ...              ...
522        355        34    1   29        1                4                9
523        979       120    0   28        1                0                9
524       1012        65    1   36        0                4               10
525        661        60    1   35        1                3               10
526       1368        85    1   39        1                7                7

     total_members  gained_asset  durable_asset  ...  incoming_salary  \
0                2      93596368       21140288  ...                0
1                5      28912201       22861940  ...                0
2                9      12390944       18978214  ...                1
3                5      28912201       22861940  ...                0
4                5      28912201       22861940  ...                0
..             ...           ...            ...  ...              ...
522              6      40997882       17008324  ...                0
523              5      28912201       22861940  ...                0
524              6      65663017       22861940  ...                0
525              5      10325787       53411261  ...                1
526              5      28912201       22861940  ...                0

     incoming_own_farm  incoming_business  incoming_no_business  \
0                    1                  0                     0
1                    0                  0                     0
2                    0                  0                     1
3                    0                  0                     0
4                    0                  0                     0
..                 ...                ...                   ...
522                  1                  0                     1
523                  0                  0                     0
524                  0                  1                     1
525                  0                  0                     0
526                  0                  0                     0

     incoming_agricultural  farm_expenses  labor_primary  lasting_investment  \
0                 43775349       77808008              0            12402556
1                 30028818       31363432              0            28411718
2                 58322639       24212127              1             2117823
3                 30028818       31363432              0            28411718
4                 30028818       31363432              0            28411718
..                     ...            ...            ...                 ...
522               18017292       64528594              0            64892816
523               30028818       31363432              0            28411718
524               78742236       30584911              0             9364798
525               15815178       19585463              1            75428271
526               30028818       31363432              0            28411718

     no_lasting_investmen  depressed
0              71201668.0          1
1              28292707.0          1
2              29246958.0          1
3              28292707.0          1
4              28292707.0          1
..                    ...        ...
522            67444725.0          0
523            28292707.0          0
524            35089231.0          0
525            37566051.0          0
526            28292707.0          0

[527 rows x 23 columns]
```

Data Transformation

Aggregation

The attributes 'gained_asset', 'durable_asset', and 'save_asset' are aggregated to create the 'total_assets' variable, while 'living_expenses', 'farm_expenses' and 'other_expenses' are combined to form the 'total_expenses' variable. This transformation simplifies the dataset by reducing the number of features and providing a clearer overview of an individual's overall financial status. The resulting aggregated values enhance interpretability and facilitate more efficient analysis within the dataset.

Original columns:

	gained_asset	durable_asset	save_asset	living_expenses	other_expenses	farm_expenses
0	28912201	22861940	23399979	26692283	28203066	31363432
1	41303144	21925041	23399979	66730708	10890451	22243569
2	28912201	22861940	23399979	26692283	28203066	31363432
3	93596368	21140288	5925687	34566505	72469551	77808008
4	28912201	22861940	23399979	26692283	28203066	31363432
...	...	...	...	...	...	...
545	28912201	22861940	23399979	26692283	28203066	31363432
546	65663017	22861940	10070598	501548	83279924	30584911
547	10325787	53411261	32030739	46711493	34192814	19585463
548	28912201	22861940	23399979	26692283	28203066	31363432
549	28912201	22861940	23399979	26692283	28203066	31363432

550 rows × 6 columns

Aggregated Columns:

```
Updated Data with Aggregated Columns:
```

	total_assets	total_expenses
0	120662343	184844064
1	75174120	86258781
2	111446005	66439318
3	75174120	86258781
4	75174120	86258781
...	...	...
522	72420039	132500494
523	75174120	86258781
524	98595555	114366383
525	95767787	100489770
526	75174120	86258781

527 rows × 2 columns

Normalization

We applied Min-Max normalization to the 'total_assets', 'total_expenses', 'incoming_agricultural', 'lasting_investment', 'no_lasting_investmen', and 'education_level' variables. This technique rescales the values to a range between 0 and 1, facilitating comparisons between these financial metrics. By transforming the data in this way, we ensure that all attributes are on a consistent scale, which is essential for many machine learning algorithms and statistical analyses. This normalization process enhances the interpretability of the data and mitigates the influence of extreme values.

Original Columns:

	total_assets	total_expenses	incoming_agricultural	lasting_investment	no_lasting_investmen	education_level
0	120662343	184844064	43775349	12402556	71201668.0	12
1	75174120	86258781	30028818	28411718	28292707.0	7
2	111446005	66439318	58322639	2117823	29246958.0	4
3	75174120	86258781	30028818	28411718	28292707.0	12
4	75174120	86258781	30028818	28411718	28292707.0	9
...	...	...	...	...	...	...
522	72420039	132500494	18017292	64892816	67444725.0	9
523	75174120	86258781	30028818	28411718	28292707.0	9
524	98595555	114366383	78742236	9364798	35089231.0	10
525	95767787	100489770	15815178	75428271	37566051.0	10
526	75174120	86258781	30028818	28411718	28292707.0	7

527 rows × 6 columns

Normalized Columns:

Normalized Data:

	total_assets	total_expenses	incoming_agricultural	lasting_investment	no_lasting_investmen	education_level
0	0.559892	0.787992	0.431686	0.124061	0.735517	0.647059
1	0.291340	0.324202	0.292214	0.285164	0.291477	0.352941
2	0.505481	0.230962	0.579282	0.020564	0.301352	0.176471
3	0.291340	0.324202	0.292214	0.285164	0.291477	0.647059
4	0.291340	0.324202	0.292214	0.285164	0.291477	0.470588
...	...	...	...	...	...	...
522	0.275081	0.541744	0.170345	0.652279	0.696638	0.470588
523	0.291340	0.324202	0.292214	0.285164	0.291477	0.470588
524	0.429615	0.456433	0.786459	0.093492	0.361811	0.529412
525	0.412921	0.391151	0.148003	0.758299	0.387442	0.529412
526	0.291340	0.324202	0.292214	0.285164	0.291477	0.352941

527 rows × 6 columns

Discretization

Only the Age attribute is discretized into three distinct categories: 'Youth', 'Adult', and 'Senior'. This transformation simplifies the analysis by grouping individuals based on age ranges. The defined age ranges are as follows:

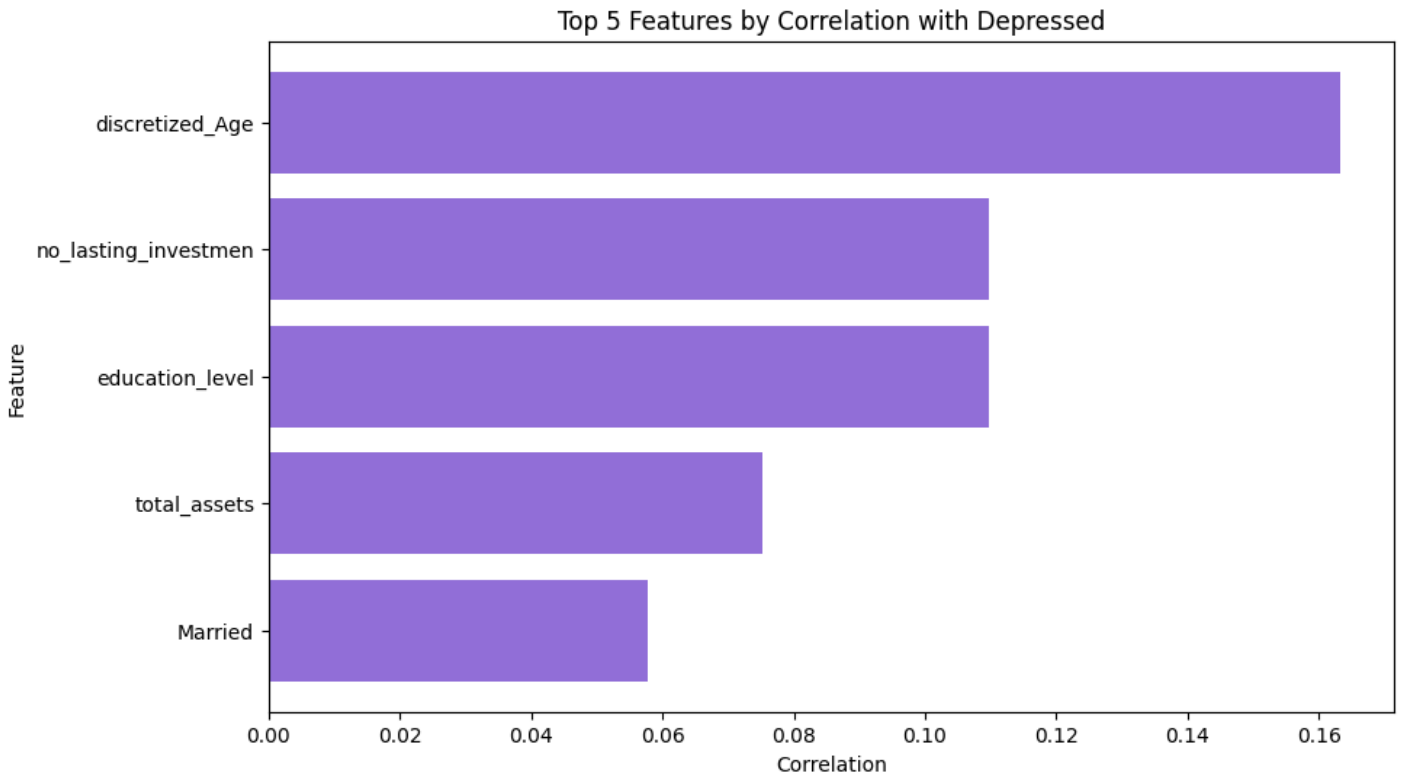
- Youth(0): Ages 0 to 24
- Adult(1): Ages 25 to 59
- Senior(2): Ages 60 and above

```
Original DataFrame with discretized column:
     Age  discretized_Age
0     56                1
1     40                0
2     43                1
3     51                1
4     20                0
..   ...              ...
522   29                0
523   28                0
524   36                0
525   35                0
526   39                0

[527 rows x 2 columns]
```

Feature Selection

We applied filter feature selection (correlation-based), we calculated the correlation between features and the target variable depressed. Then computed the absolute correlation of each feature with the target and selects the top 5 features based on their correlation. Finally, we visualized these top features using a bar plot, but the original dataset remains unchanged as this is only done for the correlation calculation.



Raw Dataset vs Processed dataset

Raw Dataset

Survey_id	Ville_id	sex	Age	Married	Number_children	education_level	total_members	gained_asset	durable_asset	save_asset	living_expenses	other_expen
926	91	1	28	1	4	10	5	28912201	22861940	23399979	26692283	28203066
747	57	1	23	1	3	8	5	28912201	22861940	23399979	26692283	28203066
1190	115	1	22	1	3	9	5	28912201	22861940	23399979	26692283	28203066
1065	97	1	27	1	2	10	4	52667108	19698904	49647648	397715	44042267
806	42	0	59	0	4	10	6	82606287	17352654	23399979	80877619	74503502
483	25	1	35	1	6	10	8	35937466	736707	23399979	30696127	11531066
849	130	0	34	0	1	9	3	41303144	21925041	23399979	66730708	10890451
1386	72	1	21	1	2	10	4	12013633	20323505	48046108	80076849	58456101
930	195	1	32	1	7	9	9	11087568	25224208	80076851	30162281	67184479
390	33	1	29	1	4	10	5	28912201	22861940	23399979	26692283	28203066
540	52	1	84	0	0	1	5	28912201	22861940	23399979	26692283	28203066
557	93	1	59	0	2	9	3	1018915	47245342	23399979	262919	30108896
1280	232	1	38	1	4	10	6	12390944	19186414	23399979	10810375	498078
1195	92	1	27	1	4	10	6	16521259	37155658	23399979	21220366	10506083
603	100	1	56	1	0	12	2	93596368	21140288	5925687	34566505	72469551
729	54	1	24	1	2	10	5	1108353	12219727	1601537	38169963	37860336
770	102	1	25	1	3	10	5	37172832	75432396	80076847	40705733	40278656
76	15	1	44	1	5	12	5	28912201	22861940	23399979	26692283	28203066
1374	267	1	32	1	4	9	5	28912201	22861940	23399979	26692283	28203066
379	22	1	26	1	2	7	4	82606287	20419597	23399979	25357668	99295292
1001	207	1	40	0	0	7	5	28912201	22861940	23399979	26692283	28203066
1356	198	1	55	0	0	6	1	17142671	83440079	24023056	16495831	42985252
137	9	1	34	1	3	10	5	28912201	1905829	23399979	49380727	30221003
840	102	1	43	1	4	4	9	12390944	18978214	80076847	18684598	23542593
309	25	1	51	1	2	12	5	28912201	22861940	23399979	26692283	28203066

Show 25 per page

1 2 10 50 58

wn_farm	incoming_business	incoming_no_business	incoming_agricultural	farm_expenses	labor_primary	lasting_investment	no_lasting_investmen	depressed
	0	0	30028818	31363432	0	28411718	28292707	0
	0	0	30028818	31363432	0	28411718	28292707	1
	0	0	30028818	31363432	0	28411718	28292707	0
	0	1	22288055	18751329	0	7781123	69219765	0
	0	0	53384566	20731006	1	20100562	43419447	0
	0	1	22688441	18907036	0	4442561	76629095	0
	0	0	26692283	22243569	0	22562288	55608922	1
	1	0	9275569	36979933	0	33922659	54600174	0
	0	0	32564587	28738691	1	14018381	15117619	0
	0	0	30028818	31363432	0	28411718	28292707	0
	0	0	30028818	31363432	0	28411718	28292707	1
	0	0	66730709	13968961	1	15714453	20214956	0
	0	0	72069163	56721101	0	20745816	15708408	0
	0	0	3109651	22688441	0	62405292	12144989	0
	0	0	43775349	77808008	0	12402556	71201668	1
	0	0	21353827	37814063	0	23991919	48624439	0
	0	0	6406148	44843035	0	11596846	12491988	0
	0	0	30028818	31363432	0	28411718	28292707	0
	0	0	30028818	31363432	0	28411718	28292707	0
	0	1	42707653	26247411	0	26450653	36790862	0
	0	0	30028818	31363432	0	28411718	28292707	1
	0	0	48046109	58456101	0	25742926	12091604	1
	0	0	30028818	31363432	0	249039	26558821	0
	0	1	58322639	24212127	1	2117823	29246958	1
	0	0	30028818	31363432	0	28411718	28292707	1

Show 25 per page

1 2 10 50 58

Processed Dataset

discretized_Age	no_lasting_investmen	education_level	total_assets	Married	depressed
1	0.73551682	0.64705882	0.55989234	1	1
0	0.29147736	0.35294118	0.29134034	0	1
1	0.30135234	0.17647059	0.5054812	1	1
1	0.29147736	0.64705882	0.29134034	1	1
0	0.29147736	0.47058824	0.29134034	1	1
0	0.8733979	0.76470588	0.82537982	1	1
1	0.12471945	0.47058824	0.15912112	0	1
0	0.85866604	0.52941176	0.32023619	1	1
0	0.29147736	0.52941176	0.29134034	1	1
0	0.29147736	0.64705882	0.29134034	0	1
0	0.29147736	0.35294118	0.29134034	1	1
0	0.39000829	0.35294118	0.1926624	1	1
1	0.24839815	0.0	0.66505339	1	1
0	0.20931265	0.52941176	0.47630025	1	1
0	0.39000833	0.41176471	0.22775476	1	1
1	0.29147736	0.41176471	0.29134034	1	1
0	0.43429595	0.29411765	0.59858958	1	1
1	0.29147736	0.29411765	0.29134034	0	1
0	0.49576127	0.35294118	0.72262601	0	1
0	0.65529705	0.17647059	0.07412954	1	1
1	0.54423258	0.76470588	0.15252135	0	1
1	0.24539423	0.17647059	0.49527796	1	1
0	0.10370351	1.0	0.55540049	1	1
0	0.24591215	0.47058824	0.85515366	1	1
1	0.32555633	0.47058824	0.78538472	0	1

Show 25 per page

1 2 10 20 22

Data Mining Technique

We applied both supervised and unsupervised learning to our data using classification and clustering techniques.

Classification

We performed a comprehensive evaluation of Decision Tree classification on our dataset. Our model will predict the class label (depressed) which has two classes: 0 and 1, the prediction is made on the rest attributes:(discretized_Age, no_lasting_investmen, education_level, total_assets, Married).

As we mentioned on before, we used classification that is a type of supervised learning, so we need training data to train the model, so we split our dataset into two subsets which are **training data** and **testing data**.

We tried 3 different sizes of training subsets which are 70%, 60%, and 80% and used two attribute selection measures (IG (entropy) ,Gini index). To evaluate our model and determine the best partitioning, we examine its metrics: accuracy ,sensitivity, specificity, precision, and error rate. Also, the confusion matrix that evaluates the performance by comparing predicted and actual labels.

Python packages used for classification

1. scikit-learn (sklearn):

- DecisionTreeClassifier: Used to build and train decision tree models for classification.
- train_test_split: Splits the dataset into training and testing sets.
- Metrics: Includes accuracy_score, precision_score, recall_score, f1_score, and confusion_matrix to evaluate model performance.
- StandardScaler: Standardizes features to have a mean of 0 and a variance of 1.
- LabelEncoder: Encodes categorical labels into numerical values.

2. yellowbrick:

- SilhouetteVisualizer: Visualizes the quality of clustering but not directly used for classification.

These tools help train, validate, and evaluate the performance of classification models, ensuring robust and reliable analysis.

Clustering

In the clustering process, which is a type of unsupervised learning, we omitted the "depressed" class label attribute since it does not use class labels, This allows the clustering algorithm to group data based on feature similarities alone, without using predefined labels, enabling unbiased clustering. Instead, we utilized all other attributes such as:(discretized_Age, no_lasting_investmen, education_level, total_assets, Married). All of which are numeric and require no conversion prior to clustering.

We employed the K-means algorithm. KMeans is a clustering algorithm that groups data points into k clusters based on feature similarity. It assigns points to the nearest cluster center, then recalculates centers until clusters stabilize, helping to reveal patterns in the data.

For cluster validation, the silhouette score was used here to measure how similar data points are within a cluster compared to other clusters. Higher scores indicate better-defined clusters.

Additionally, we used the elbow method and WSS method to compare three different cluster sizes to determine the optimal number by assessing the separation and compactness of the clusters.

Python packages used for Clustering

1. scikit-learn (sklearn):

- KMeans: Implements the K-Means clustering algorithm, a popular method for partitioning data into clusters.
- silhouette_samples and silhouette_score: Calculate silhouette scores to assess the quality of clustering. Silhouette scores measure how similar a data point is to its own cluster compared to other clusters.
- calinski_harabasz_score: Measures the dispersion within and between clusters to evaluate clustering performance.

2. yellowbrick (yellowbrick):

- SilhouetteVisualizer: Provides a visual interpretation of silhouette analysis, helping assess how well data points fit within their clusters.

These tools are essential for performing, evaluating, and visualizing clustering tasks in Python, enabling more effective analysis of patterns and group structures in data.

Evaluation And Comparison

Classification

Classification was applied to predict depression in individuals based on features in the dataset. The Decision Tree algorithm was employed due to its interpretability and efficiency in handling categorical and numerical data. Two attribute selection measures—Information Gain (Entropy) and

Gini Index—were used to construct and evaluate the model. The data was split into three distinct partitions for training and testing: 60-40, 70-30, and 80-20. This ensures robust evaluation of the model's performance across different configurations.

1. 60-40 Split

Entropy:

Accuracy: 56%.

Sensitivity: 35%

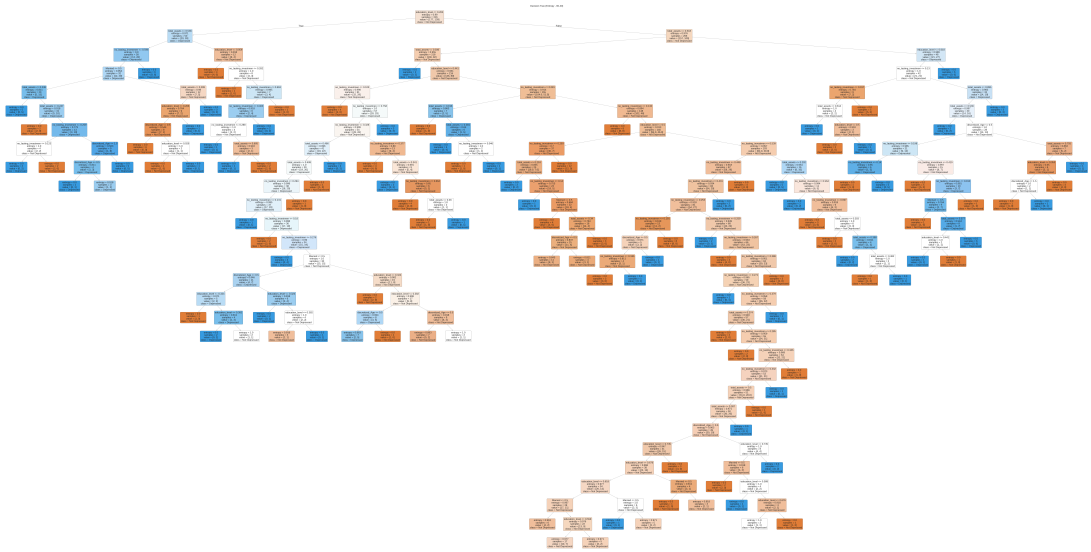
Specificity: 70%

Precision: 43%

Error Rate: 44%

Confusion matrix:

Actual \ Predicted	Positive	Negative
Positive	29	54
Negative	38	90



Gini:

Accuracy: 61%

Sensitivity: 43%

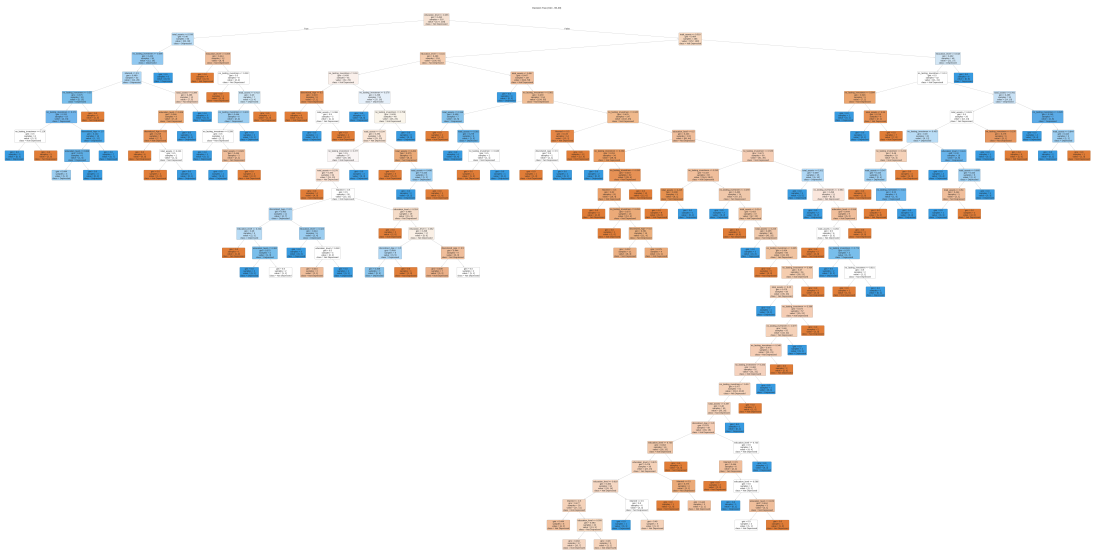
Specificity: 72%

Precision: 50%

Error Rate: 39%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	36	47
Negative	36	92



Analysis: Gini performs better than Entropy for the 60-40 split, achieving higher accuracy and specificity. However, sensitivity remains low, indicating the model struggles to identify positive cases.

2. 70-30 Split

Entropy:

Accuracy: 63%

Sensitivity: 42%

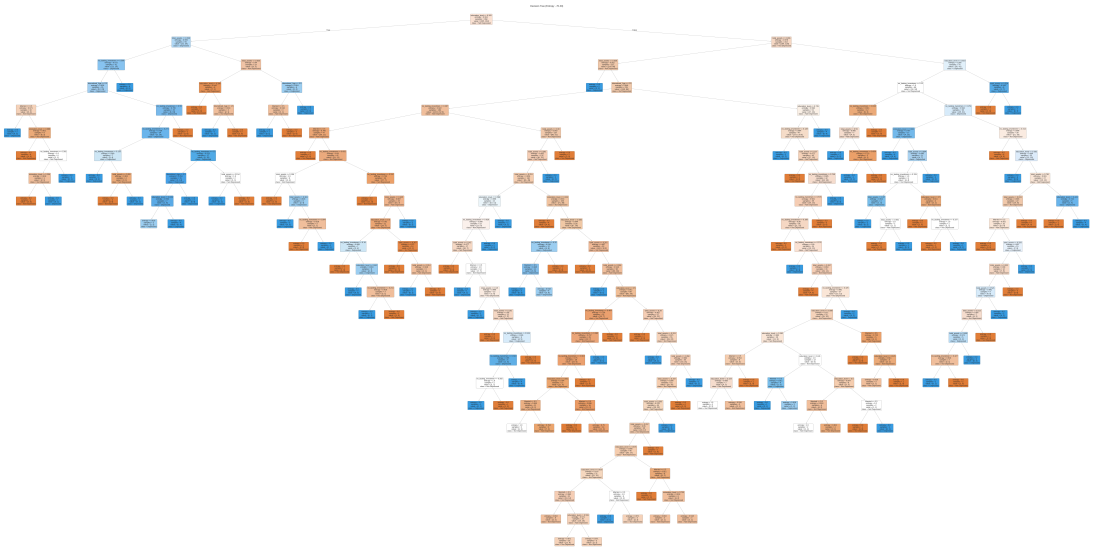
Specificity: 75%

Precision: 50%

Error Rate: 37%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	25	34
Negative	25	75



Gini: Accuracy: 53%

Sensitivity: 41%

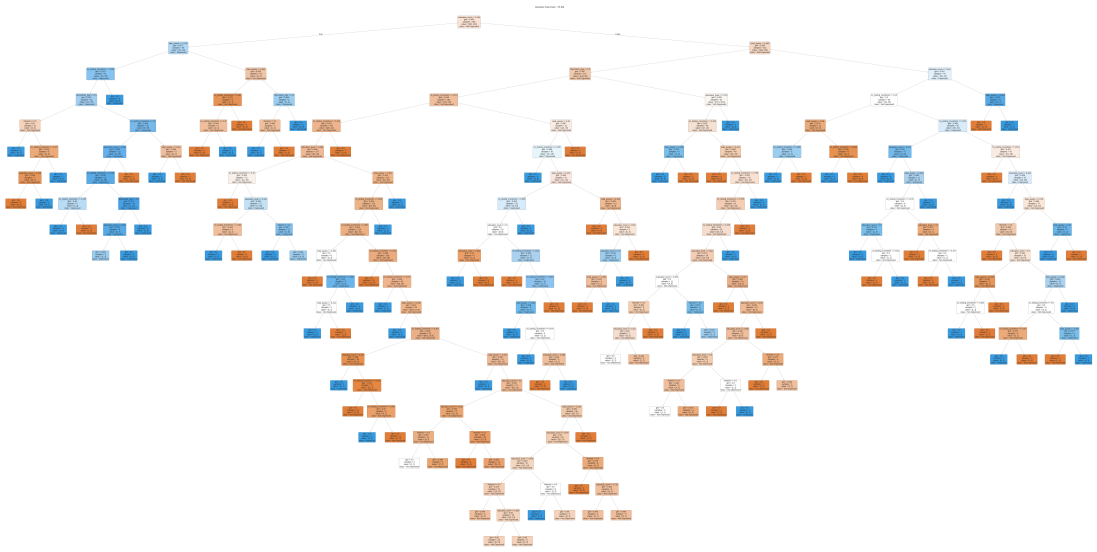
Specificity: 61%

Precision: 38%

Error Rate: 47%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	24	35
Negative	39	61



Analysis: Entropy outperforms Gini in the 70-30 split, achieving the highest accuracy (63%) among all configurations. This configuration balances sensitivity and specificity better, making it the best-performing split for Entropy.

3. 80-20 Split

Entropy:

Accuracy: 53%

Sensitivity: 36%

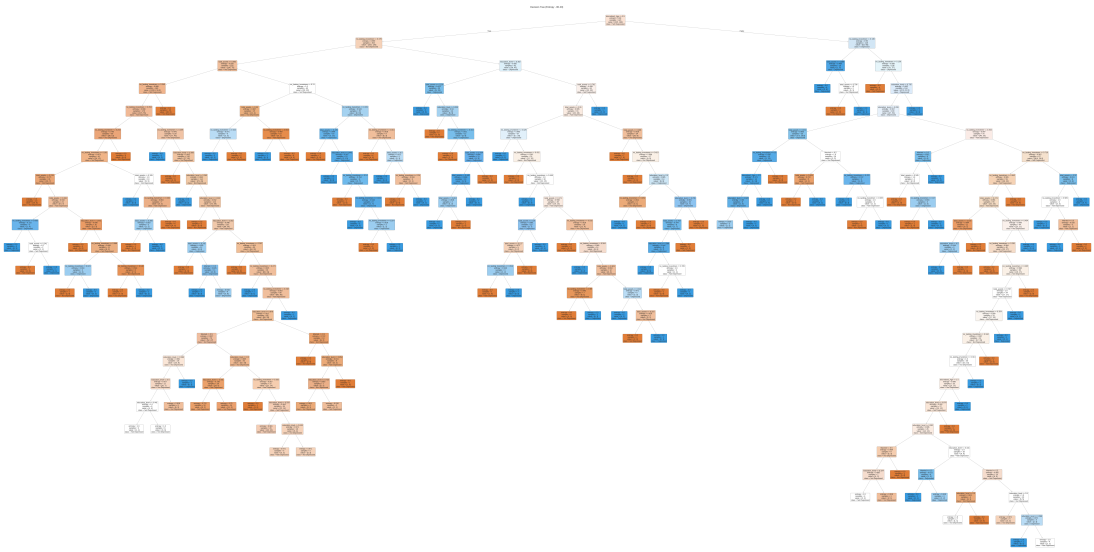
Specificity: 61%

Precision: 43%

Error Rate: 47%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	13	28
Negative	27	43



Gini:

Accuracy: 54%

Sensitivity: 39%

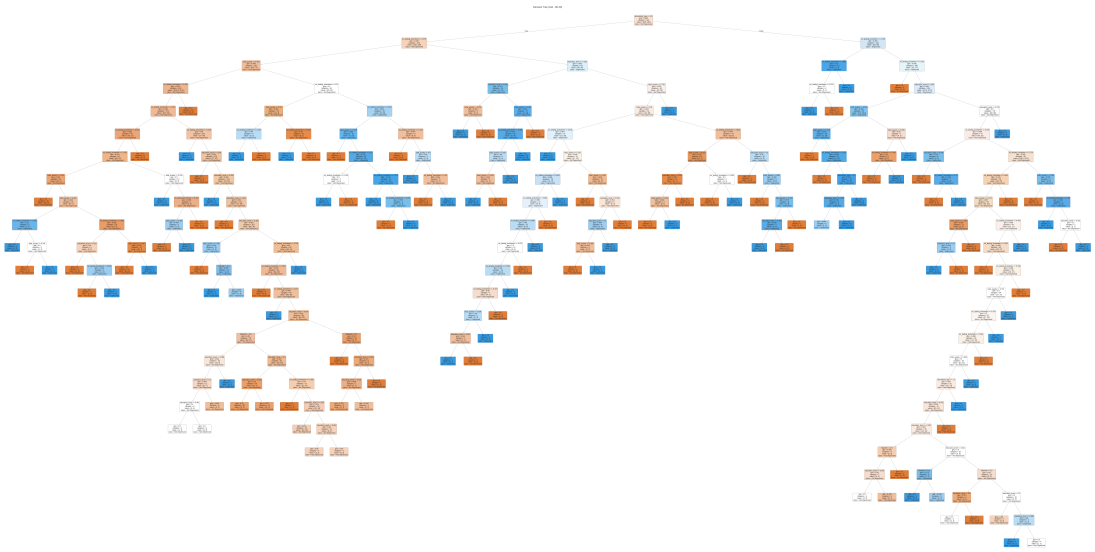
Specificity: 61%

Precision: 34%

Error Rate: 46%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	14	22
Negative	27	43



Analysis: Both Entropy and Gini exhibit the lowest performance in this split. Accuracy and specificity are relatively low, and the number of misclassifications is higher compared to other splits.

Performance Metrics Summary:

Entropy

Metric	60-40 Split	70-30 Split	80-20 Split
Accuracy	56.39%	62.83%	52.83%
Sensitivity (Recall)	35%	42%	36%
Specificity	70%	75%	61%
Precision	43%	50%	43%
Error Rate	44%	37%	47%

Gini

Metric	60-40 Split	70-30 Split	80-20 Split
Accuracy	60.65%	53.49%	53.77%
Sensitivity (Recall)	43%	41%	39%
Specificity	72%	61%	61%
Precision	50%	38%	34%
Error Rate	39%	47%	46%

Findings and Conclusion

Best Split for Entropy: The 70-30 split achieved the highest accuracy (63%) and the best balance between sensitivity (42%) and specificity (75%), making it the most effective configuration.

Best Split for Gini: The 60-40 split performed the best for Gini, with the highest accuracy (61%) and specificity (72%). However, sensitivity (43%) remains a concern.

Overall Best Configuration: When comparing both measures, Entropy with the 70-30 split emerges as the best overall, providing the most balanced performance.

Training-Testing Split	**60% Training, 40% Testing**		**70% Training, 30% Testing**		**80% Training, 20% Testing**	
Attribute Selection	IG	Gini Index	IG	Gini Index	IG	Gini Index
Accuracy (%)	56.39%	60.65%	62.83%	53.49%	52.83%	53.77%

Clustering

Clustering was applied to group individuals in the dataset based on shared characteristics. The aim was to identify patterns and meaningful groupings, which can provide insight into potential factors related to depression or anxiety. The K-means clustering algorithm was selected due to its ability to effectively partition data into non-overlapping groups.

- Algorithm: K-means clustering.
- Metrics for Evaluation:
    - Average Silhouette Width: Indicates cluster cohesion and separation, where higher values are better.
    - WCSS (Within-Cluster Sum of Squares): Measures cluster compactness; lower values indicate tighter groups.
- Cluster Numbers Tested: (K = 5, 6, 7).

Clustring Trial

We tested clustering for three different values of K=5,6,7:

Based on the plots:

For (K = 5):

- Clusters are broad, with some overlapping.
- Good for grouping data into larger, general categories.
- Less precise, misses finer details in the data.

For (K = 6):

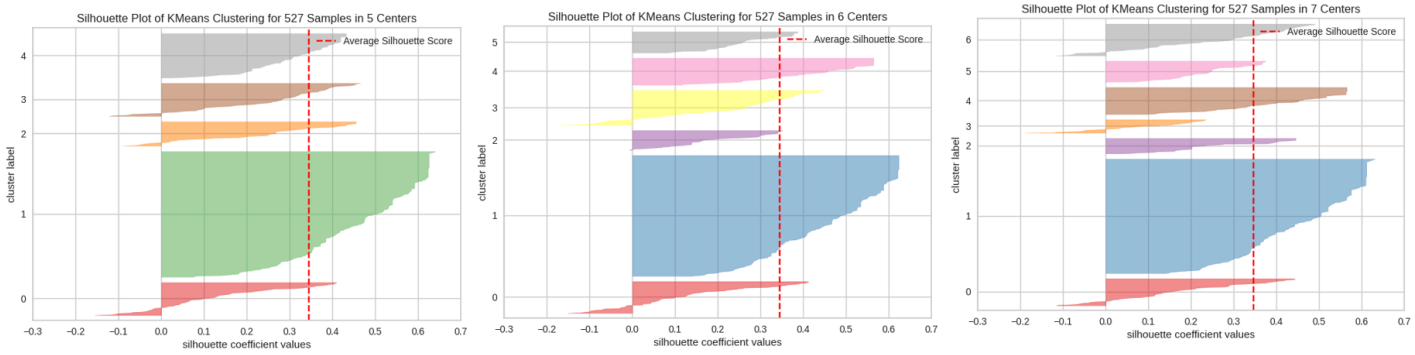
- Clusters are more distinct and balanced.
- Good separation between groups with well-positioned centroids.
- Likely the best choice as it captures meaningful patterns without over-segmentation.

For (K = 7):

- Clusters are smaller and tighter.
- Some clusters might represent outliers or very specific patterns.
- Risks over-segmentation, creating too many small groups.

Recommendation:

- (K = 6) is the best option, offering clear and meaningful clusters.



K = 5

- **Cluster Centers**:
  - There are 5 cluster centroids distributed across the feature space.
  - The centers suggest varied characteristics among the clusters.
- **Cluster Labels**:
  - The data points are grouped into 5 clusters, with some clusters appearing more dominant in size.
- **Silhouette Score**:
  - The average silhouette score is **0.3449**, which indicates moderate cluster cohesion and separation.
  - Some points may be assigned to the wrong clusters or overlap between clusters.

K = 6

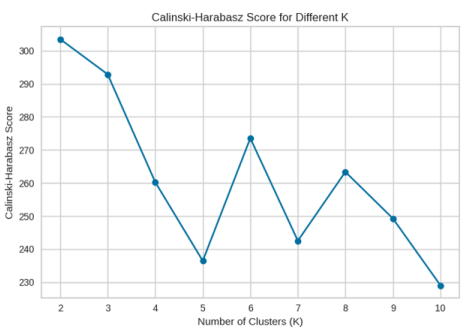
- **Cluster Centers**:
  - Adding a 6th cluster refines the groupings, with more separation in the clusters.
  - Some clusters split into smaller, more specific groups.
- **Cluster Labels**:
  - The labels reflect better grouping, with more balance compared to (K = 5).
- **Silhouette Score**:
  - The average silhouette score is **0.3446**, similar to (K = 5), indicating no significant improvement.
  - The clustering is more detailed but does not significantly improve quality.

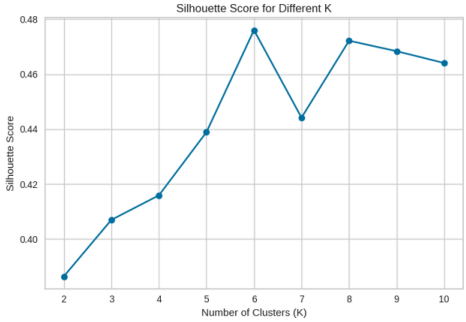
K = 7

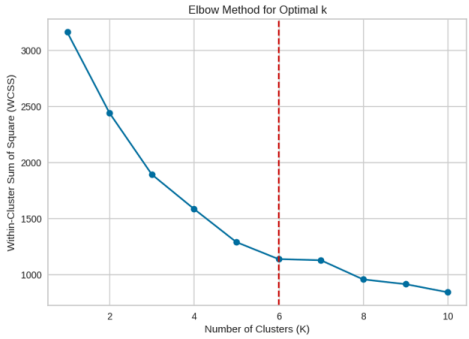
- **Cluster Centers**:
  - Adding a 7th cluster further divides the groups, creating smaller clusters.
  - Some clusters may represent outliers or very specific patterns.
- **Cluster Labels**:
  - Labels show over-segmentation, with some clusters potentially unnecessary.
- **Silhouette Score**:
  - The average silhouette score is **0.3459**, slightly higher than (K = 5) and (K = 6).
  - While the score improves slightly, the added complexity of (K = 7) may not justify the improvement.

Recommendation:

- **K = 6** provides a good balance of detail and interpretability, but there is little improvement in silhouette score compared to (K = 5).
- If simplicity is preferred, stick with (K = 5).
- Avoid (K = 7) unless highly detailed segmentation is necessary, as it risks over-segmentation.







The results for each trial are summarized below:

	K=5	K=6(BEST)	K=7
Average Silhouette score	0.44	0.48	0.44
Total within-cluster sum of square (WCSS)	1581.77	1226.94	1170.98
Calinski-Harabasz Score Analysis	238	274	242

Comparison and Optimal Cluster Selection

Average Silhouette score: Measures the separation of clusters.

- The highest silhouette score was achieved at K=6 (0.48), indicating well-separated and balanced clusters.
- At K=7, the silhouette score drops back to ~0.44, which is the same as K=5. This suggests that adding a 7th cluster leads to weaker-defined clusters and potential over-segmentation.

Total Within-Cluster Sum of Squares (WCSS):

The decrease in WCSS between:

- K=5 to K=6: 1581.77 − 1226.94 = 354.83
- K=6 to K=7: 1226.94 − 1170.98 = 55.96

Interpretation: The significant drop in WCSS occurs between K=5 and K=6 (354.83). The decrease in WCSS from K=6 to K=7 is minimal (55.96), which suggests that increasing K beyond 6 results in diminishing returns and over-segmentation.

Calinski-Harabasz Score Analysis:

- K=6: The highest Calinski-Harabasz Score (274) is observed, suggesting that clusters are well-separated and compact.

- K=5: A lower score (238), indicating less optimal clustering compared to K=6.
- K=7: A slight decrease to 242, which aligns with the observation that adding more clusters does not improve the clustering structure significantly.

Optimal Number of Clusters:

Based on the *majority rule* (considering all metrics), the optimal K was determined to be K=6, offering a balance between cluster separation and compactness.

- Based on the Average Silhouette score:

  K=6 provides the highest Average Silhouette score (0.48), which is the key metric for evaluating the separation and quality of clusters.

- Based on the Elbow Method:

  The optimal number of clusters is K=6, as it provides a good balance between compactness (WCSS) and simplicity while avoiding over-segmentation.

- Based on the Calinski-Harabasz Score Analysis:

  K=6 is optimal as it maximizes the quality of clustering by achieving the highest score (274).

Findings

In this section, we analyze the results obtained from applying different data mining techniques to the selected dataset. The goal of this analysis is to evaluate the effectiveness of the models used and determine their suitability for solving the problem at hand. Classification and clustering techniques were utilized to uncover patterns and valuable insights within the data. This section provides a comprehensive discussion of the findings, a comparison of different models, and an interpretation of their performance based on the evaluation criteria. The primary aim is to identify the best model for solving the problem and to present a clear understanding of the relationships revealed within the data

Classification:

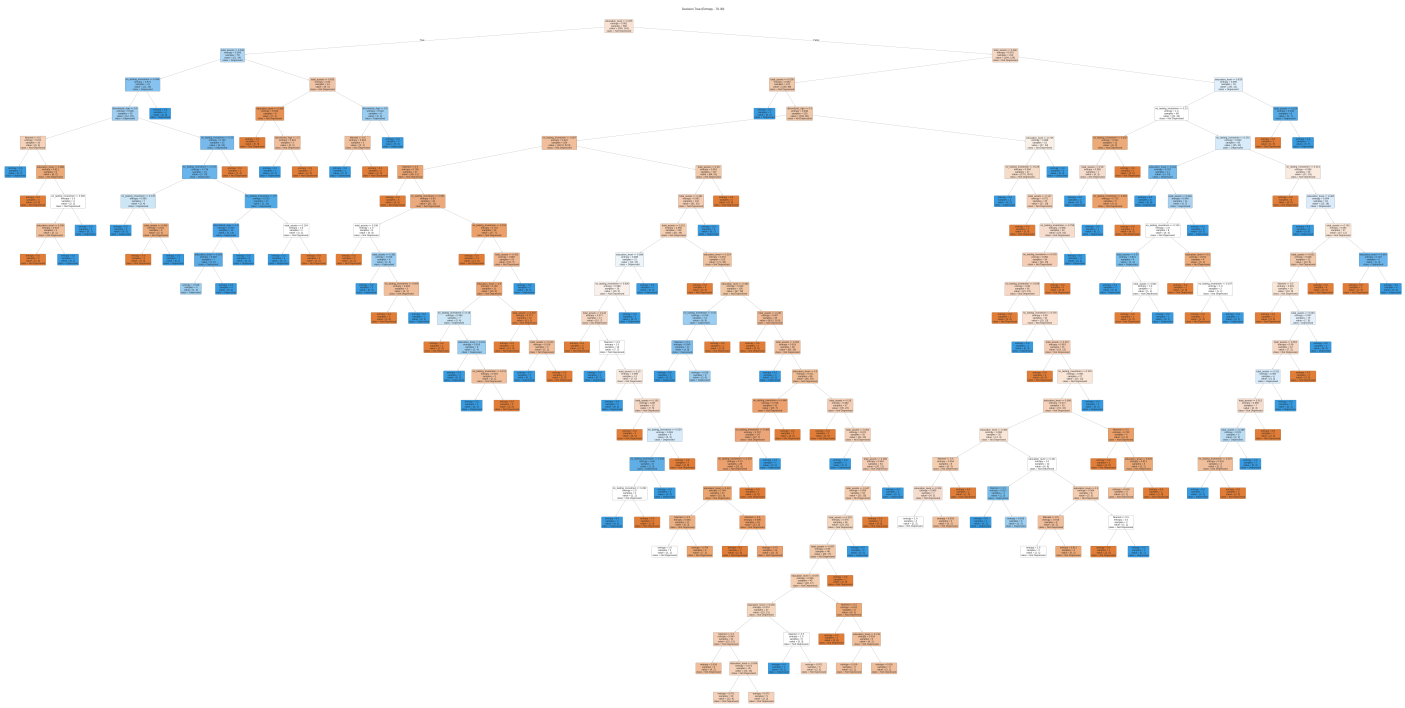
In this part of the findings, we focus on the results obtained from applying the Decision Tree classification technique. To evaluate the model's performance, we split the data into three different training and testing sets. Compared to the other splits, the 70-30 configuration achieved the highest

accuracy and demonstrated a better balance between false positives and false negatives, making it the strongest predictor among the three splits. Here are the detailed metrics for the **70-30 split:**

- Accuracy: ~62.83%
- True Negatives: 75
- True Positives: 25
- False Negatives: 34
- False Positives: 25

Attribute Selection Measures& Building Decision Tree

We utilized two attribute selection measures, Gini Index and Entropy, to construct the Decision Tree. However, we focus on the results obtained from the Decision Tree classification technique using the Entropy criterion, as it demonstrated the best performance. The model built with entropy achieved an accuracy of 62.83%, compared to 53.49% for Gini, demonstrating its superiority in classifying the data.



Root Node:

- Feature: education_level <= 0.265.
- The root node is based on the feature education_level, which is the primary splitting factor.
- Samples: 368.
- 205 samples: "Not Depressed".

- 163 samples: "Depressed".
- Entropy: 0.991 (indicates high overlap between classes at the root node).

Key Influential Features:

- education_level: The most significant feature, appearing at the root node.
- total_assets : Plays a critical role in subsequent nodes as an economic indicator.
- no_lasting_investmen: Indicates long-term financial stability and appears in important splits.
- discretized_Age: Appears in lower levels as a secondary factor influencing classification.

Tree Performance:

- The tree effectively separates the data based on the identified features, progressively reducing entropy.
- Leaves provide the final classification into "Depressed" and "Not Depressed".
- Several leaves have a small number of samples, reflecting fine-grained detail in the model.
  Tree Complexity:
- The tree contains many branches, making it precise but potentially prone to overfitting.
- The significant depth of the tree improves predictions but reduces interpretability.

Results:

- The tree highlights that education and financial assets (e.g. total_assets) are strong indicators of mental health status.
- Economic and social factors play a critical role in classification.
- Entropy decreases progressively at each tree level, indicating improved classification accuracy.
- While attributes like education_level and total_assets were highly influential, other attributes like Married and discretized_Age showed limited or secondary impact. This analysis suggests that mental health in this dataset is primarily driven by educational and economic factors, while demographic attributes play a smaller role.