CSE506: Data Mining

Assignment 1

Group 19.
Group Members: Ananya Kansal (2019458), Avishi Gupta (2019155), Jahnvi Kumari (2019469), Manvi Goel (2019472), Prachi Goyal (2019186)

Exploratory Data Analysis and Data Preprocessing

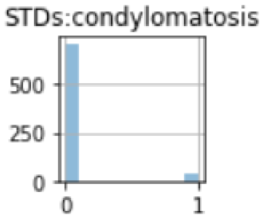
Dataset 1

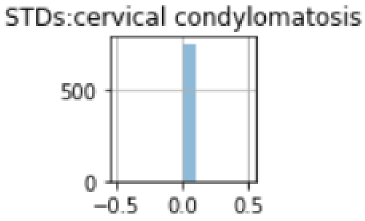
The cervical cancer dataset contains indicators and risk factors for predicting whether a woman will get cervical cancer. The task is **binary classification** to predict biopsy as healthy or cancerous. The dataset has **858 objects** and **35 features**. The preprocessing steps followed are as follows →

1. Replacing '?' with Nan
2. Dropping columns with more than 50% missing data
3. Dropping rows with more than 40% missing data
4. Changing the datatype of features from object to int/float
5. Analyzing distribution using histogram and unique values
6. Dropping columns with just one unique value
7. Analyzing correlation using the heatmap
8. Filling NA values using bfill

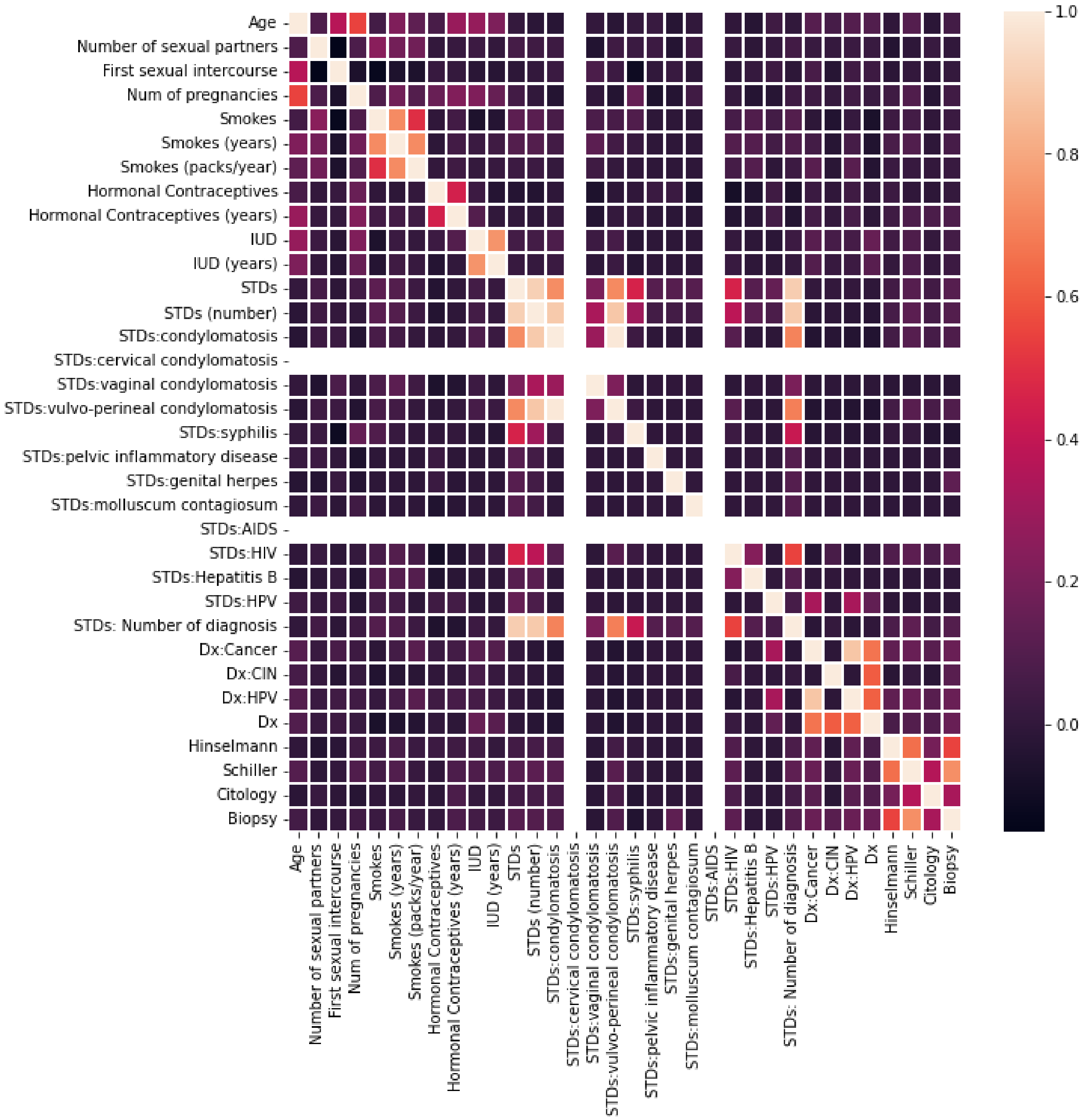
Dataset size after preprocessing → **753 objects** and **32 features**.

Histogram of 2 features, from the second distribution, we can infer that STDs: cervical condylomatosis just has one unique value for all the objects, which is '0', and hence the model doesn't have anything to learn from this feature for the classification task. As a result, we drop this column **(step 6)**.





Initially, the dataset had **55 cancer-positive** samples and **803 noncancerous** samples. Cancer prediction datasets mostly have non-cancerous samples as the majority. From the exploratory data analysis, we could understand that most of the missing values were present in samples labeled as non-cancerous. After all the preprocessing steps, the final count of cancer-positive samples was 53, only two less than the initial count, whereas the count of healthy samples was 700, 103 less than the initial count.



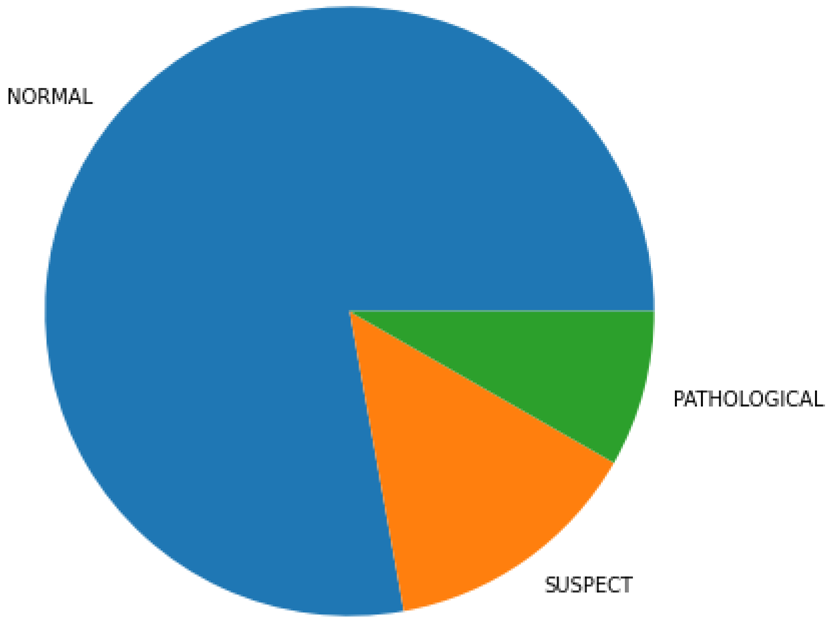
Heatmap helps us to understand the **correlation** between features as well as between features and labels. We want a **minimum correlation between features** (or repeated information) and a maximum correlation between features and labels so that the model can learn something from each feature. Here the lightly colored blocks represent a higher correlation between the two features.

Dataset 2

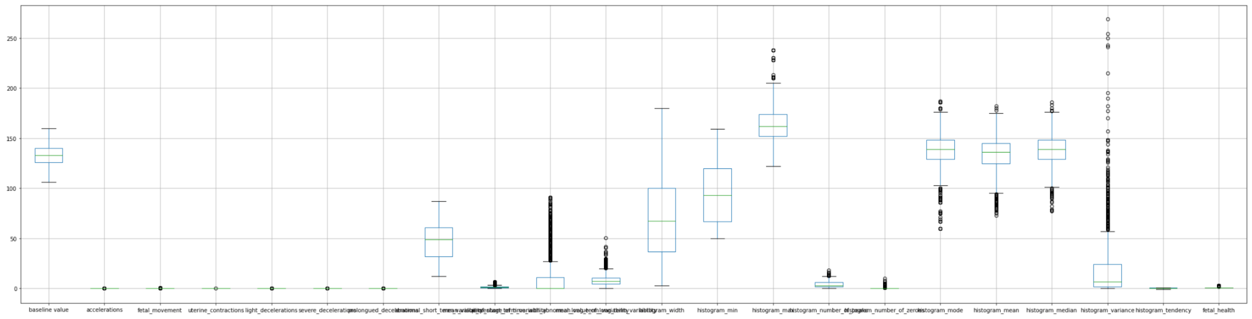
This dataset contains 2126 records of features extracted from Cardiotocogram exams to classify fetal health into 3 categories - normal, suspect and pathological using 21 measures like Baseline Fetal Heart Rate and uterine contractions per second.

There were no null values in the dataset, and we were not required to fill data or drop rows.

The following shows the distribution of the target column. We see a skewed distribution in the pie chart.



We also plotted a box plot to see the distribution and variance of the numerical columns.



After observing the dataset, we perform the following preprocessing steps to create the final dataset.

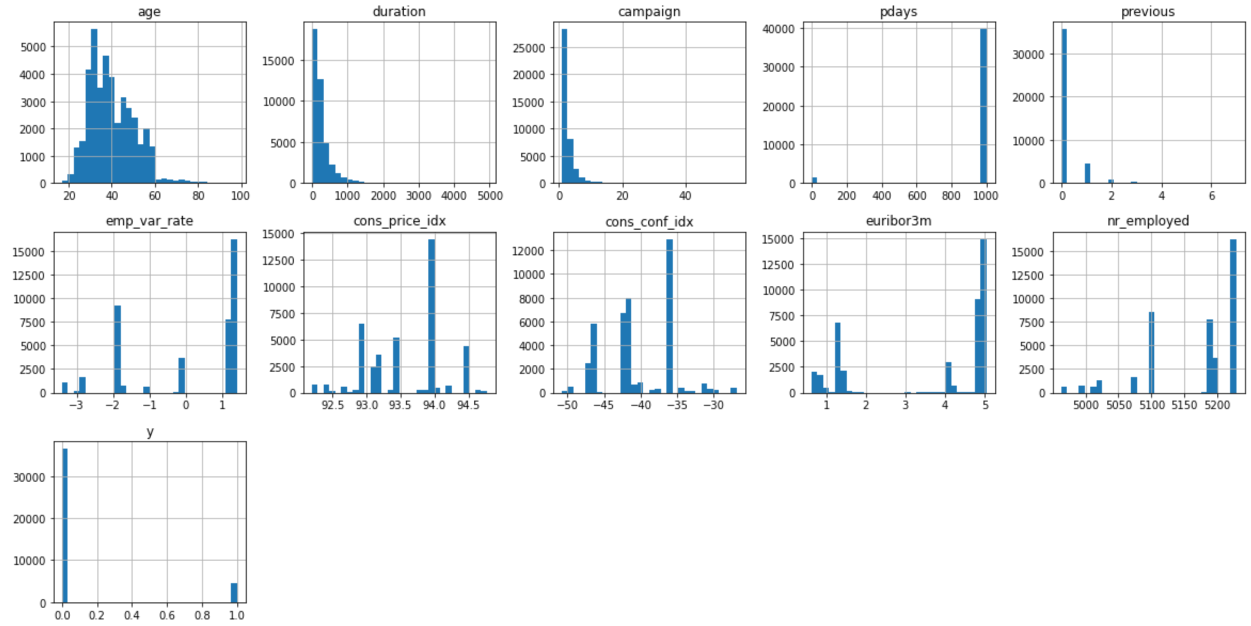
1. We then binarize the columns with categorical features.
2. We used a Standard scaler to standardize the continuous variables.

Dataset 3 - Banking Dataset.

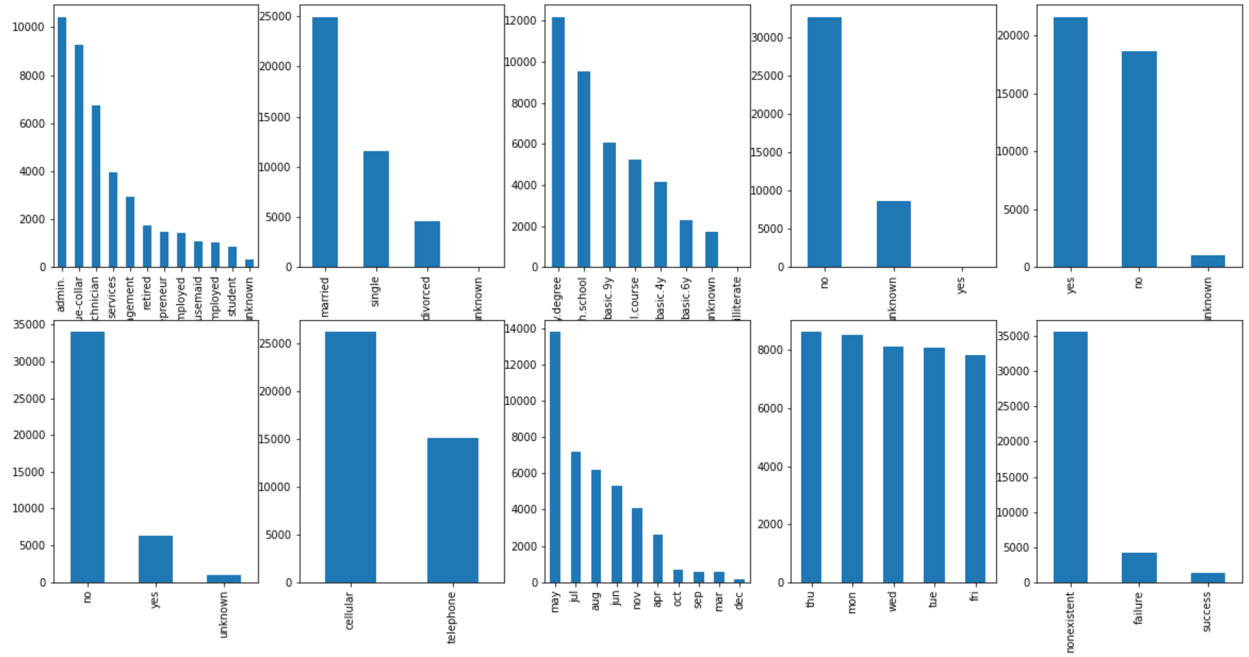
Size of dataset - 41188 samples and 20 attributes. The target column is ‘y’ and contains two possible values. Hence the problem is binary classification.

Null Values in the dataset - None. There are no null values in the dataset, hence we do not need to handle them.
Next, we plot histograms for both the numerical and categorical columns of the dataset.

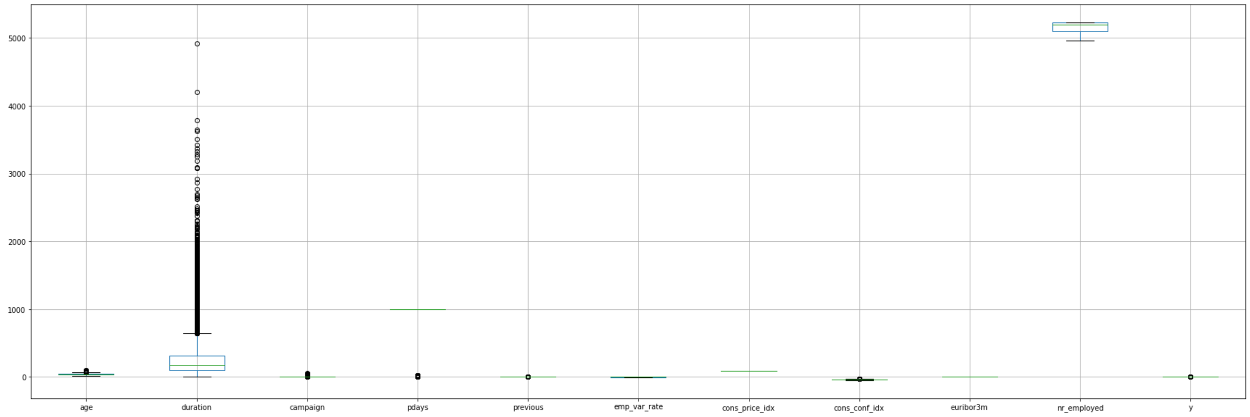
Histogram for Numerical Attributes.



Histogram for Categorical Attributes



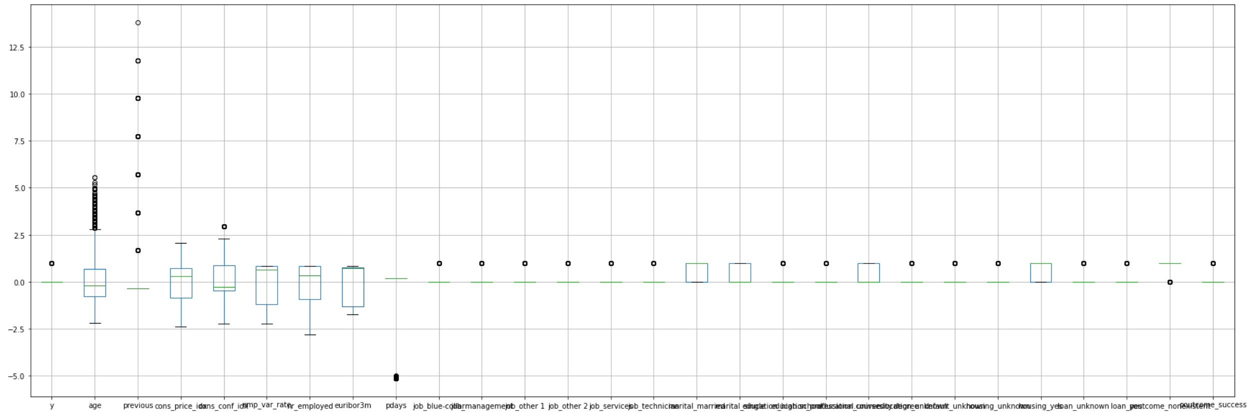
BoxPlot



After observing the dataset, we perform the following preprocessing steps to create the final dataset.

1. We first remove the columns contact, month, day of the week, duration, and campaign.
2. We then merge a few categories from the categorical columns to improve the category rations.
3. We then binarize the columns with categorical features.
4. We used a Standard scaler to standardize the continuous variables.

The box plot of the cleaned dataset is. We can see the difference between the two box plots to see that scaling has improved the bias in the dataset.



Model Implementation (Q1)

Logit Tree is a
type of decision tree that splits data at nodes using Logistic Regression.

Construction:
Consider n data points at a node (X, y). We implement a logistic regression model (let’s name it model) at the node and calculate training loss (l).

Splitting at nodes with one attribute
Next, we check if splitting the node into two results in a lower loss achievable. To get the best split, we first iterate over all attributes. For each attribute, we implement a logistic regression model over X[:, attribute] and calculate the loss. Attribute corresponding to the minimum loss is chosen (let’s call it attr).

Next, to find a threshold, we consider two ways to initialize a search space. Consider unique values that attr takes in X. If num of unique values exceeds 100, then we consider 100 uniform values between minimum and maximum values of attr. We iterate over this search space. For each threshold, data points (X, y) are split into two datasets.

Left datapoints (X_1, y_1) → X[:, attr] <= threshold

Right datapoints (X_2, y_2) → X[:, attr] > threshold

Where $|X_1| = n_1$ and,
$|X_2| = n_2$ and,
$n_1 + n_2 = n$

Splitting at nodes with two attribute
For splitting a node using two attributes, we consider all unordered pairs of attributes possible. For each pair (attribute_1, attribute_2), we implement a logistic regression model over concatenate(X[:, attribute_1], X[:, attribute_2]). The attribute pair corresponding to the minimum loss is chosen (let’s call it (attr_1, attr_2)).

Next, to find thresholds for (attr_1, attr_2), we initialize a search space as above for attr_1 and attr_2. The final search space is a cross-product of the search space for attr_1 and attr_2. We iterate over this search space. For each threshold pair (thresh_1, thresh_2), datapoints (X, y) are split into two datasets.

If X[:, attr_1] <= thresh_1 and X[:, attr_2] <= thresh_2, then point belongs to Left dataset (X_1, y_1)

else belongs to Right dataset (X_2, y_2)

Where $|X_1| = n_1$ and,

$|X_2| = n_2$ and,
$n_1 + n_2 = n$

Checking for split

For each (X_1, y_1) and (X_2, y_2), we build logistic regression models ($model_1$ and $model_2$) and calculate the training losses (l_1 and l_2), respectively.

We calculate the average loss after split as: loss_after_split = $(n_1 * l_1 + n_2 * l_2) / (n_1 + n_2)$
We consider the minimum possible loss_after_split over all thresholds.

If minimum_loss_after_split < l, we split the node into datasets (X_1, y_1) and (X_2, y_2) with models $model_1$ and $model_{2,}$ respectively.

Parameters for LogitTree class:

1. `max_depth:` sets the maximum depth of the decision tree. Default value = 5.
2. `min_leaf_samples:` minimum number of training samples at each leaf node. Default value = 10.
3. `num_attributes_split:` number of attributes considered at each node for splitting in the decision tree. Can only take values 1 or 2. Default value = 1.

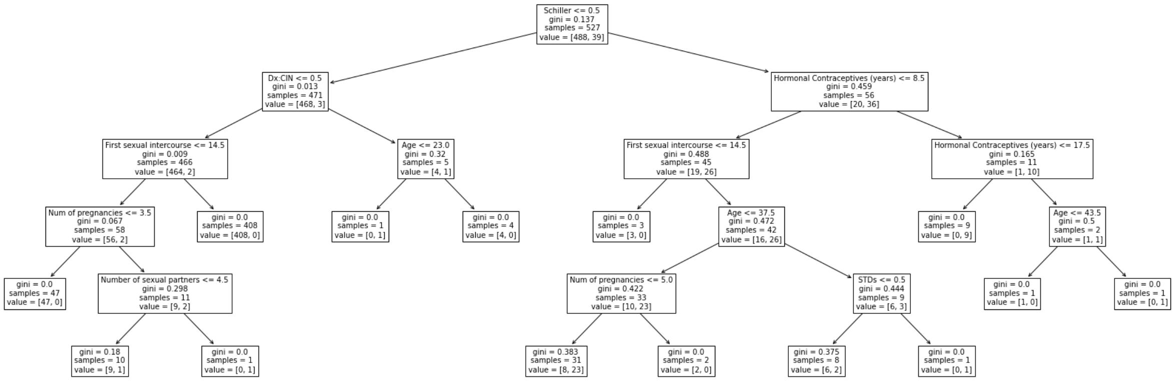
Assumptions

1. The number of attributes considered at each node for splitting in the decision tree can either be 1 or 2.
2. Data takes only numerical values.

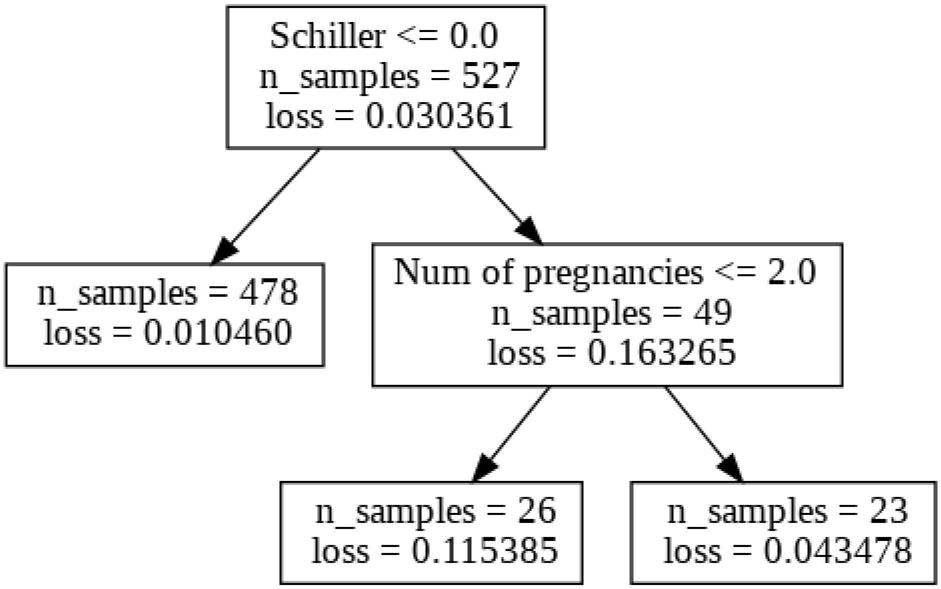
Visualizations (Q2)

Drive Link

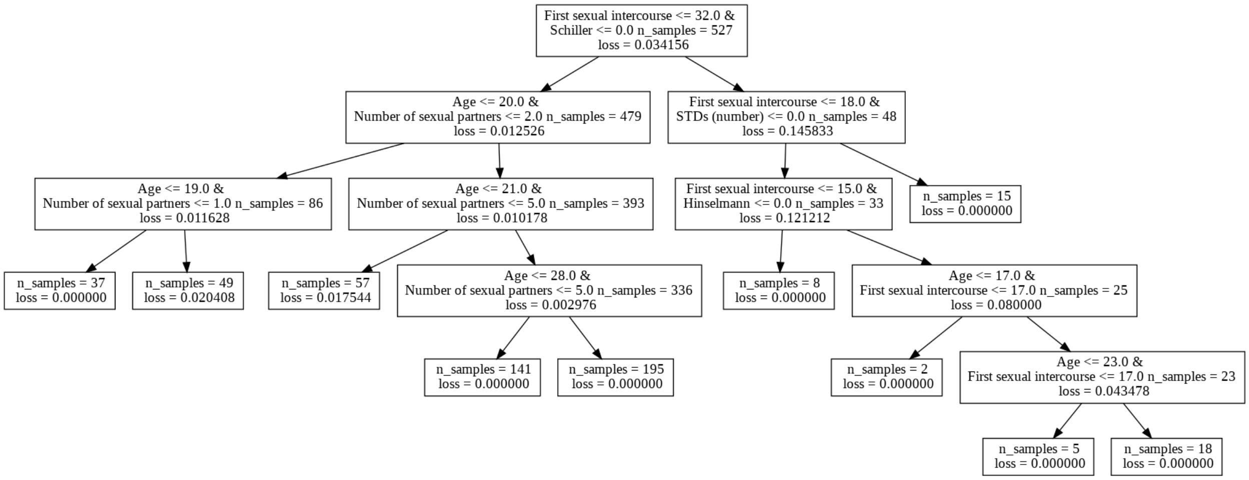
Dataset 1: Vanilla Decision Tree


Schiller <= 0.5 gini = 0.137 samples = 527 value = [488, 39]
Dx:CIN <= 0.5 gini = 0.013 samples = 471 value = [468, 3]
Hormonal Contraceptives (years) <= 8.5 gini = 0.459 samples = 56 value = [20, 36]
First sexual intercourse <= 14.5 gini = 0.009 samples = 466 value = [464, 2]
Age <= 23.0 gini = 0.32 samples = 5 value = [4, 1]
First sexual intercourse <= 14.5 gini = 0.488 samples = 45 value = [19, 26]
Hormonal Contraceptives (years) <= 17.5 gini = 0.165 samples = 11 value = [1, 10]
Num of pregnancies <= 3.5 gini = 0.067 samples = 58 value = [56, 2]
gini = 0.0 samples = 408 value = [408, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 4 value = [4, 0]
gini = 0.0 samples = 3 value = [3, 0]
Age <= 37.5 gini = 0.472 samples = 42 value = [16, 26]
gini = 0.0 samples = 9 value = [0, 9]
Age <= 43.5 gini = 0.5 samples = 2 value = [1, 1]
gini = 0.0 samples = 47 value = [47, 0]
Number of sexual partners <= 4.5 gini = 0.298 samples = 11 value = [9, 2]
Num of pregnancies <= 5.0 gini = 0.422 samples = 33 value = [10, 23]
STDs <= 0.5 gini = 0.444 samples = 9 value = [6, 3]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.18 samples = 10 value = [9, 1]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.383 samples = 31 value = [8, 23]
gini = 0.0 samples = 2 value = [2, 0]
gini = 0.375 samples = 8 value = [6, 2]
gini = 0.0 samples = 1 value = [0, 1]

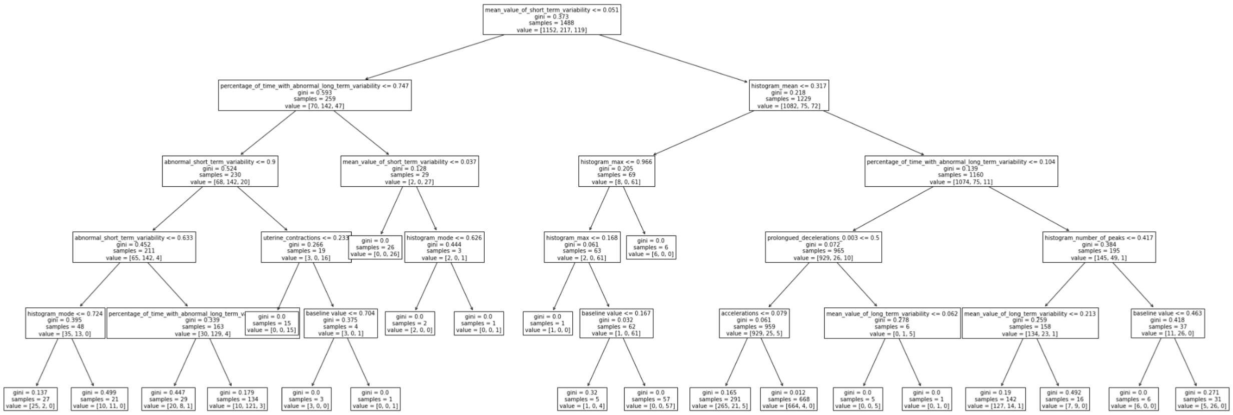
Dataset 1: single attribute


Schiller <= 0.0 n_samples = 527 loss = 0.030361
n_samples = 478 loss = 0.010460
Num of pregnancies <= 2.0 n_samples = 49 loss = 0.163265
n_samples = 26 loss = 0.115385
n_samples = 23 loss = 0.043478

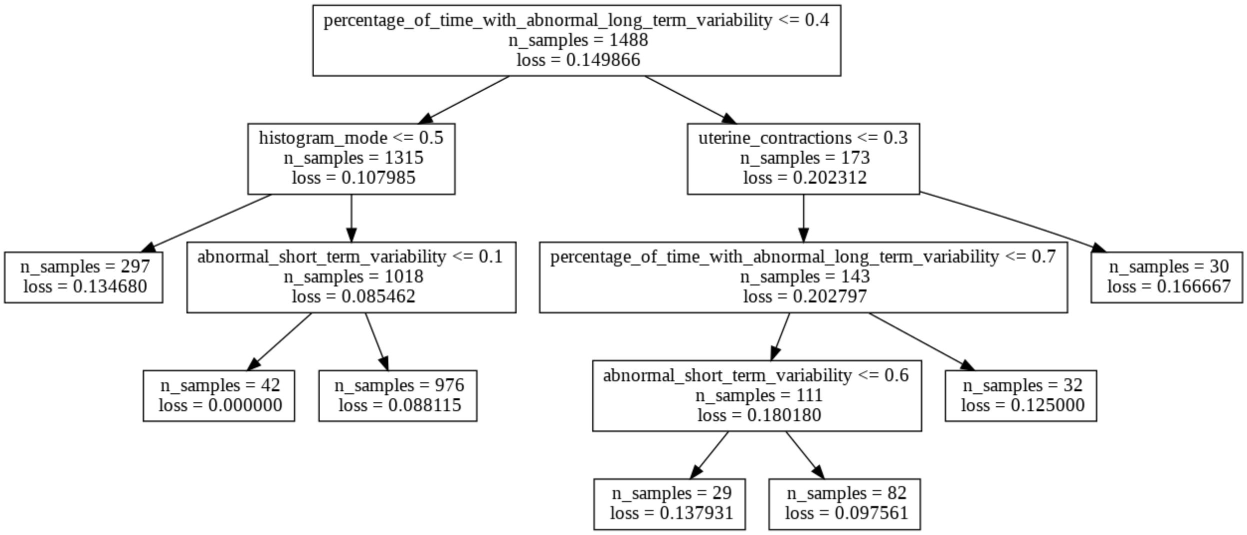
Dataset 1: double attribute


First sexual intercourse <= 32.0 & Schiller <= 0.0 n_samples = 527 loss = 0.034156
Age <= 20.0 & Number of sexual partners <= 2.0 n_samples = 479 loss = 0.012526
First sexual intercourse <= 18.0 & STDs (number) <= 0.0 n_samples = 48 loss = 0.145833
Age <= 19.0 & Number of sexual partners <= 1.0 n_samples = 86 loss = 0.011628
Age <= 21.0 & Number of sexual partners <= 5.0 n_samples = 393 loss = 0.010178
First sexual intercourse <= 15.0 & Hinselmann <= 0.0 n_samples = 33 loss = 0.121212
n_samples = 15 loss = 0.000000
n_samples = 37 loss = 0.000000
n_samples = 49 loss = 0.020408
n_samples = 57 loss = 0.017544
Age <= 28.0 & Number of sexual partners <= 5.0 n_samples = 336 loss = 0.002976
n_samples = 8 loss = 0.000000
Age <= 17.0 & First sexual intercourse <= 17.0 n_samples = 25 loss = 0.080000
n_samples = 141 loss = 0.000000
n_samples = 195 loss = 0.000000
n_samples = 2 loss = 0.000000
Age <= 23.0 & First sexual intercourse <= 17.0 n_samples = 23 loss = 0.043478
n_samples = 5 loss = 0.000000
n_samples = 18 loss = 0.000000

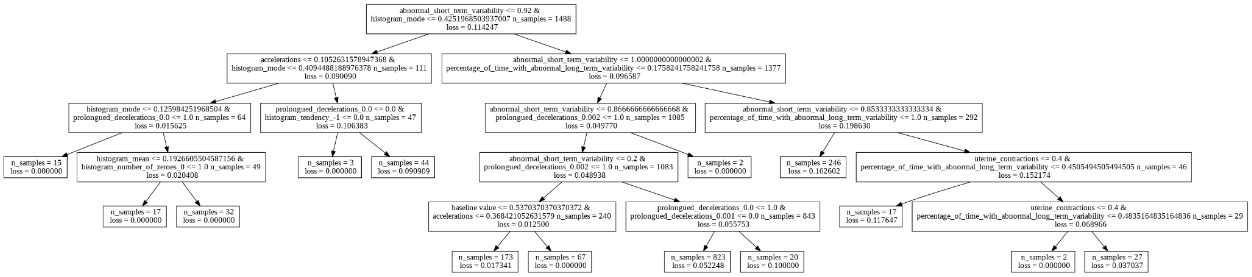
Dataset 2: Vanilla Decision Tree


mean_value_of_short_term_variability <= 0.051 gini = 0.373 samples = 1488 value = [1152, 217, 119]
percentage_of_time_with_abnormal_long_term_variability <= 0.747 gini = 0.593 samples = 259 value = [70, 142, 47]
histogram_mean <= 0.317 gini = 0.218 samples = 1229 value = [1082, 75, 72]
abnormal_short_term_variability <= 0.9 gini = 0.524 samples = 230 value = [68, 142, 20]
mean_value_of_short_term_variability <= 0.037 gini = 0.128 samples = 29 value = [2, 0, 27]
histogram_max <= 0.966 gini = 0.205 samples = 69 value = [8, 0, 61]
percentage_of_time_with_abnormal_long_term_variability <= 0.104 gini = 0.139 samples = 1160 value = [1074, 75, 11]
abnormal_short_term_variability <= 0.633 gini = 0.452 samples = 211 value = [65, 142, 4]
uterine_contractions <= 0.233 gini = 0.266 samples = 19 value = [3, 0, 16]
gini = 0.0 samples = 26 value = [0, 0, 26]
histogram_mode <= 0.626 gini = 0.444 samples = 3 value = [2, 0, 1]
histogram_max <= 0.168 gini = 0.061 samples = 63 value = [2, 0, 61]
gini = 0.0 samples = 6 value = [6, 0, 0]
prolongued_decelerations_0.003 <= 0.5 gini = 0.072 samples = 965 value = [929, 26, 10]
histogram_number_of_peaks <= 0.417 gini = 0.384 samples = 195 value = [145, 49, 1]
histogram_mode <= 0.724 gini = 0.395 samples = 48 value = [35, 13, 0]
percentage_of_time_with_abnormal_long_term_v gini = 0.339 samples = 163 value = [30, 129, 4]
gini = 0.0 samples = 15 value = [0, 0, 15]
baseline value <= 0.704 gini = 0.375 samples = 4 value = [3, 0, 1]
gini = 0.0 samples = 2 value = [2, 0, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.0 samples = 1 value = [1, 0, 0]
baseline value <= 0.167 gini = 0.032 samples = 62 value = [1, 0, 61]
accelerations <= 0.079 gini = 0.061 samples = 959 value = [929, 25, 5]
mean_value_of_long_term_variability <= 0.062 gini = 0.278 samples = 6 value = [0, 1, 5]
mean_value_of_long_term_variability <= 0.213 gini = 0.259 samples = 158 value = [134, 23, 1]
baseline value <= 0.463 gini = 0.418 samples = 37 value = [11, 26, 0]
gini = 0.137 samples = 27 value = [25, 2, 0]
gini = 0.499 samples = 21 value = [10, 11, 0]
gini = 0.447 samples = 29 value = [20, 8, 1]
gini = 0.179 samples = 134 value = [10, 121, 3]
gini = 0.0 samples = 3 value = [3, 0, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.32 samples = 5 value = [1, 0, 4]
gini = 0.0 samples = 57 value = [0, 0, 57]
gini = 0.165 samples = 291 value = [265, 21, 5]
gini = 0.012 samples = 668 value = [664, 4, 0]
gini = 0.0 samples = 5 value = [0, 0, 5]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.19 samples = 142 value = [127, 14, 1]
gini = 0.492 samples = 16 value = [7, 9, 0]
gini = 0.0 samples = 6 value = [6, 0, 0]
gini = 0.271 samples = 31 value = [5, 26, 0]

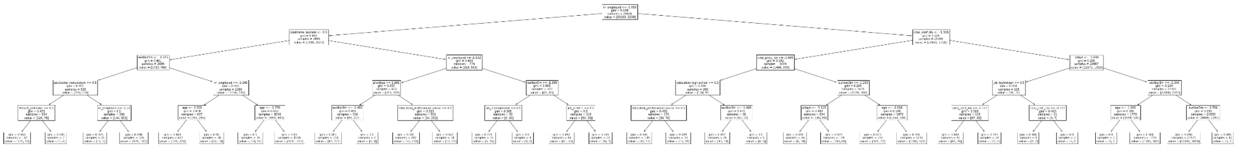
Dataset 2: single attribute



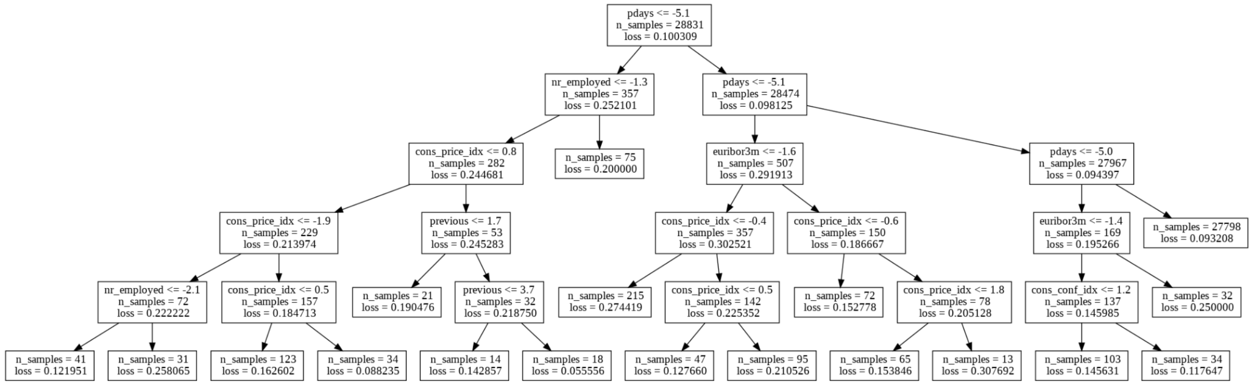
Dataset 2: double attribute


abnormal_short_term_variability <= 0.92 &
histogram_mode <= 0.4251968503937007 n_samples = 1488
loss = 0.114247
accelerations <= 0.1052631578947368 &
histogram_mode <= 0.4094488188976378 n_samples = 111
loss = 0.090090
abnormal_short_term_variability <= 1.0000000000000002 &
percentage_of_time_with_abnormal_long_term_variability <= 0.1758241758241758 n_samples = 1377
loss = 0.096587
histogram_mode <= 0.125984251968504 &
prolongued_decelerations_0.0 <= 1.0 n_samples = 64
loss = 0.015625
prolongued_decelerations_0.0 <= 0.0 &
histogram_tendency_-1 <= 0.0 n_samples = 47
loss = 0.106383
abnormal_short_term_variability <= 0.8666666666666668 &
prolongued_decelerations_0.002 <= 1.0 n_samples = 1085
loss = 0.049770
abnormal_short_term_variability <= 0.8533333333333334 &
percentage_of_time_with_abnormal_long_term_variability <= 1.0 n_samples = 292
loss = 0.198630
n_samples = 15
loss = 0.000000
histogram_mean <= 0.1926605504587156 &
histogram_number_of_zeroes_0 <= 1.0 n_samples = 49
loss = 0.020408
n_samples = 3
loss = 0.000000
n_samples = 44
loss = 0.090909
abnormal_short_term_variability <= 0.2 &
prolongued_decelerations_0.002 <= 1.0 n_samples = 1083
loss = 0.048938
n_samples = 2
loss = 0.000000
n_samples = 246
loss = 0.162602
uterine_contractions <= 0.4 &
percentage_of_time_with_abnormal_long_term_variability <= 0.4505494505494505 n_samples = 46
loss = 0.152174
n_samples = 17
loss = 0.000000
n_samples = 32
loss = 0.000000
baseline value <= 0.5370370370370372 &
accelerations <= 0.368421052631579 n_samples = 240
loss = 0.012500
prolongued_decelerations_0.0 <= 1.0 &
prolongued_decelerations_0.001 <= 0.0 n_samples = 843
loss = 0.055753
n_samples = 17
loss = 0.117647
uterine_contractions <= 0.4 &
percentage_of_time_with_abnormal_long_term_variability <= 0.4835164835164836 n_samples = 29
loss = 0.068966
n_samples = 173
loss = 0.017341
n_samples = 67
loss = 0.000000
n_samples = 823
loss = 0.052248
n_samples = 20
loss = 0.100000
n_samples = 2
loss = 0.000000
n_samples = 27
loss = 0.037037

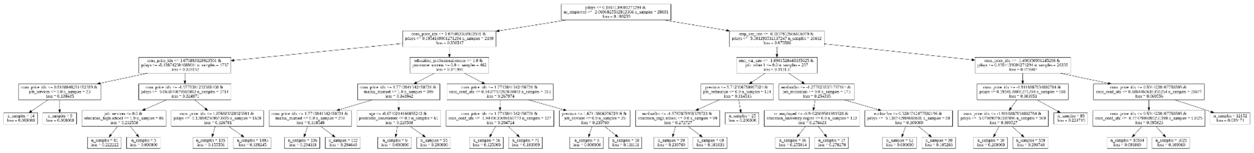
Dataset 3: Vanilla Decision Tree



Dataset 3: single attribute



Dataset 3: double attribute



Interpreting rules output

We use the visualizations of trees to compare and contrast the rules for the decision trees.
We make the following observations.

We find from the three models, that the results of Model 0 and Model 1 are comparable. But the results for Model 2 are slightly lower than both. The visualization also shows that the model for two attributes is showing near 0 loss for leaf nodes, which suggests overfitting of the data. This overfitting may be one of the reasons for the lower performance of Model 2.

Dataset 1

Schiller's research enables physicians to diagnose cervical cancer early, helping women receive treatment quicker and ultimately helping to popularize annual diagnostic exam (ref) - this is why Schiller is the leading cause in splitting root node in all three models (Model 0, 1 & 2).
The following factors play significant role as splitting criterion in the subsequent depths: age in years, number of sexual partners, first sexual intercourse (age in years), number of pregnancies, and hormonal contraceptives. These factors all contribute to information related to sexual health.

Studies say that multiple sexual partners and early onset of sexual activity are important factors linked to cervical cancer since it is found to spread through sexual intercourse (ref). Additionally, researches show that early sexual intercourse can be a risk factor for cervical cancer in young women, since it is a risk factor for HPV infection (ref).

Comparison with vanilla decision tree -

DTC with 1 attribute split: Depth of DTC with logistic regression models is significantly lesser than vanilla DTC since logistic regression model also classifies data. The split at the root node is similar as it uses the same attribute (Schiller) with similar threshold value and split ratio. DTC with logistic regression is more efficient in splitting in deeper nodes which results in lesser overfitting and better accuracy results than vanilla decision tree.

DTC with 2 attribute split: Depth of DTC with logistic regression models is comparable to vanilla DTC. The split at the root node is similar as it uses the same attribute (Schiller) with similar threshold value and split ratio. Further, they have similar splitting attribute values and splitting ratio. Leaf node models in DTC with logistic regression models can be seen having very less loss values indicating overfitting in some cases.

Dataset 2

When abnormal long term variability happens less than 40% of times, then the chances of a fetus surviving is higher. This is the first attribute we split on in a single attribute decision tree. However, the double attribute decision tree splits based on weather the fetus shows abnormal short term variability and even the vanilla decision tree splits on the same criteria. We also observe that the three models achieve similar accuracies with the vanilla decision tree having a depth of 6, the single split decision tree having a depth of 5 and the double split decision tree having a depth of 6.

Dataset 3.

This dataset is related to the direct marketing campaigns of a Portuguese banking institution based on phone calls, and the classification goal is to predict if the client will subscribe to a term deposit. We see that p_days, which is the number of days that passed by after the client was last contacted from a previous campaign, explains the frequent appearance of the attribute in the Logit Tree, while nr_employed is the number of employees in the campaign. The number of days since the last communication is an important indicator of the success of the call, as frequent calls may already have a bias toward the last communication, and fewer contacts during the campaign is a positive sign.

Nr.employees appears in all three models, which is an important economic context indicator and shows the economic state of society and is useful along with other economic indicators like the cons.price index which is the consumer price index. We note these attributes appear frequently in the tree. This shows the economic situation of society is an important indicator for deposits. This is also reflected in economics.

Comparison of Model 0 and Logit Tree.

We see that the depth of all trees is fixed using max_depth and all reach this height. We see that while Model 0 uses other economic indicators as the Logit Trees, p_days only appears once at the bottom of the tree. This shows the stark difference in calculating information gain in Gini and Logistic Regression.

Comparison of Model 1 and 2.

From the visualization of the two decision trees of one attribute and two attributes, we can see that the important variables, such as p_days and nr_employed appear at similar positions regardless of the model. In Model 1, p_days is the topmost split, and nr_employed is on the second level, while in Model 2, p_days and nr_employed appear in the same topmost node. But the threshold values are different, which can be due to different weights in the logistic regression model and the difference between one-variable and two-variable regression.

We notice this trend throughout the tree, where the rule attribute (in Model 1) appears in one higher node for Model 2, which may explain the overfitting.

We also see that some attributes which were completely missing in Model 1, make an appearance in Model 2, which shows that some variables were useful in combination with other variables.

Results

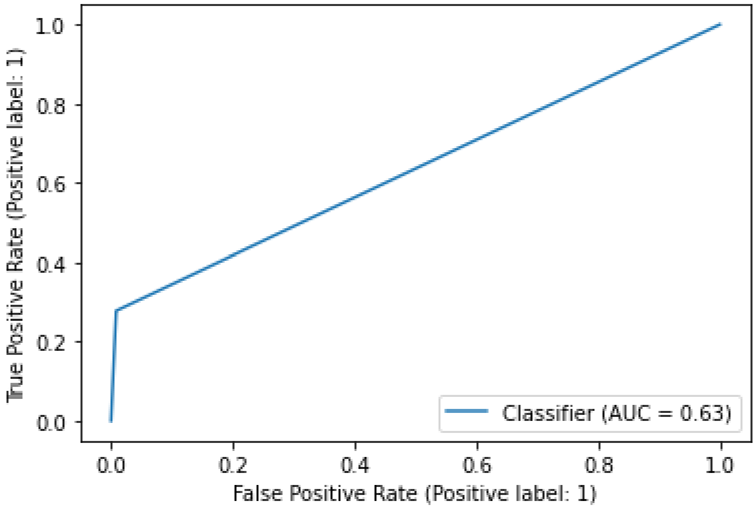
The evaluation results for 0.7 (train) and 0.3 (test) split on each dataset are as follows:

Assumptions.

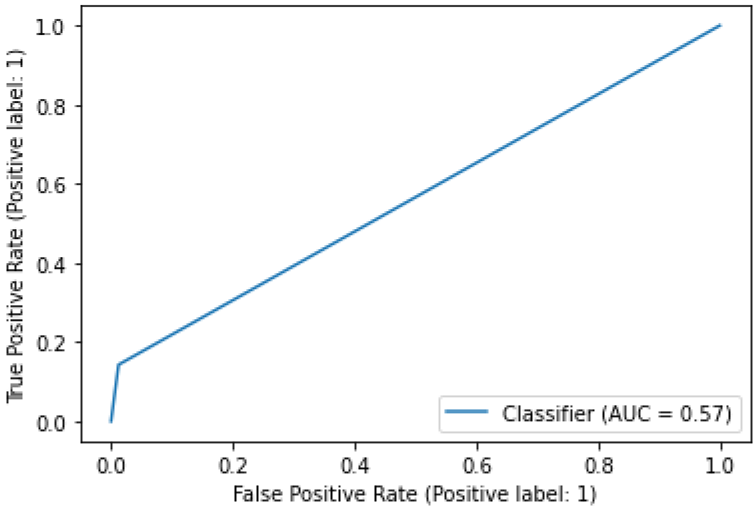
1. Model 0 is a Decision Tree from the sklearn library (Max Depth = 5)
2. Model 1 is Logit Tree with one attribute split (Max Depth = 5 and Min Leaf Nodes = 10)
3. Model 2 is Logits Tree with two attributes split (Max Depth = 5 and Min Leaf Nodes = 2)

Dataset 1

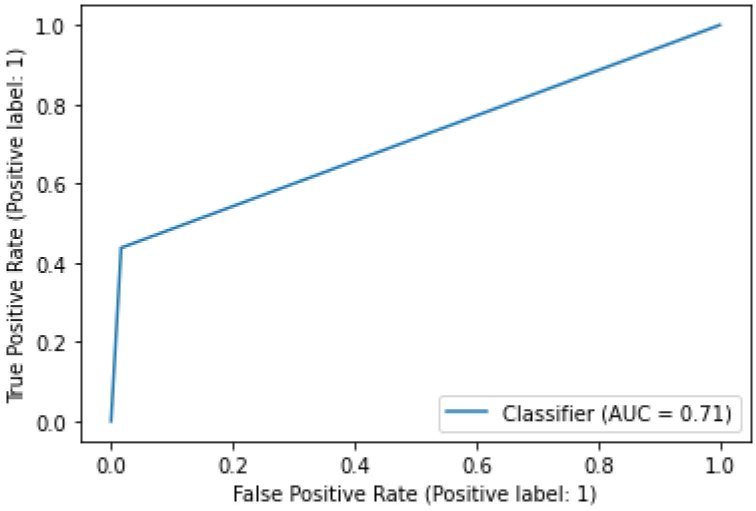
- Model 0
  Accuracy=0.9513274336283186,
  Precision=0.9590471345416687,
  Recall=0.9513274336283186



- Model 1
  Accuracy=0.9513274336283186,
  Precision=0.9557156972248475,
  Recall=0.9513274336283186



- Model 2
  Accuracy=0.915929203539823,
  Precision=0.9180656527814341,
  Recall=0.915929203539823



Dataset 2

- Model 0
  Accuracy=0.915929203539823,
  Precision=0.9032301364346734,
  Recall=0.915929203539823

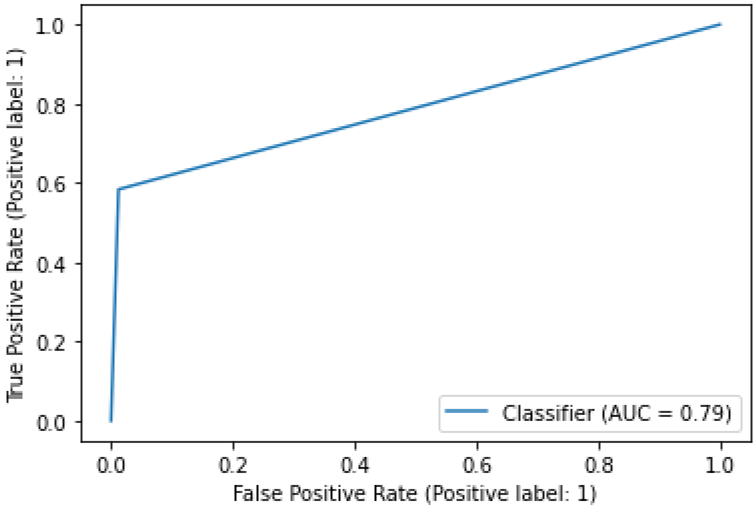
- Model 1
  Accuracy=0.8996865203761756,
  Precision=0.8957883233616476,

Recall=0.8996865203761756

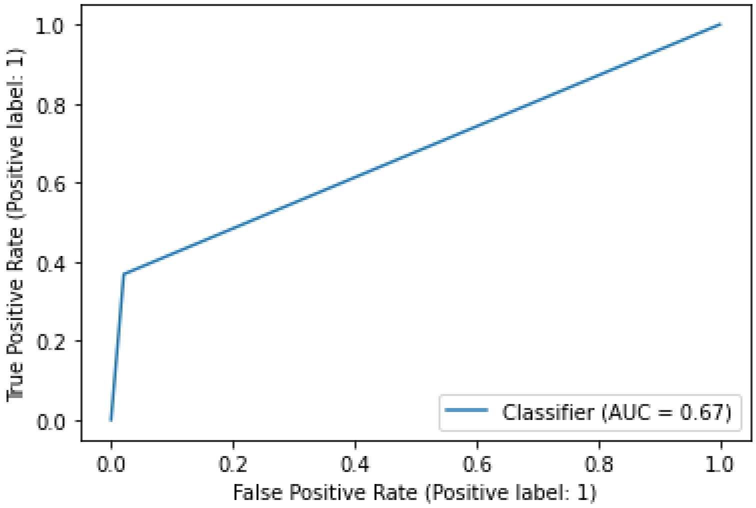
- Model 2
  Accuracy=0.890282131661442,
  Precision=0.8843458693073448,
  Recall=0.890282131661442

Dataset 3

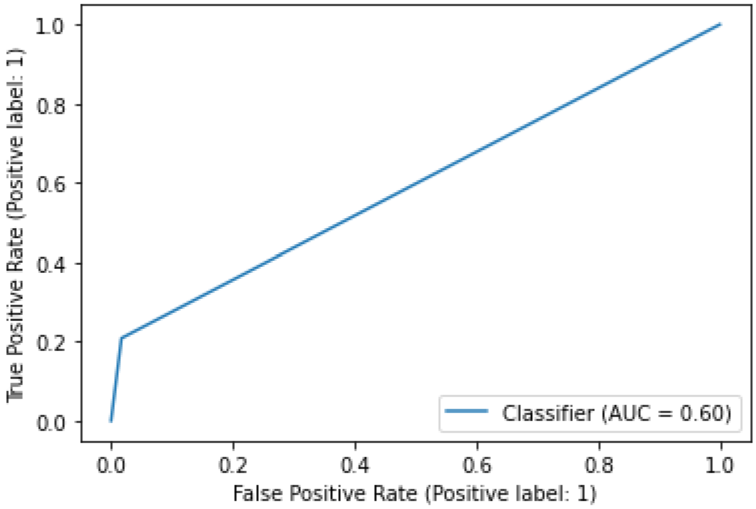
- Model 0
  Accuracy=0.9601769911504425,
  Precision=0.9592417226872965,
  Recall=0.9601769911504425



- Model 1
  Accuracy=0.9341085271317829,
  Precision=0.9241271191781686,
  Recall=0.9341085271317829



- Model 2
  Accuracy=0.8962531358744031,
  Precision=0.8756076381393768,
  Recall=0.8962531358744031



Confusion matrices

A confusion matrix is a summary of prediction results on a classification problem. Since all three datasets were classification problems, we made confusion matrices for them. The Confusion Matrix created has four different quadrants:

True Negative (Top-Left Quadrant), False Positive (Top-Right Quadrant),

False Negative (Bottom-Left Quadrant), and True Positive (Bottom-Right Quadrant)

The confusion matrices for each dataset and model were as follows:

Dataset 1

```
Confusion matrix for model 1 on dataset 1
 [[205   6]
 [  2  13]]

Confusion matrix for model 2 on dataset 1
 [[208   3]
 [  3  12]]
```

Dataset 2

```
Confusion matrix for model 1 on dataset 2
 [[480   6   2]
 [ 34  57   1]
 [  4  13  41]]
Confusion matrix for model 2 on dataset 2
 [[473  13   2]
 [ 28  62   2]
 [  0   7  51]]
```

Dataset 3

```
Confusion matrix for model 1 on dataset 3
 [[10843   143]
 [ 1019   352]]
Confusion matrix for model 2 on dataset 3
 [[10847   139]
 [ 1000   371]]
```

Cross-validation (Q3)

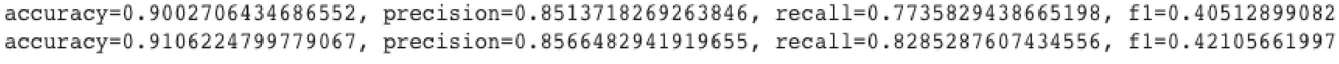
K-fold cross-validation is a method to evaluate which model works better. It consists of splitting the dataset into K parts (we have set K=5), training the model on K-1 parts, and testing the one leftover part. This process is repeated K times, taking each Kth part as the testing set.

The cross-validation accuracy, precision, recall, and f1 score for each dataset and model are as follows. Note that model 1 corresponds to the single attribute split and model 2 corresponds to the double attribute split.

Dataset 1
Model 1
cross-validation mean accuracy=0.942896247240618, precision=0.7967631505491032, recall=0.7878617761868284, f1=0.39508759387184356

Model 2
cross-validation mean accuracy=0.9428874172185431, precision=0.7693193266878412, recall=0.7764798150961909, f1=0.3862026193284925

Dataset 2



```
accuracy=0.9002706434686552, precision=0.8513718269263846, recall=0.7735829438665198, f1=0.40512899082
accuracy=0.9106224799779067, precision=0.8566482941919655, recall=0.8285287607434556, f1=0.42105661997
```

Dataset 3
Model 1
cross-validation mean accuracy=0.8973730350528732, precision=0.7766697391883491, recall=0.5940005971858753, f1=0.3364543807216242

Model 2
cross-validation mean accuracy=0.8970816668362895, precision=0.7666705377685022, recall=0.600057085820769, f1=0.3365914726349441

Statistical Tests

Assumptions:-

1. We are using Dataset 1 to train both decision trees to ensure the shortest training time.
2. We use accuracy as proportions to our population distribution.

We chose the statistical tests mentioned in the paper, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms as they cover the area of interest for our models.

1. Two Proportions Test

   Given two proportions from two populations, the null hypothesis is that the difference between the proportions has a mean equal to 0.

   Here,

   $$z = (p_a - p_b)/\sqrt{2p(1-p)/n}$$

   Proportions test output for all datasets.

   z statistic: 1.5151191043234233, p-value: 0.9351288777783295

   z statistic: 0.5478657417173515, p-value: 0.7081079532510551

   z statistic: 0.10439027873045763, p-value: 0.541570181508626

   But this test assumes the independence of samples, which is not true in our case, as the accuracies are calculated on the same test data.

2. Resampled Paired T-Test.

   Taking into account the variance of the test set, we can also use the resampled paired t-test. That is, we use the paired student t-test and the number of iterations as 5.

   Here,

   $$t = \frac{p\sqrt{n}}{\sqrt{\Sigma_{i=1}^{n}(p^{i} - p)^{2}/(n-1)}}$$

   $$\text{where } p^{i} = p^{i}_{A} - p^{i}_{B}$$

   For our models,

   t statistic: 0.3885143449429059, p-value: 0.717428359914488

   Here, due to time constraints, we could only train the model 5 times. Also, the assumption of normal distribution would not hold in our case, as there is an overlap between the training and testing sets.

3. Cross-Validated Paired T-Test
   Using a similar method as the previous one, we will use KFold here instead of the holdout test set. The formula is the same as above.
   For our models,
   t statistic: 0.0012708973635689395, p-value: 0.9990468272980619

   This solves the issue of overlap in the test sets, but we still face the issue of overlap in training data.

4. McNamar's Test
   Here, we create a contingency table and use the null hypothesis that both algorithms should have the same error rate. The contingency table contains the hit-and-miss count of both models.

	Model 2 Hit	Model 2 Miss
Model 1 Hit	A	B
Model 1 Miss	C	D

   We do a chi-squared test using the table with the formula,

$$X^{2} = \frac{((|b-c|)-1)^{2}}{b+c}$$

   For our models,
   chi² statistic: 1.4545454545454546 and p-value: 0.22779999398822554

5. 5x2 Cross Validation Test
   We run 2-fold cross-validation five times and generate ten estimations. Using the values obtained, we define the following proportions.

$$p^{1} = p^{1}_{A} - p^{1}_{B}$$
$$p^{2} = p^{2}_{A} - p^{2}_{B}$$
$$p = (p^{1} + p^{2})/2$$
$$s^{2} = (p^{1} - p)^{2} + (p^{2} - p)$$

$$t = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s^2_i / 5}}$$

For our models,
t statistic: 0.0, p-value: 1.0
Given the shortcomings of the other methods, we will base our hypothesis testing on the last two tests, which are also classified as the second best and best tests in the domain, respectively.

We consider the null hypothesis as both populations are the same. We observe that, with a confidence level of 95%, that p should be less than 0.05 for the result to be statistically significant. We realize that the p-value remains high in both tests.

Thus, we conclude that the null hypothesis is true. The models do not show any statistically significant differences, and the difference between the accuracies can be random.

6. Wilcoxon Test
   It is a non-parametric alternative to the t-test. Doesn't assume normal distribution and Outliers have less effect compared to the t-test. Used for comparison of models on all 3 datasets using two arrays where each array contains 3 values of accuracy achieved by the respective model on the 3 datasets.
   p-value = 0.25
   Thus, we conclude that the null hypothesis is true. The models do not show any statistically significant differences, and the difference between the accuracies can be random.

References

1. Statistical Tests - Wilcoxon
2. Statistical Tests - McNamar and 5x2 Cross-Validation Test.
3. https://graphviz.readthedocs.io/en/stable/manual.html - Graphviz documentaion (for visualization).
4. Darby, Alexis, "Walter Schiller (1887–1960)". Embryo Project Encyclopedia (2021-08-12). ISSN: 1940-5030 http://embryo.asu.edu/handle/10776/13305.
5. Early Age at First Sexual Intercourse is Associated with Higher Prevalence of High-grade Squamous Intraepithelial Lesions
6. Impact of cervical cancer on the sexual and physical health of women diagnosed with cervical cancer in Ghana: A qualitative phenomenological study
7. Decision Tree Classifier - sklearn