EECS738
Madhu Peduri (3006758)
pmspraju@ku.edu

Question 1:

Total number of records: 768

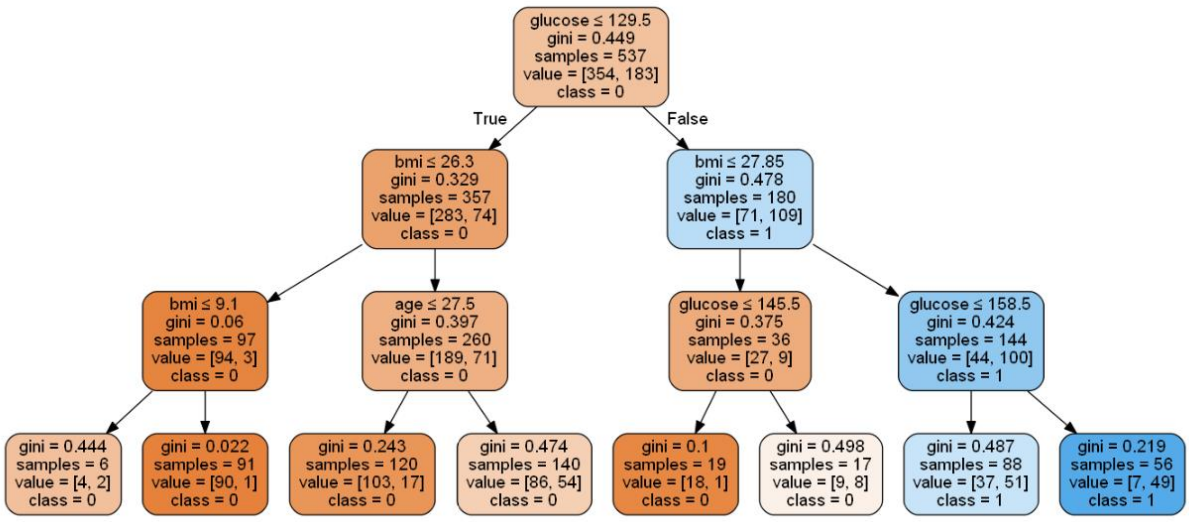
Total number of features: 9

Training set has 537 samples.

Testing set has 231 samples.

Times for Training, Prediction: 0.00174, 0.00122

Accuracy for Training, Test sets: 0.76350, 0.75758



Question 2:

Pre processing steps:

1) In 'Insurance fraud' dataset, we have few attribute 'Id' that is continuous and sequential. This is not useful to train our model. We can drop this attribute
2) Another attribute 'Insurance type' is same across the dataset. This is not useful to train our model. We can drop this attribute.
3) After few test run, I found attribute 'Marital status' is not contributing much to our model. We can drop this attribute. Inclusion of this attribute does not have effect on accuracy of the model, but omission of this is reducing the elapsed time.
4) We have another attribute 'Claim amount Received' is having high values with more variance. This makes it as continuous variable. Because of this model is getting trained on this attribute mainly and getting overfitted. So we used Kbins-discretizer to make it discrete.
5) We have attributes 'Income of policy holder' and 'Total Claimed amount' having high values and zeroes consistently. As zeros are outliers in this case and this makes this attribute skewed. So we use logarithmic transformation to make the outliers even.
6) We have another attribute 'Claim Amount' which has high values consistently which makes it outlying attribute compare to other attributes. We use minMaxscaler to make it uniform.
7) We have categorical and text attributes 'Injury Type', 'Overnight stay'. Since model deals with numeric value, we use one-hot encoding to change them to numeric attributes.

- In this case, we use max_depth =3 which is would sufficient given the number of features = 11. Increase in this parameter might result in overfitting.
- We use very less percentage of data to train. This might cause our model less exposure to the complexity of the data. This might cause underfitting.

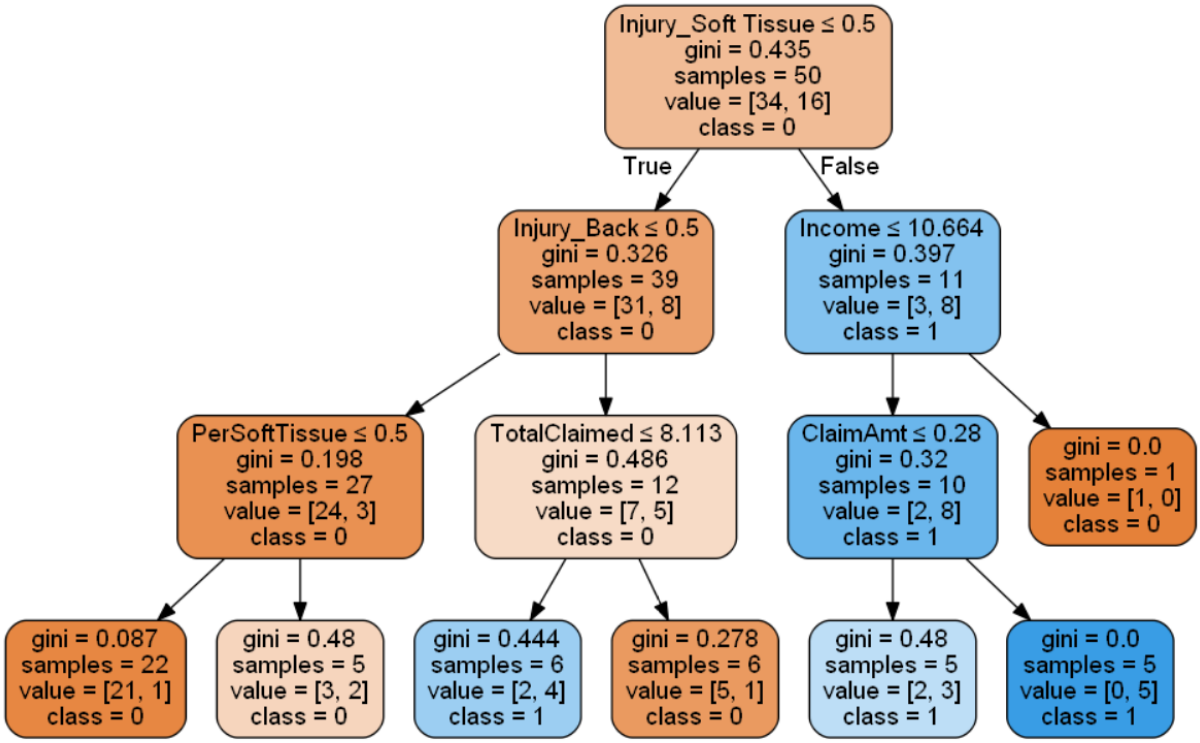
Train – 10% Test – 90% max_depth = 3 random_state = 1

Total number of records: 500
Total number of features: 14

---

14 total features after one-hot encoding.
Training set has 50 samples.
Testing set has 450 samples.

Times for Training, Prediction: 0.00275, 0.00132
Accuracy for Training, Test sets for Gini measure: 0.84000, 0.64444



Question 3:

1) Pre-processing steps will remain same.

Total number of records: 500
Total number of features: 14

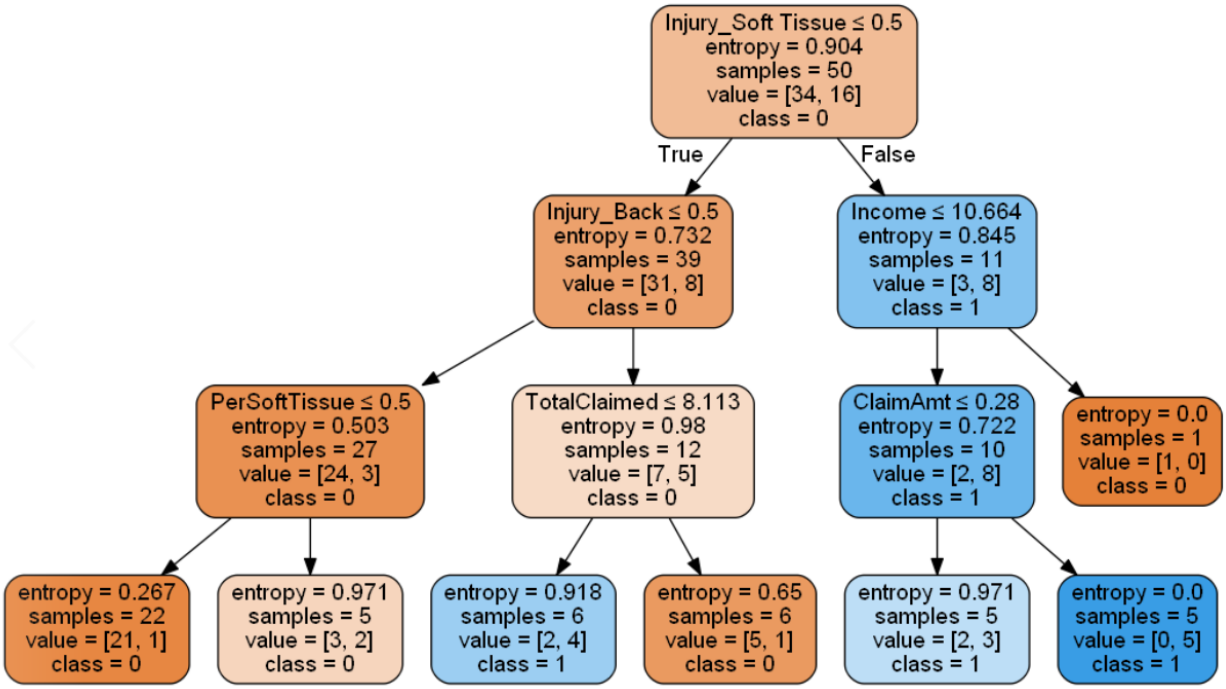
---

14 total features after one-hot encoding.
Training set has 50 samples.
Testing set has 450 samples.

Times for Training, Prediction: 0.00275, 0.00132
Accuracy for Training, Test sets: 0.84000, 0.64444

Times for Training, Prediction: 0.00250, 0.00128
Accuracy for Training, Test sets: 0.84000, 0.64444

- Accuracy did not change even if we change the criterion to ‘Entropy’.
- Tree is also did not change. Only there is slight improvement in the elapsed time.
- For the given parameters, I think both Gini and Entropy criterion perform same.



Question 4:

1) Pre-processing steps remain same

Total number of records: 500
Total number of features: 14

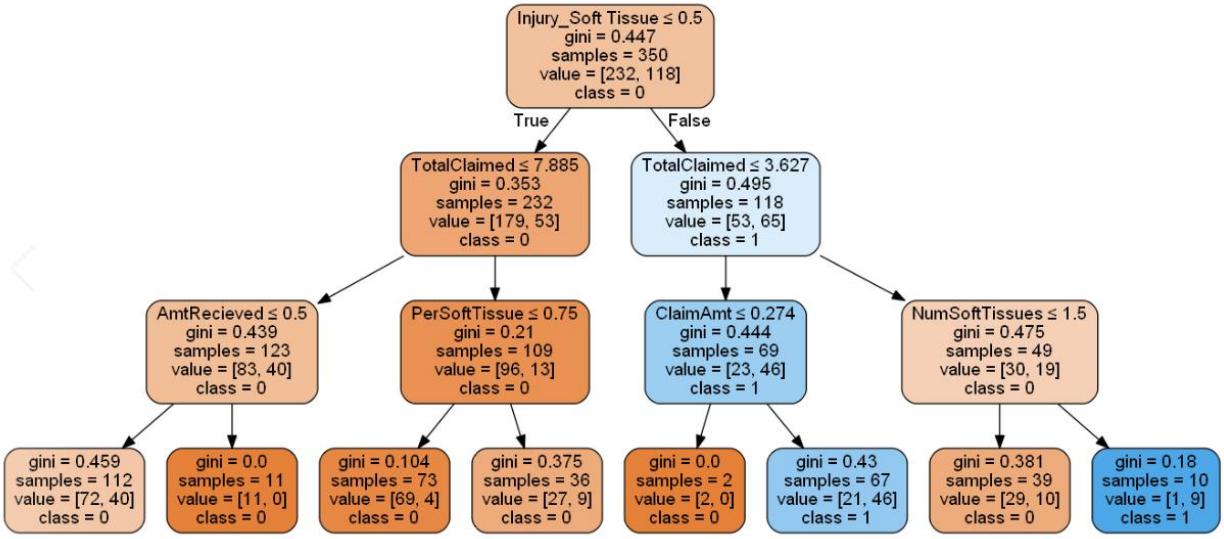
---

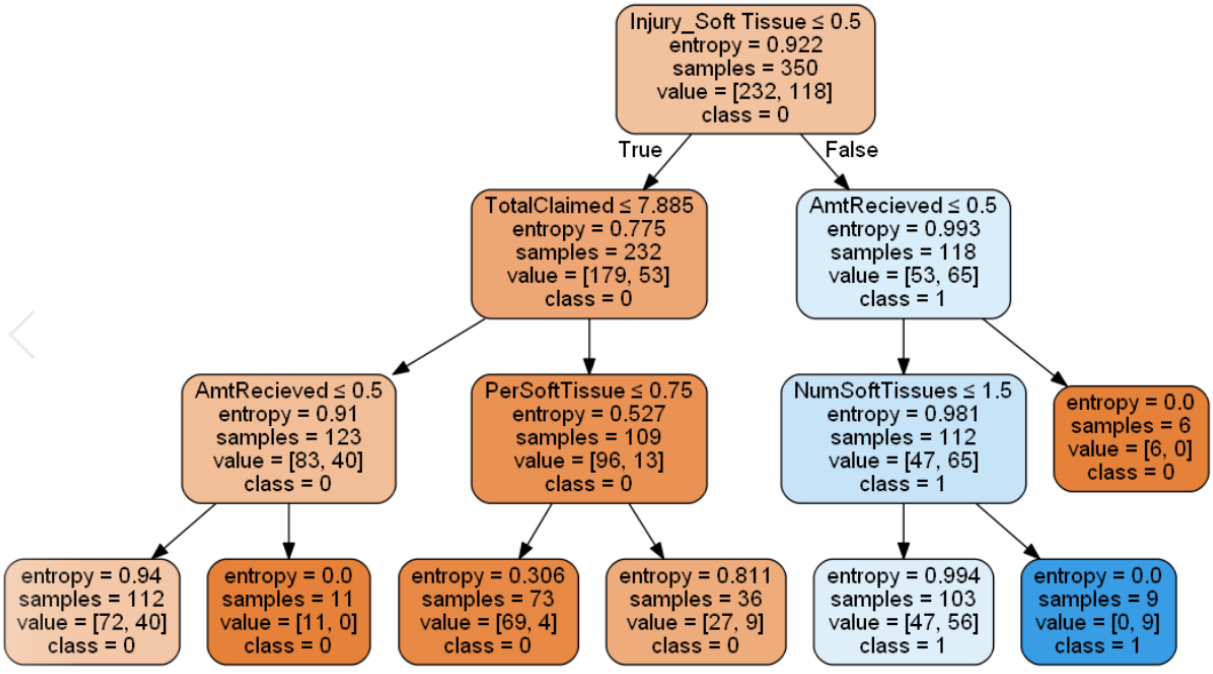
14 total features after one-hot encoding.
Training set has 350 samples.
Testing set has 150 samples.

Times for Training, Prediction: 0.00282, 0.00125
Accuracy for Training, Test sets: 0.75714, 0.72000

Times for Training, Prediction: 0.00292, 0.00129
Accuracy for Training, Test sets: 0.71429, 0.70000

- We can see an improvement in the performance. By increasing the training data for training. With less training data, there is a chance that model overfit on the training data and underfit for Test data.
- The split 70-30, between training and test, is better than the 10-90
- With these parameters, Gini measure performed well than the Entropy criteria.
- We can see the difference in the tree structures. Tree for entropy is evenly distributed than the one from Entropy method.


Injury_Soft Tissue ≤ 0.5
gini = 0.447
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
gini = 0.353
samples = 232
value = [179, 53]
class = 0
TotalClaimed ≤ 3.627
gini = 0.495
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
gini = 0.439
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
gini = 0.21
samples = 109
value = [96, 13]
class = 0
ClaimAmt ≤ 0.274
gini = 0.444
samples = 69
value = [23, 46]
class = 1
NumSoftTissues ≤ 1.5
gini = 0.475
samples = 49
value = [30, 19]
class = 0
gini = 0.459
samples = 112
value = [72, 40]
class = 0
gini = 0.0
samples = 11
value = [11, 0]
class = 0
gini = 0.104
samples = 73
value = [69, 4]
class = 0
gini = 0.375
samples = 36
value = [27, 9]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.43
samples = 67
value = [21, 46]
class = 1
gini = 0.381
samples = 39
value = [29, 10]
class = 0
gini = 0.18
samples = 10
value = [1, 9]
class = 1


Injury_Soft Tissue ≤ 0.5
entropy = 0.922
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
entropy = 0.775
samples = 232
value = [179, 53]
class = 0
AmtRecieved ≤ 0.5
entropy = 0.993
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
entropy = 0.91
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
entropy = 0.527
samples = 109
value = [96, 13]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.981
samples = 112
value = [47, 65]
class = 1
entropy = 0.0
samples = 6
value = [6, 0]
class = 0
entropy = 0.94
samples = 112
value = [72, 40]
class = 0
entropy = 0.0
samples = 11
value = [11, 0]
class = 0
entropy = 0.306
samples = 73
value = [69, 4]
class = 0
entropy = 0.811
samples = 36
value = [27, 9]
class = 0
entropy = 0.994
samples = 103
value = [47, 56]
class = 1
entropy = 0.0
samples = 9
value = [0, 9]
class = 1

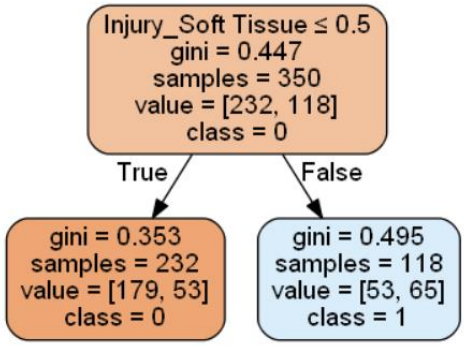
Question 5:

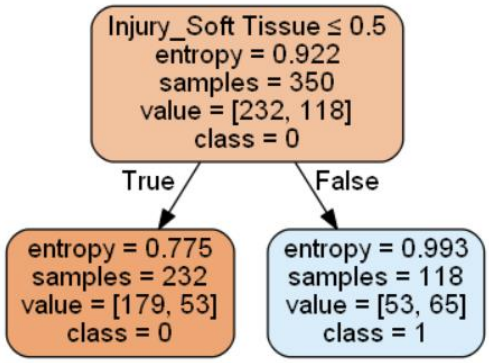
1) Pre-processing steps remain same.

Max_depth	1	2	3	4	5
Acc_Gini_Train	0.69714	0.72857	0.75714	0.77714	0.79143
Acc_Gini_Test	0.68	0.72667	0.72	0.67333	0.66667
Acc_Entr_Train	0.69714	0.71429	0.71429	0.76571	0.77714
Acc_Entr_Test	0.68	0.7	0.7	0.73333	0.71333

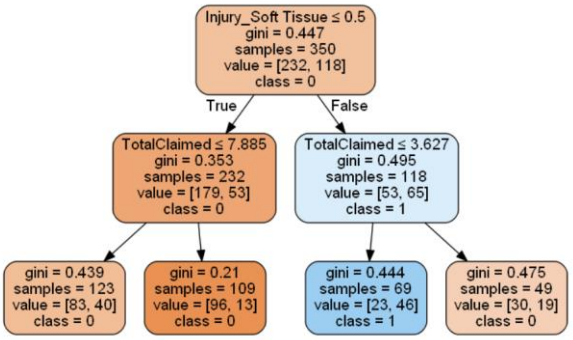
- We can see as max_depth increases, Training accuracy is also increasing
- Similarly, Test accuracy is increasing till max_depth = 2, then it is decreasing For Gini.
- For Entropy method, as max_depth increases, accuracies also increases.
- I think, if max_depth >3, model is overfitting.

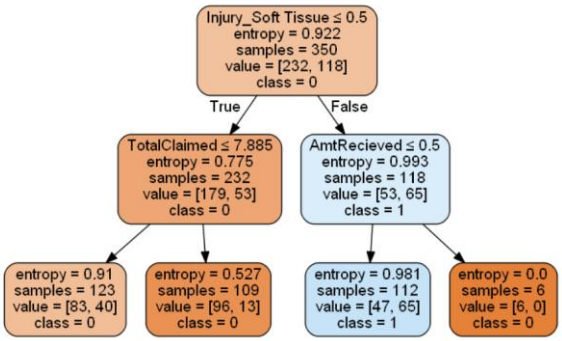
Max_depth = 1



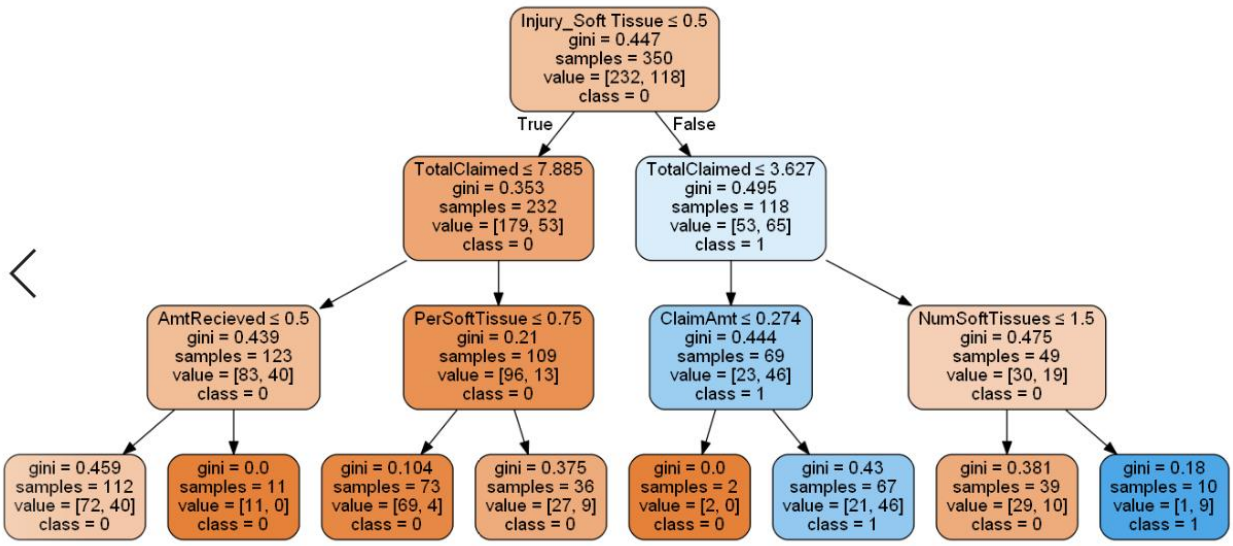


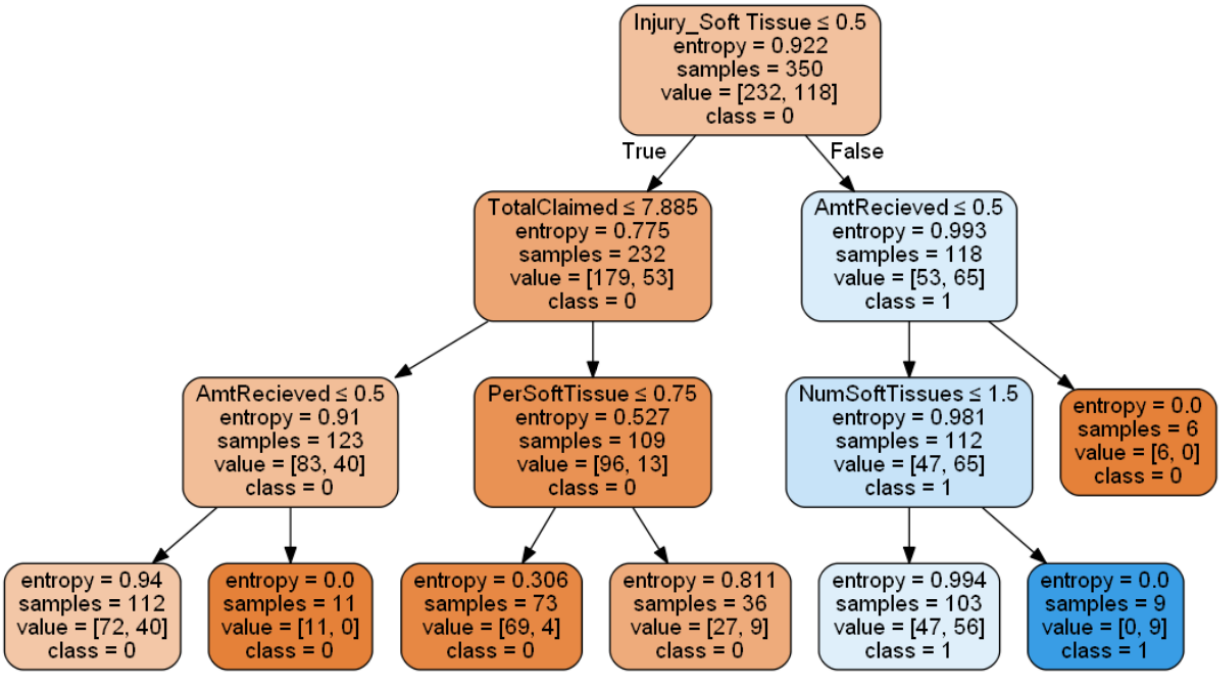
Max_depth = 2



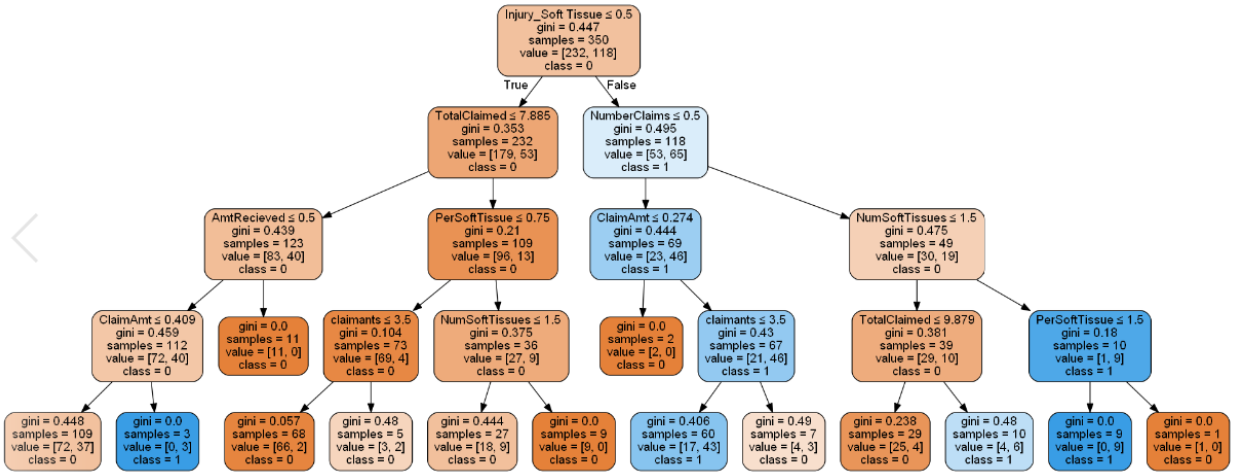


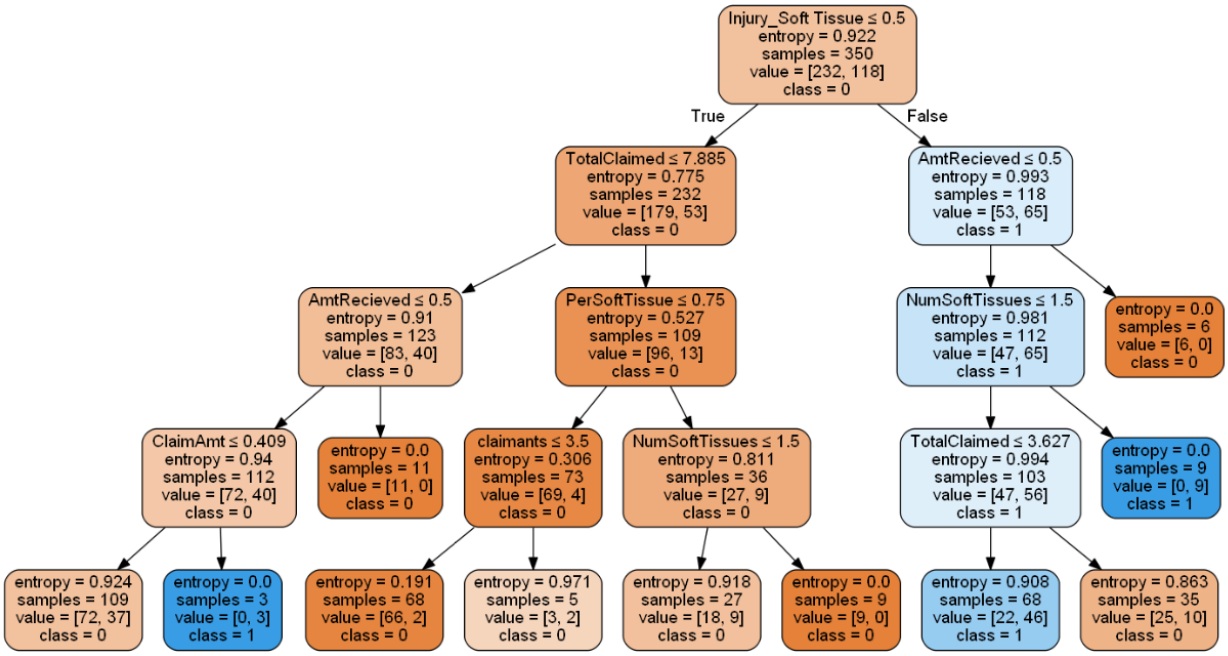
Max_depth = 3


Injury_Soft Tissue ≤ 0.5
gini = 0.447
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
gini = 0.353
samples = 232
value = [179, 53]
class = 0
TotalClaimed ≤ 3.627
gini = 0.495
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
gini = 0.439
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
gini = 0.21
samples = 109
value = [96, 13]
class = 0
ClaimAmt ≤ 0.274
gini = 0.444
samples = 69
value = [23, 46]
class = 1
NumSoftTissues ≤ 1.5
gini = 0.475
samples = 49
value = [30, 19]
class = 0
gini = 0.459
samples = 112
value = [72, 40]
class = 0
gini = 0.0
samples = 11
value = [11, 0]
class = 0
gini = 0.104
samples = 73
value = [69, 4]
class = 0
gini = 0.375
samples = 36
value = [27, 9]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.43
samples = 67
value = [21, 46]
class = 1
gini = 0.381
samples = 39
value = [29, 10]
class = 0
gini = 0.18
samples = 10
value = [1, 9]
class = 1

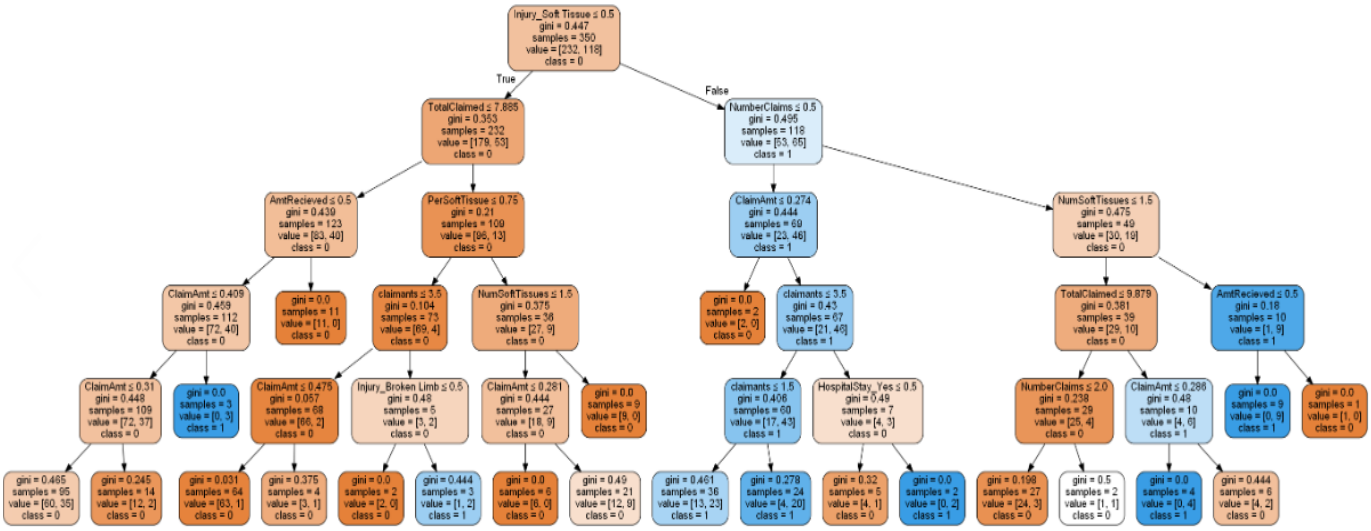

Injury_Soft Tissue ≤ 0.5
entropy = 0.922
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
entropy = 0.775
samples = 232
value = [179, 53]
class = 0
AmtRecieved ≤ 0.5
entropy = 0.993
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
entropy = 0.91
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
entropy = 0.527
samples = 109
value = [96, 13]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.981
samples = 112
value = [47, 65]
class = 1
entropy = 0.0
samples = 6
value = [6, 0]
class = 0
entropy = 0.94
samples = 112
value = [72, 40]
class = 0
entropy = 0.0
samples = 11
value = [11, 0]
class = 0
entropy = 0.306
samples = 73
value = [69, 4]
class = 0
entropy = 0.811
samples = 36
value = [27, 9]
class = 0
entropy = 0.994
samples = 103
value = [47, 56]
class = 1
entropy = 0.0
samples = 9
value = [0, 9]
class = 1

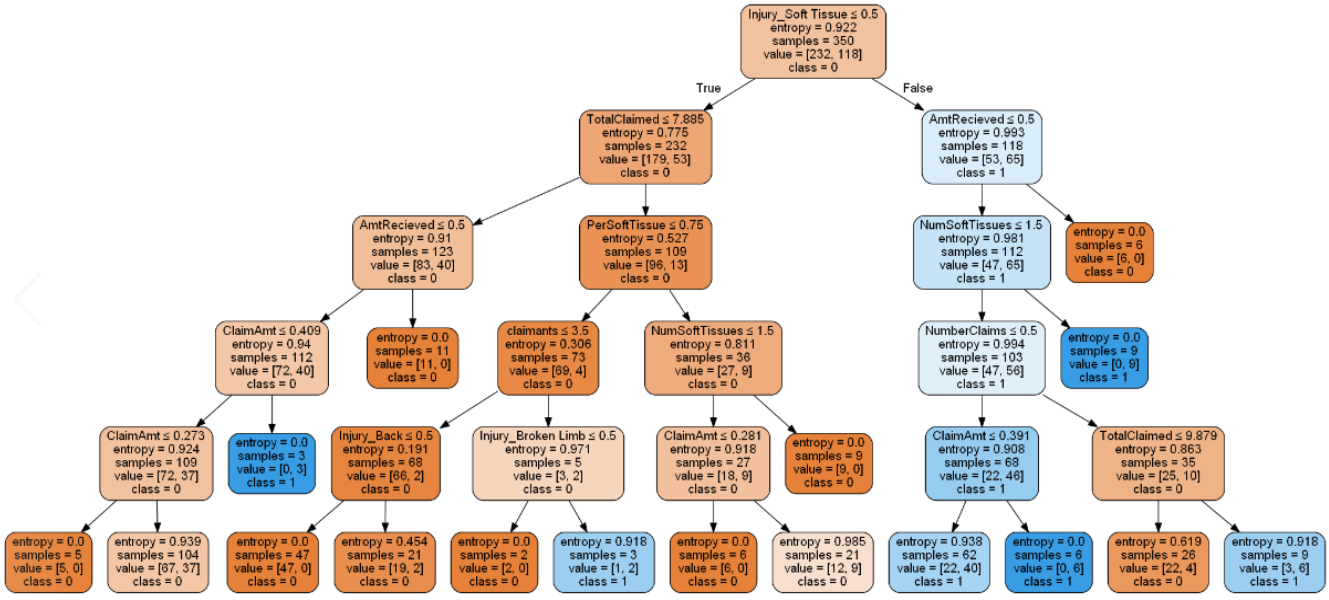
Max_depth = 4


Injury_Soft Tissue ≤ 0.5
gini = 0.447
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
gini = 0.353
samples = 232
value = [179, 53]
class = 0
NumberClaims ≤ 0.5
gini = 0.495
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
gini = 0.439
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
gini = 0.21
samples = 109
value = [96, 13]
class = 0
ClaimAmt ≤ 0.274
gini = 0.444
samples = 69
value = [23, 46]
class = 1
NumSoftTissues ≤ 1.5
gini = 0.475
samples = 49
value = [30, 19]
class = 0
ClaimAmt ≤ 0.409
gini = 0.459
samples = 112
value = [72, 40]
class = 0
gini = 0.0
samples = 11
value = [11, 0]
class = 0
claimants ≤ 3.5
gini = 0.104
samples = 73
value = [69, 4]
class = 0
NumSoftTissues ≤ 1.5
gini = 0.375
samples = 36
value = [27, 9]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
claimants ≤ 3.5
gini = 0.43
samples = 67
value = [21, 46]
class = 1
TotalClaimed ≤ 9.879
gini = 0.381
samples = 39
value = [29, 10]
class = 0
PerSoftTissue ≤ 1.5
gini = 0.18
samples = 10
value = [1, 9]
class = 1
gini = 0.448
samples = 109
value = [72, 37]
class = 0
gini = 0.0
samples = 3
value = [0, 3]
class = 1
gini = 0.057
samples = 68
value = [66, 2]
class = 0
gini = 0.48
samples = 5
value = [3, 2]
class = 0
gini = 0.444
samples = 27
value = [18, 9]
class = 0
gini = 0.0
samples = 9
value = [9, 0]
class = 0
gini = 0.406
samples = 60
value = [17, 43]
class = 1
gini = 0.49
samples = 7
value = [4, 3]
class = 0
gini = 0.238
samples = 29
value = [25, 4]
class = 0
gini = 0.48
samples = 10
value = [4, 6]
class = 1
gini = 0.0
samples = 9
value = [0, 9]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0


Injury_Soft Tissue ≤ 0.5
entropy = 0.922
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
entropy = 0.775
samples = 232
value = [179, 53]
class = 0
AmtRecieved ≤ 0.5
entropy = 0.993
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
entropy = 0.91
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
entropy = 0.527
samples = 109
value = [96, 13]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.981
samples = 112
value = [47, 65]
class = 1
entropy = 0.0
samples = 6
value = [6, 0]
class = 0
ClaimAmt ≤ 0.409
entropy = 0.94
samples = 112
value = [72, 40]
class = 0
entropy = 0.0
samples = 11
value = [11, 0]
class = 0
claimants ≤ 3.5
entropy = 0.306
samples = 73
value = [69, 4]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.811
samples = 36
value = [27, 9]
class = 0
TotalClaimed ≤ 3.627
entropy = 0.994
samples = 103
value = [47, 56]
class = 1
entropy = 0.0
samples = 9
value = [0, 9]
class = 1
entropy = 0.924
samples = 109
value = [72, 37]
class = 0
entropy = 0.0
samples = 3
value = [0, 3]
class = 1
entropy = 0.191
samples = 68
value = [66, 2]
class = 0
entropy = 0.971
samples = 5
value = [3, 2]
class = 0
entropy = 0.918
samples = 27
value = [18, 9]
class = 0
entropy = 0.0
samples = 9
value = [9, 0]
class = 0
entropy = 0.908
samples = 68
value = [22, 46]
class = 1
entropy = 0.863
samples = 35
value = [25, 10]
class = 0

Max_depth = 5


Injury_Soft Tissue ≤ 0.5
gini = 0.447
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
gini = 0.353
samples = 232
value = [179, 53]
class = 0
NumberClaims ≤ 0.5
gini = 0.495
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
gini = 0.439
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
gini = 0.21
samples = 109
value = [96, 13]
class = 0
ClaimAmt ≤ 0.274
gini = 0.444
samples = 69
value = [23, 46]
class = 1
NumSoftTissues ≤ 1.5
gini = 0.475
samples = 49
value = [30, 19]
class = 0
ClaimAmt ≤ 0.409
gini = 0.459
samples = 112
value = [72, 40]
class = 0
gini = 0.0
samples = 11
value = [11, 0]
class = 0
claimants ≤ 3.5
gini = 0.104
samples = 73
value = [69, 4]
class = 0
NumSoftTissues ≤ 1.5
gini = 0.375
samples = 36
value = [27, 9]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
claimants ≤ 3.5
gini = 0.43
samples = 67
value = [21, 46]
class = 1
TotalClaimed ≤ 9.879
gini = 0.381
samples = 39
value = [29, 10]
class = 0
AmtRecieved ≤ 0.5
gini = 0.18
samples = 10
value = [1, 9]
class = 1
ClaimAmt ≤ 0.31
gini = 0.448
samples = 109
value = [72, 37]
class = 0
gini = 0.0
samples = 3
value = [0, 3]
class = 1
ClaimAmt ≤ 0.475
gini = 0.057
samples = 68
value = [66, 2]
class = 0
Injury_Broken Limb ≤ 0.5
gini = 0.48
samples = 5
value = [3, 2]
class = 0
ClaimAmt ≤ 0.281
gini = 0.444
samples = 27
value = [18, 9]
class = 0
gini = 0.0
samples = 9
value = [9, 0]
class = 0
claimants ≤ 1.5
gini = 0.406
samples = 60
value = [17, 43]
class = 1
HospitalStay_Yes ≤ 0.5
gini = 0.49
samples = 7
value = [4, 3]
class = 0
NumberClaims ≤ 2.0
gini = 0.238
samples = 29
value = [25, 4]
class = 0
ClaimAmt ≤ 0.286
gini = 0.48
samples = 10
value = [4, 6]
class = 1
gini = 0.0
samples = 9
value = [0, 9]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.465
samples = 95
value = [60, 35]
class = 0
gini = 0.245
samples = 14
value = [12, 2]
class = 0
gini = 0.031
samples = 64
value = [63, 1]
class = 0
gini = 0.375
samples = 4
value = [3, 1]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.444
samples = 3
value = [1, 2]
class = 1
gini = 0.0
samples = 6
value = [6, 0]
class = 0
gini = 0.49
samples = 21
value = [12, 9]
class = 0
gini = 0.461
samples = 36
value = [13, 23]
class = 1
gini = 0.278
samples = 24
value = [4, 20]
class = 1
gini = 0.32
samples = 5
value = [4, 1]
class = 0
gini = 0.0
samples = 2
value = [0, 2]
class = 1
gini = 0.198
samples = 27
value = [24, 3]
class = 0
gini = 0.5
samples = 2
value = [1, 1]
class = 0
gini = 0.0
samples = 4
value = [0, 4]
class = 1
gini = 0.444
samples = 6
value = [4, 2]
class = 0


Injury_Soft Tissue ≤ 0.5
entropy = 0.922
samples = 350
value = [232, 118]
class = 0
True
False
TotalClaimed ≤ 7.885
entropy = 0.775
samples = 232
value = [179, 53]
class = 0
AmtRecieved ≤ 0.5
entropy = 0.993
samples = 118
value = [53, 65]
class = 1
AmtRecieved ≤ 0.5
entropy = 0.91
samples = 123
value = [83, 40]
class = 0
PerSoftTissue ≤ 0.75
entropy = 0.527
samples = 109
value = [96, 13]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.981
samples = 112
value = [47, 65]
class = 1
entropy = 0.0
samples = 6
value = [6, 0]
class = 0
ClaimAmt ≤ 0.409
entropy = 0.94
samples = 112
value = [72, 40]
class = 0
entropy = 0.0
samples = 11
value = [11, 0]
class = 0
claimants ≤ 3.5
entropy = 0.306
samples = 73
value = [69, 4]
class = 0
NumSoftTissues ≤ 1.5
entropy = 0.811
samples = 36
value = [27, 9]
class = 0
NumberClaims ≤ 0.5
entropy = 0.994
samples = 103
value = [47, 56]
class = 1
entropy = 0.0
samples = 9
value = [0, 9]
class = 1
ClaimAmt ≤ 0.273
entropy = 0.924
samples = 109
value = [72, 37]
class = 0
entropy = 0.0
samples = 3
value = [0, 3]
class = 1
Injury_Back ≤ 0.5
entropy = 0.191
samples = 68
value = [66, 2]
class = 0
Injury_Broken Limb ≤ 0.5
entropy = 0.971
samples = 5
value = [3, 2]
class = 0
ClaimAmt ≤ 0.281
entropy = 0.918
samples = 27
value = [18, 9]
class = 0
entropy = 0.0
samples = 9
value = [9, 0]
class = 0
ClaimAmt ≤ 0.391
entropy = 0.908
samples = 68
value = [22, 46]
class = 1
TotalClaimed ≤ 9.879
entropy = 0.863
samples = 35
value = [25, 10]
class = 0
entropy = 0.0
samples = 5
value = [5, 0]
class = 0
entropy = 0.939
samples = 104
value = [67, 37]
class = 0
entropy = 0.0
samples = 47
value = [47, 0]
class = 0
entropy = 0.454
samples = 21
value = [19, 2]
class = 0
entropy = 0.0
samples = 2
value = [2, 0]
class = 0
entropy = 0.918
samples = 3
value = [1, 2]
class = 1
entropy = 0.0
samples = 6
value = [6, 0]
class = 0
entropy = 0.985
samples = 21
value = [12, 9]
class = 0
entropy = 0.938
samples = 62
value = [22, 40]
class = 1
entropy = 0.0
samples = 6
value = [0, 6]
class = 1
entropy = 0.619
samples = 26
value = [22, 4]
class = 0
entropy = 0.918
samples = 9
value = [3, 6]
class = 1

Question 6:

1) If we include all features, 'Claim Amount Received' has most entropy. But as it looked like a continuous variable and causing overfit.
2) After omitting above feature, 'Injury Type' has more entropy.

Question 7:

1) Dataset consists of some discrepancies. We have few attributes with below drawbacks
    a. Columns having missing data
    b. Columns having string data
    c. Columns with skewed data
    d. Columns that are continuous
2) Decision tree is sensitive to data with above properties. We can tune the overfit behavior by pre-processing above columns.
3) Tuning parameters of the Decision tree model are important. Parameters like, Max_depth and random_state will tune the model to suit the data better.
4) Splitting of Training and test data plays important role. By using less training data, model can overfit the training data and underfit on test data.
5) Accuracy does not represent all aspects of performances. Increase in accuracy, sometimes, points to overfitting of the data.

Question 8:

Yes. We have skewed features like 'Income of Policy holder' , 'Total Claimed amount' and 'Claim amount received. We have outliers (mostly zeros) in these features.

a. Either we can use a mean or median value of the feature and replace the zero with that or doing some imputation.
b. We can use clamp transformation using upper and lower threshold values.

Question 9:

Yes. I think.

Question 10:

1) I do not see any problems in the mentioned paper.
2) I think, having different amount of data will result in to two different trees, with in a given set of model parameters, only if domain knowledge of the dataset changes.
3) Maybe we can use the statistical measures that are sensitive to the changes happen in a feature. Like variance in both the datasets for a given feature.