**Decision Tree Classification – Lab 06**

**Q1. What is the data about? (10’)**

The dataset is about diabetes, specifically focusing on predicting whether a patient has diabetes based on various diagnostic measurements. The attributes include:

* preg : Number of times pregnant
* plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* pres: Diastolic blood pressure (mm Hg)
* skin: Triceps skin fold thickness (mm)
* insu: 2-Hour serum insulin (mm U/ml)
* mass: Body mass index (weight in kg/(height in m)^2)
* pedi: Diabetes pedigree function
* age: Age (years)
* class: Class variable (tested\_positive or tested\_negative)

The goal from the dataset is to predict whether an individual has tested positive or negative for diabetes based on provided features (binary classification problem)

**Q2. Run a classification with an unpruned tree setting. Show the tree structure. What are the other settings you used in the process, e.g. minimum leaf size, cross-validation, etc. (30’)**

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -U -M 2*

*Relation: pima\_diabetes*

*Instances: 768*

*Attributes: 9*

*preg*

*plas*

*pres*

*skin*

*insu*

*mass*

*pedi*

*age*

*class*

*Test mode: 10-fold cross-validation*

*=== Classifier model (full training set) ===*

*J48 unpruned tree*

*------------------*

*plas <= 127*

*| mass <= 26.4*

*| | preg <= 7: tested\_negative (117.0/1.0)*

*| | preg > 7*

*| | | mass <= 0: tested\_positive (2.0)*

*| | | mass > 0: tested\_negative (13.0)*

*| mass > 26.4*

*| | age <= 28: tested\_negative (180.0/22.0)*

*| | age > 28*

*| | | plas <= 99: tested\_negative (55.0/10.0)*

*| | | plas > 99*

*| | | | pedi <= 0.56: tested\_negative (84.0/34.0)*

*| | | | pedi > 0.56*

*| | | | | preg <= 6*

*| | | | | | age <= 30: tested\_positive (4.0)*

*| | | | | | age > 30*

*| | | | | | | age <= 34: tested\_negative (7.0/1.0)*

*| | | | | | | age > 34*

*| | | | | | | | mass <= 33.1: tested\_positive (6.0)*

*| | | | | | | | mass > 33.1: tested\_negative (4.0/1.0)*

*| | | | | preg > 6: tested\_positive (13.0)*

*plas > 127*

*| mass <= 29.9*

*| | plas <= 145: tested\_negative (41.0/6.0)*

*| | plas > 145*

*| | | age <= 25: tested\_negative (4.0)*

*| | | age > 25*

*| | | | age <= 61*

*| | | | | mass <= 27.1: tested\_positive (12.0/1.0)*

*| | | | | mass > 27.1*

*| | | | | | pres <= 82*

*| | | | | | | pedi <= 0.396: tested\_positive (8.0/1.0)*

*| | | | | | | pedi > 0.396: tested\_negative (3.0)*

*| | | | | | pres > 82: tested\_negative (4.0)*

*| | | | age > 61: tested\_negative (4.0)*

*| mass > 29.9*

*| | plas <= 157*

*| | | pres <= 61: tested\_positive (15.0/1.0)*

*| | | pres > 61*

*| | | | age <= 30: tested\_negative (40.0/13.0)*

*| | | | age > 30: tested\_positive (60.0/17.0)*

*| | plas > 157: tested\_positive (92.0/12.0)*

*Number of Leaves : 22*

*Size of the tree : 43*

*Time taken to build model: 0.01 seconds*

*=== Stratified cross-validation ===*

*=== Summary ===*

*Correctly Classified Instances 558 72.6563 %*

*Incorrectly Classified Instances 210 27.3438 %*

*Kappa statistic 0.3992*

*Mean absolute error 0.3146*

*Root mean squared error 0.4544*

*Relative absolute error 69.2099 %*

*Root relative squared error 95.3354 %*

*Total Number of Instances 768*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0.788 0.388 0.791 0.788 0.790 0.399 0.736 0.790 tested\_negative*

*0.612 0.212 0.607 0.612 0.610 0.399 0.736 0.556 tested\_positive*

*Weighted Avg. 0.727 0.327 0.727 0.727 0.727 0.399 0.736 0.708*

*=== Confusion Matrix ===*

*a b <-- classified as*

*394 106 | a = tested\_negative*

*104 164 | b = tested\_positive*

**A screenshot of a computer

Description automatically generated**

A diagram of a tree

Description automatically generated

**Q3. What is the error rate? (10’)**

Correctly Classified Instances 558 72.6563 %

Incorrectly Classified Instances 210 27.3438 %

Error Rate = 27.3438 %

**Q4. Explain the confusion matrix.(10’)**

=== Confusion Matrix ===

a b <-- classified as

394 106 | a = tested\_negative

104 164 | b = tested\_positive

* In the top-left quadrant, there are 394 true negatives (TN), which are instances correctly predicted as not having diabetes (tested\_negative).
* In the top-right, there 106 false positives (FP), which are instances incorrectly predicted as having diabetes (tested\_positive) when they do not.
* In the bottom-left, there are 104 false negatives (FN), which are instances incorrectly predicted as not having diabetes (tested\_negative) when they do.
* In the bottom right quadrant, there are 164 true positives (TP), which are instances correctly predicted as having diabetes (tested\_positive).

**Q5. Do at least three more runs, such as a pruned version, a version with different setting on leaf size limit, and different setting on percentage split. Explain the differences between these models. (20’)**

**Pruned Version**

**A screenshot of a computer

Description automatically generated**

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2*

*Relation: pima\_diabetes*

*Instances: 768*

*Attributes: 9*

*preg*

*plas*

*pres*

*skin*

*insu*

*mass*

*pedi*

*age*

*class*

*Test mode: 10-fold cross-validation*

*=== Classifier model (full training set) ===*

*J48 pruned tree*

*------------------*

*plas <= 127*

*| mass <= 26.4: tested\_negative (132.0/3.0)*

*| mass > 26.4*

*| | age <= 28: tested\_negative (180.0/22.0)*

*| | age > 28*

*| | | plas <= 99: tested\_negative (55.0/10.0)*

*| | | plas > 99*

*| | | | pedi <= 0.56: tested\_negative (84.0/34.0)*

*| | | | pedi > 0.56*

*| | | | | preg <= 6*

*| | | | | | age <= 30: tested\_positive (4.0)*

*| | | | | | age > 30*

*| | | | | | | age <= 34: tested\_negative (7.0/1.0)*

*| | | | | | | age > 34*

*| | | | | | | | mass <= 33.1: tested\_positive (6.0)*

*| | | | | | | | mass > 33.1: tested\_negative (4.0/1.0)*

*| | | | | preg > 6: tested\_positive (13.0)*

*plas > 127*

*| mass <= 29.9*

*| | plas <= 145: tested\_negative (41.0/6.0)*

*| | plas > 145*

*| | | age <= 25: tested\_negative (4.0)*

*| | | age > 25*

*| | | | age <= 61*

*| | | | | mass <= 27.1: tested\_positive (12.0/1.0)*

*| | | | | mass > 27.1*

*| | | | | | pres <= 82*

*| | | | | | | pedi <= 0.396: tested\_positive (8.0/1.0)*

*| | | | | | | pedi > 0.396: tested\_negative (3.0)*

*| | | | | | pres > 82: tested\_negative (4.0)*

*| | | | age > 61: tested\_negative (4.0)*

*| mass > 29.9*

*| | plas <= 157*

*| | | pres <= 61: tested\_positive (15.0/1.0)*

*| | | pres > 61*

*| | | | age <= 30: tested\_negative (40.0/13.0)*

*| | | | age > 30: tested\_positive (60.0/17.0)*

*| | plas > 157: tested\_positive (92.0/12.0)*

*Number of Leaves : 20*

*Size of the tree : 39*

*Time taken to build model: 0.01 seconds*

*=== Stratified cross-validation ===*

*=== Summary ===*

*Correctly Classified Instances 567 73.8281 %*

*Incorrectly Classified Instances 201 26.1719 %*

*Kappa statistic 0.4164*

*Mean absolute error 0.3158*

*Root mean squared error 0.4463*

*Relative absolute error 69.4841 %*

*Root relative squared error 93.6293 %*

*Total Number of Instances 768*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0.814 0.403 0.790 0.814 0.802 0.417 0.751 0.811 tested\_negative*

*0.597 0.186 0.632 0.597 0.614 0.417 0.751 0.572 tested\_positive*

*Weighted Avg. 0.738 0.327 0.735 0.738 0.736 0.417 0.751 0.727*

*=== Confusion Matrix ===*

*a b <-- classified as*

*407 93 | a = tested\_negative*

*108 160 | b = tested\_positive*

**Version with different setting on leaf size limit**

*A screenshot of a computer

Description automatically generated*

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -C 0.25 -M 6*

*Relation: pima\_diabetes*

*Instances: 768*

*Attributes: 9*

*preg*

*plas*

*pres*

*skin*

*insu*

*mass*

*pedi*

*age*

*class*

*Test mode: 10-fold cross-validation*

*=== Classifier model (full training set) ===*

*J48 pruned tree*

*------------------*

*plas <= 127*

*| mass <= 26.4: tested\_negative (132.0/3.0)*

*| mass > 26.4*

*| | age <= 28: tested\_negative (180.0/22.0)*

*| | age > 28*

*| | | plas <= 99: tested\_negative (55.0/10.0)*

*| | | plas > 99*

*| | | | pedi <= 0.56: tested\_negative (84.0/34.0)*

*| | | | pedi > 0.56*

*| | | | | preg <= 6*

*| | | | | | insu <= 120: tested\_negative (11.0/4.0)*

*| | | | | | insu > 120: tested\_positive (10.0/2.0)*

*| | | | | preg > 6: tested\_positive (13.0)*

*plas > 127*

*| mass <= 29.9: tested\_negative (76.0/24.0)*

*| mass > 29.9*

*| | plas <= 157*

*| | | pres <= 61: tested\_positive (15.0/1.0)*

*| | | pres > 61*

*| | | | age <= 30: tested\_negative (40.0/13.0)*

*| | | | age > 30: tested\_positive (60.0/17.0)*

*| | plas > 157: tested\_positive (92.0/12.0)*

*Number of Leaves : 12*

*Size of the tree : 23*

*Time taken to build model: 0.01 seconds*

*=== Stratified cross-validation ===*

*=== Summary ===*

*Correctly Classified Instances 570 74.2188 %*

*Incorrectly Classified Instances 198 25.7813 %*

*Kappa statistic 0.4185*

*Mean absolute error 0.3161*

*Root mean squared error 0.4301*

*Relative absolute error 69.5542 %*

*Root relative squared error 90.2293 %*

*Total Number of Instances 768*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0.830 0.422 0.786 0.830 0.807 0.420 0.779 0.836 tested\_negative*

*0.578 0.170 0.646 0.578 0.610 0.420 0.779 0.640 tested\_positive*

*Weighted Avg. 0.742 0.334 0.737 0.742 0.739 0.420 0.779 0.768*

*=== Confusion Matrix ===*

*a b <-- classified as*

*415 85 | a = tested\_negative*

*113 155 | b = tested\_positive*

**Version with different test mode (percentage split)**

*A screenshot of a computer

Description automatically generated*

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -C 0.25 -M 6*

*Relation: pima\_diabetes*

*Instances: 768*

*Attributes: 9*

*preg*

*plas*

*pres*

*skin*

*insu*

*mass*

*pedi*

*age*

*class*

*Test mode: split 80.0% train, remainder test*

*=== Classifier model (full training set) ===*

*J48 pruned tree*

*------------------*

*plas <= 127*

*| mass <= 26.4: tested\_negative (132.0/3.0)*

*| mass > 26.4*

*| | age <= 28: tested\_negative (180.0/22.0)*

*| | age > 28*

*| | | plas <= 99: tested\_negative (55.0/10.0)*

*| | | plas > 99*

*| | | | pedi <= 0.56: tested\_negative (84.0/34.0)*

*| | | | pedi > 0.56*

*| | | | | preg <= 6*

*| | | | | | insu <= 120: tested\_negative (11.0/4.0)*

*| | | | | | insu > 120: tested\_positive (10.0/2.0)*

*| | | | | preg > 6: tested\_positive (13.0)*

*plas > 127*

*| mass <= 29.9: tested\_negative (76.0/24.0)*

*| mass > 29.9*

*| | plas <= 157*

*| | | pres <= 61: tested\_positive (15.0/1.0)*

*| | | pres > 61*

*| | | | age <= 30: tested\_negative (40.0/13.0)*

*| | | | age > 30: tested\_positive (60.0/17.0)*

*| | plas > 157: tested\_positive (92.0/12.0)*

*Number of Leaves : 12*

*Size of the tree : 23*

*Time taken to build model: 0.01 seconds*

*=== Evaluation on test split ===*

*Time taken to test model on test split: 0 seconds*

*=== Summary ===*

*Correctly Classified Instances 115 74.6753 %*

*Incorrectly Classified Instances 39 25.3247 %*

*Kappa statistic 0.4066*

*Mean absolute error 0.3078*

*Root mean squared error 0.4366*

*Relative absolute error 68.7067 %*

*Root relative squared error 93.4163 %*

*Total Number of Instances 154*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0.829 0.429 0.806 0.829 0.817 0.407 0.782 0.861 tested\_negative*

*0.571 0.171 0.609 0.571 0.589 0.407 0.782 0.568 tested\_positive*

*Weighted Avg. 0.747 0.347 0.743 0.747 0.745 0.407 0.782 0.768*

*=== Confusion Matrix ===*

*a b <-- classified as*

*87 18 | a = tested\_negative*

*21 28 | b = tested\_positive*

**Differences:**

* **Uses ‘-M 2’ meaning that a leaf node is created if it has at least 2 instances. The second and third run uses ‘-M 6’ increasing minimum instances required to form a leaf node. This will then reduce overfitting as the tree does not grow too deep. The size and the number of leaves also decrease. Specifically, the size of the tree decreased from 39 to 23, and the number of leaves reduced from 20 to 12.**
* **The accuracy improved a bit with ‘-M’ (2 to 6), which increased from 73.8281% to 74.2188% in cross-validation, and demonstrated a similar trend in the percentage split evaluation with an accuracy of 74.6753%**
* **The third run introduces a different test mode, switching from 10-fold cross-validation to splitting the dataset into 80% for training and 20% for testing. There is only a slightly higher accuracy in the test split**

**Q6. Which model is finally chosen and why? (10’)**

The final model that should be chosen is the one with the different settings on the leaf size limit (‘-M 6’) which used the 10-fold cross validation. This is why:

1. Accuracy increased from 73.8281% to 74.2188%.
2. It has fewer leaves (12 vs 20) and a smaller tree size (23 vs 39), which suggests it is simpler and potentially more robust against overfitting.
3. There is also a consistent performance among other evaluation metrics (precision, recall, and F-measure)

**Q7. Create a new record with arbitrary values and use the chosen model to predict the class for this new record. What is the new record and what is the result? (10’)**

**Record**

*3,122,70,30,88,35.9,0.452,28,tested\_negative*

*2,110,64,20,0,28.1,0.259,22,tested\_positive*

*5,136,68,0,0,32.3,0.855,35,tested\_negative*

*0,107,76,0,0,45.3,0.678,23,tested\_negative*

*1,90,60,18,0,23.5,0.191,25,tested\_positive*

**Result**

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -C 0.25 -M 6*

*Relation: pima\_diabetes*

*Instances: 768*

*Attributes: 9*

*preg*

*plas*

*pres*

*skin*

*insu*

*mass*

*pedi*

*age*

*class*

*Test mode: user supplied test set: size unknown (reading incrementally)*

*=== Classifier model (full training set) ===*

*J48 pruned tree*

*------------------*

*plas <= 127*

*| mass <= 26.4: tested\_negative (132.0/3.0)*

*| mass > 26.4*

*| | age <= 28: tested\_negative (180.0/22.0)*

*| | age > 28*

*| | | plas <= 99: tested\_negative (55.0/10.0)*

*| | | plas > 99*

*| | | | pedi <= 0.56: tested\_negative (84.0/34.0)*

*| | | | pedi > 0.56*

*| | | | | preg <= 6*

*| | | | | | insu <= 120: tested\_negative (11.0/4.0)*

*| | | | | | insu > 120: tested\_positive (10.0/2.0)*

*| | | | | preg > 6: tested\_positive (13.0)*

*plas > 127*

*| mass <= 29.9: tested\_negative (76.0/24.0)*

*| mass > 29.9*

*| | plas <= 157*

*| | | pres <= 61: tested\_positive (15.0/1.0)*

*| | | pres > 61*

*| | | | age <= 30: tested\_negative (40.0/13.0)*

*| | | | age > 30: tested\_positive (60.0/17.0)*

*| | plas > 157: tested\_positive (92.0/12.0)*

*Number of Leaves : 12*

*Size of the tree : 23*

*Time taken to build model: 0.01 seconds*

*=== Evaluation on test set ===*

*Time taken to test model on supplied test set: 0 seconds*

*=== Summary ===*

*Correctly Classified Instances 2 40 %*

*Incorrectly Classified Instances 3 60 %*

*Kappa statistic -0.3636*

*Mean absolute error 0.5632*

*Root mean squared error 0.6737*

*Relative absolute error 119.8698 %*

*Root relative squared error 136.7798 %*

*Total Number of Instances 5*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0.667 1.000 0.500 0.667 0.571 -0.408 0.167 0.533 tested\_negative*

*0.000 0.333 0.000 0.000 0.000 -0.408 0.167 0.325 tested\_positive*

*Weighted Avg. 0.400 0.733 0.300 0.400 0.343 -0.408 0.167 0.450*

*=== Confusion Matrix ===*

*a b <-- classified as*

*2 1 | a = tested\_negative*

*2 0 | b = tested\_positive*