Chapter 1

Preprocessing

1.1 Technique And Design

The universal database contains raw data of academic performance of the students.There are plenty of problems in this Database which has been discussed in *Chapter3*.A rigorous preprocessing technique is highly needed for turn the database into a suitable one for the classification implementation.The classification algorithm demands a clean and quality dataset without which it can't proceed efficiently.In this procedure,there were several steps taken for the data preprocessing.

1.1.1 Data Cleaning

- One of the main problems with the universal database was,there are several attributes where no value was not inserted at all.In result,there are plenty of blank in the data sheet. As the first step of cleaning,these missing values were filled up using intuition and global constant.For example, data sheet 1 had some blank attributes like the following:

Serial	Department	...	CourseName	PartAMark	PartBMark	...
65	CSE	...	CSE100	-	-	...
65	CSE	...	CSE210	-	-	...
...	...	...	...	...	...	...

Table 1.1: Blank attributes in the Universal Database

The *PartAMark & PartBMark* fields are blank because the tuples are

for sessional courses and only theory courses have Part A and Part B factor.
In this case,these two fields have been filled up with 0.After the process,table looked like this,

Serial	Department	...	CourseName	PartAMark	PartBMark	...
65	CSE	...	CSE100	0	0	...
65	CSE	...	CSE210	0	0	...
...	...	...	...	...	...	...

Table 1.2: Universal database after filling up blank attributes

- There are several attributes in the dataset which are not really needed for the classification implementation.As we are going to implement the excellency of a student in academic career,attributes like *District,Thana, PlaceOfBirth,AdmissionDate,StartingDate* are not going to affect the classification algorithm anyway.So,all the database has been got rid of all the redundant attributes. Before this procedure,the database was like the following:

Serial	Department	District	Thana	CourseName	LetterGrade	...
4478	NAME	Dhaka	Dhamrai	MATH141	A	...
14251	MME	Satkhira	koloroa	PHY143	B+	...
...	...	...	...	...	...	...

Table 1.3: Redundant attributes in Database

After data cleaning process,the redundant attributes have been taken care of.

Serial	Department	CourseName	LetterGrade	...
4478	NAME	MATH141	A	...
14251	MME	PHY143	B+	...
...	...	...	...	...

Table 1.4: Database after cleaning of redundant attributes

- The universal database is filled with plenty of inconsistent data.These data types are not compatible with the procedure of ID3 classification.Some steps were taken to change these inconsistent data type to a consistent one.
  The *Gender* and *HallStatus* fields are filled with binary value.*Gender* is represented as *Male* or *Female* and *HallStatus* is represented with *Resident* or *Attached*.

Serial	Department	Gender	HallStatus	...
4478	NAME	Male	Attached	...
14251	MME	Male	Resident	...
...	...	...	...	...

Table 1.5: Database wth binary values

The values are converted to numeric from binary so that the data set becomes compatible with the ID3 procedure.

Serial	Department	Gender	HallStatus	...
4478	NAME	0	1	...
14251	MME	0	0	...
...	...	...	...	...

Table 1.6: Database after conversion of binary attributes to numeric value

- A crucial fault in the universal database is,it doesn't have any *CourseCreditHour* attribute and some data sheets don't have the correct value of *CGPA* as well.The *CGPA* field is rounded in some data sheets.Which does not represent the accurate value.But for proper calculation,accurate *CGPA* and *GPA* is must.

Serial	Department	CourseName	GPA	CGPA	...
4478	NAME	MATH141	3	3	...
14251	MME	PHY143	3	3	...
...	...	...	...	...	...

Table 1.7: Database before Data cleaning

In this case,with the help of other attributes of the database, *CourseCreditHour*, *GPA* and *CGPA* is calculated.
At first, *Numberpercentage* is calculated from *LetterGrade* attribute.

LetterGrade	NumberPercentage
A+	80%-100%
A	75%-79%
A-	70%-74%
B+	65%-69%
B	60%-64%
B-	55%-59%
C+	50%-54%
C	45%-49%
D	40%-44%

Table 1.8: Letter Grade and it's respective percentage of number

Then,the *CourseCreditHour* is calculated from the following equation.

$$CourseCreditHour = \frac{TotalNumber}{NumberPercentage} \tag{1.1}$$

After the modification in *CourseCreditHour*,the database looks like the following:

Serial	Department	CourseName	GPA	CourseCreditHour	...
4478	NAME	MATH141	3.5	4	...
14251	MME	PHY143	3.25	3	...
...	...	...	...	...	...

Table 1.9: Database after addition of CourseCreditHour attribute

In the ID3 implementation,the final *CGPA* of student is used only.It is calculated using the *CourseCreditHour* and *GPA* attribute of all courses the student is enrolled to.

$$CGPA = \frac{\sum(CourseCreditHour * CourseGPA)}{\sum CourseCreditHour} \tag{1.2}$$

After the modification in *CGPA*,the database looks like the following:

Serial	Department	CGPA	...
4478	NAME	3.46	...
14251	MME	3.1	...
...	...	...	...

Table 1.10: Database after calculation of final CGPA

In this phase,one student has only one tuple in the database.

1.1.2 Data Normalization

1.2 Results

After implementing some crucial data preprocessing procedures,the modified database is at last suitable for the next step,implementation of ID3 algorithm using decision tree.The final test data has only the relevant attributes in compatible format.The attributes are *Serial,Department,HallStatus,Gender, ClassTestMark,AttendanceMark,CGPA and TotalCourseCompleted.*
After the Data Cleaning and Data Normalization the final test data:

Serial	Department	Hall Status	Gender	Attendance	Class Test	CGPA	Course Completed
4478	6	0	1	0.82	0.72	3.8	1
4480	8	0	0	0.96	0.66	2.99	1
14251	2	1	1	0.64	0.67	3.11	1
...	...	...	...	...	...	...	...

Table 1.11: Final test data after preprocessing