EXPERIMENT 5

Title :	Perform exploratory Data Analysis and preprocessing for the data set.
Problem :	Perform following Data pre-processing activities using WEKA: 1. Normalization 2. Discretization 3. Duplicate Removal 4. Standardization 5. Handling of Missing Data.
Theory:	Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. **Missing Data:** This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are: Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple. Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value. **Data Transformation:** This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways: **Normalization:** It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0) **Attribute Selection:** In this strategy, new attributes are constructed from the given set of attributes to help the mining process. **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels. **Concept Hierarchy Generation:** Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Performance	1. Load the ***diabetes.arff*** file into WEKA 2. Study the dataset and answer following Number of instances = ? Number of Attributes = ? Types of attributes other than class = ? Type of class Attribute = ? 3. Study the Class labels. Do you think if there is any class imbalance. What are the class labels? Which class is dominant? Paste the screen shot – depicting the number of classes 4. Study the spread of at-least three numeric attributes and provide statistical measures. 5. Discretize Age attribute – choose appropriate number of bins for discretization 6. Normalize plas attribute using min-max normalization – use normalization range as (0 ,1) 7. Remove Duplicate instances if any 8. Impute Missing values using mean value imputation.
Deliverables:	A PDF document with screen shots and appropriate description depicting every activity mentioned in the performance section.
Conclusion:	Students will write the conclusion in their own words summarizing the understandings from the practical performed.