**Experiment 2.Creating new ARFF file**

**Aim: Creating a new ARFF file**

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software in WEKA, each data entry is an instance of the java class weka.core. Instance, and each instance consists of a For loading datasets in WEKA, WEKA can load ARFF files. Attribute Relation File Format has two sections:

1. The Header section defines relation (dataset) name, attribute name, and type.

2. The Data section lists the data instances.

The figure above is from the textbook that shows an ARFF file for the weather data. Lines beginning with a % sign are comments. And there are three basic keywords:

The external representation of an Instances class Consists of:

 **A header:** Describes the attribute types

 **Data section:** Comma separated list of data

**Experiment 3: Pre-Processes Techniques on Data Set**

**Aim: 3a) Pre-process a given dataset based on Attribute selection**

To search through all possible combinations of attributes in the data and find which subset of attributes works best for prediction, make sure that you set up attribute evaluator to „Cfs Subset Val‟ and a search method to „Best First‟. The evaluator will determine what method to use to assign a worth to each subset of attributes. The search method will determine what style of search to perform. The options that you can set for selection in the „Attribute Selection Mode‟ fig no: 3.2

1. **Use full training set.** The worth of the attribute subset is determined using the full set of training data.

2. **Cross-validation.** The worth of the attribute subset is determined by a process of cross- validation. The „Fold‟ and „Seed‟ fields set the number of folds to use and the random seed used when shuffling the data.

Specify which attribute to treat as the class in the drop-down box below the test options. Once all the test options are set, you can start the attribute selection process by clicking on

When it is finished, the results of selection are shown on the right part of the window and entry is added to the „Result list‟.

WEKA‟s visualization allows you to visualize a 2-D plot of the current working relation. Visualization is very useful in practice; it helps to determine difficulty of the learning problem. WEKA can visualize single attributes (1-d) and pairs of attributes (2-d), rotate 3-d visualizations (Xgobi-style). WEKA has “Jitter” option to deal with nominal attributes and to detect “hidden” data points.

**Aim: B. Pre-process a given dataset based on Handling Missing Values**

**Process**: Replacing Missing Attribute Values by the Attribute Mean. This method is used for data sets with numerical attributes. An example of such a data set is presented in fig no: 3.4

In this method, every missing attribute value for a numerical attribute is replaced by the arithmetic mean of known attribute values. In Fig, the mean of known attribute values for Temperature is 99.2, hence all missing attribute values for Temperature should be replaced by The table with missing attribute values replaced by the mean is presented in fig. For symbolic attributes Headache and Nausea, missing attribute values were replaced using the most common value of the Replace Missing Values.

**Experiment 9: calculating information gain measures.**

Information gain (IG) measures how much “information” a feature gives us about the class. – Features that perfectly partition should give maximal information. – Unrelated features should give no information. It measures the reduction in entropy. CfsSubsetEval aims to identify a subset of attributes that are highly correlated with the target while not being strongly correlated with one another. It searches through the space of possible attribute subsets for the “best” one using the BestFirst search method by default, although other methods can be chosen. To use the wrapper method rather than a filter method, such as CfsSubsetEval, first select WrapperSubsetEval and then configure it by choosing a learning algorithm to apply and setting the number of cross-validation folds to use when evaluating it on each attribute subset.

Steps:

 Open WEKA Tool.

 Click on WEKA Explorer.

 Click on Preprocessing tab button.

 Click on open file button.

 Select and Click on data option button.

 Choose a data set and open file.

 Click on select attribute tab and Choose attribute evaluator, search method algorithm