	Hope Foundation's **Finolex Academy of Management and Technology, Ratnagiri** **Department of Information Technology**		
Subject name	**Business Intelligence Lab**		**Subject Code: ITL602**
Class	**TE IT**	**Semester – VI (CBCGS)**	**Academic year: 2018-19 (FH 2019)**
Name of Student			**QUIZ Score :**
Roll No		**Assignment/Experiment No:**	**02**
Title:	**To use WEKA to implement the Classifiers Algorithms**		

1. Course objectives applicable: LO3

2. Course outcomes applicable: LO3

3. Learning Objectives:

1. To learn classification techniques
2. To understand supervised learning
3. To learn to run classification algorithms in WEKA

4. Practical applications of the assignment/experiment: Prediction using classification

5. Prerequisites:

1. Types of classification algorithms
2. Supervised learning concepts

6. Hardware Requirements:

1. PC with minimum 2 GB RAM

7. Software Requirements:

1. Windows 8.1 or higher
2. WEKA 3.8 or higher

8. Viva Questions (if any): (Online Quiz will be taken separately batch-wise)

1. What is meant by labeled data?
2. What is supervised learning?
3. What are the steps in data mining using classification?
4. What is precision and recall?

9. Experiment/Assignment Evaluation:

Sr. No.	Parameters	Marks obtained	Out of
1	Technical Understanding (Assessment may be done based on Q & A **or** any other relevant method.)		6
2	Neatness/presentation		2
3	Punctuality		2

Date of performance (DOP)		**Total marks obtained**		**10**
Date of checking (DOC)		**Signature of teacher**		

10. Theory: <<handwritten work>>

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from *labeled training data* consisting of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Decision trees used in data mining are of two main types:

- **Classification tree** analysis is when the predicted outcome is the class to which the data belongs.
- **Regression tree** analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).
- Some techniques, often called ***ensemble* methods**, construct more than one decision tree

- **Boosted trees** incrementally building an ensemble by training each new instance to emphasize the training instances previously mis-modeled. A typical example is AdaBoost. These can be used for regression-type and classification-type problems.
- **Bootstrap aggregated** (or bagged) decision trees, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction
- **Random Forest** classifier is a specific type of bootstrap aggregating.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

11. Performance Steps: <<Handwritten>>

1. Start WEKA
2. Open the data set (ARFF file)
3. Select the classifier (Decision Tree, Random Forest, Naïve Bayes) and run the trial
4. Record the reading and find precision and Recall
5. Identify a subset of attributes having negative correlation with results i.e. by removing the attributes accuracy shall increase.

12. Results:

<< Add the hard-copy of output screen shots >>

13. Learning Outcomes Achieved

1. Understanding of meaning of labeled data and supervised learning.
2. Understanding of classification steps: training, testing and prediction.
3. Understanding of attribute selection process.

14. Conclusion:

1. **Applications of the studied technique in industry:** To build classification models
2. **Engineering Relevance:** To build Decision support systems using classification
3. **Skills Developed:** Understanding of classification techniques

15. References:

[1] Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition.

[2] P. N. Tan, M. Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education.

[3] Michael Berry and Gordon Linoff, "Data Mining Techniques", 2nd Edition Wiley Publications

[4] https://en.wikipedia.org/wiki/Decision_tree_learning

[5] https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/