# KAKATIYA UNIVERSITY WARANGAL

Under Graduate Courses (Under CBCS AY: 2021-2022 on words)

## B.Sc. DATA SCIENCE

**II Year: Semester-IV**

# Paper – IV: Machine Learning

[4 HPW:: 4 Credits :: 100 Marks (External:80, Internal:20)]

**Objectives:** The main objective of this course is to teach the principles and foundations of machine learning algorithms

**Outcomes:**

At the end of the course the student will be able to understand

- Basics of Machine Learning and its limitations
- Machine Learning Algorithms: supervised, unsupervised, bio-inspired
- Probabilistic Modeling and Association Rule Mining

## Unit-I

**Introduction:** What does it mean to learn, Some canonical Learning Problems, The Decision Tree Model of Learning, Formalizing the Learning Problem ID3 Algorithm [Reference1, 2]
**Limits of Learning:** Data Generating Distributions, Inductive Bias, Not Everything islearnable, Under fitting and Overfitting, Separation of training and test Data, Models, parameters and Hyperparameters, Real World Applications of Machine Learning **Geometry and Nearest Neighbours:** From Data to Feature Vectors, k-Nearest Neighbours, Decision Boundaries, k-means Clustering, High Dimensions [Reference 1]

## Unit-II

**The Perceptron:** Bio-inspired Learning, The Perceptron Algorithm, Geometric Interpretation, Interpreting Perceptron Weights, Perceptron Convergence and Linear Separability, Improved Generalization, Limitations of the Perceptron
**Practical Issues:** Importance of Good Features, Irrelevant and Redundant Features, Feature Pruning and Normalization, Combinatorial Feature Explosion, Evaluating Model Performance, Cross Validation, Hypothesis Testing and Statistical Significance, Debugging Learning Algorithms, Bias Variance tradeoff
**Linear Models:** The Optimization Framework for Linear Models, Convex Surrogate Loss Functions, Weight Regularization, Optimization and Gradient Descent, Support Vector Machines [Reference 1]

## Unit-III

**Probabilistic Modelling:** Classification by Density Estimation, Statistical Estimation, Naïve Bayes Models, Prediction [Reference 1]
**Neural Networks:** Bio-inspired Multi-Layer Networks, The Back-propagation Algorithm, Initialization and Convergence of Neural Networks, Beyond two layers, Breadth vs Depth, Basis Functions [Reference 1]

## Unit IV

**Unsupervised Learning:** Clustering Introduction, Similarity and Distance Measures,Agglomerative Algorithms, Divisive Clustering, Minimum Spanning Tree [Reference 2]
**Association Rules:** Introduction, large Itemsets, Apriori Algorithm [Reference 2]

**References:**


1. A Course in Machine Learning (CIML). Hal Daume III, 2017 (freely available online) http://ciml.info/
2. Data Mining: Introductory and Advanced Topics. Margaret H Dunham, Pearson Education, 2003


**Suggested Reading:**


3. Hands on Machine Learning with SciKit-Learn, Keras and Tensor Flow. AurélienGéron. O'Reily, 2019
4. Machine Learning with Python Cookbook. Chris Albo, O'Reily, 2018
5. Introduction to Machine Learning with Python: A guide. Andreas C Miller, Sarah Guido. O'Reily, 2017

# *Practical- 4:* Machine Learning (Lab)

[3 HPW:: 1 Credit :: 25 Marks]

**Objective:**
The main objective of this laboratory is to put into practice the various machine learning algorithms for data analysis using Python and Weka.

**ML Toolkits**
Students are expected to learn
1. Scikit-learn(https://scikit-learn.org/) an open source machine learning Python library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.
2. Weka (http://www.cs.waikato.ac.nz/ml/weka/)is another widely used ML toolkit.

**Datasets**
1. The sklearn datasets package embeds small toy datasets. It includes utilities to load these datasets. It also includes methods to load and fetch popular reference datasets and features some artificial data generators. Students are expected to study and make use of these datasets
2. Weka also has provides various data sets.

**References:**
1. Scikit-learn user guide.https://scikit-learn.org/stable//_downloads/scikit-learn-docs.pdf
2. Ian Witten, Eibe Frank, and Mark Hall, Chris Pal. DATA MINING: Practical Machine Learning Tools and Techniques, 4th Edition. Morgan Kaufmann.

**Exercises**
1. Write a Python program using Scikit-learn to split the iris dataset into 70% train data and 30% test data. Out of total 150 records, the training set will contain 120 records and the test set contains 30 of those records. Print both datasets
2. Write Python program to use sklearn's Decision Tree Classifier to build a decision tree for the sklearn's datasets. Implement functions to find the importance of a split (entropy, information gain, gini measure)
3. Write a Python program to implement your own version of the K-means algorithm. Then apply it to different datasets and evaluate the performance.
4. Design a perceptron classifier to classify handwritten numerical digits (0-9). Implement using scikit or Weka.
5. Write a Python program to classify text as spam or not spam using the Naïve Bayes Classifier
6. Use WEKA and experiment with the following classifiers: Association Rule Mining (Apriori), Agglomerative and Divisive Clustering.