NIRMA UNIVERSITY

Institute:	Institute of Technology, School of Technology
Name of Programme:	MTech (Data Science)
Course Code:	6CS363ME22
Course Title:	Advanced Statistical Learning
Course Type:	Department Elective-I
Year of Introduction:	2022-23

L	T	Practical Component				C
		LPW	PW	W	S	
3	0	2	-	-	-	4

Course Learning Outcomes (CLO):

At the end of the course, the students will be able to:

1. comprehend the fundamentals of various statistical learning methods (BL2)
2. interpret and critically evaluate the outcomes of statistical analysis (BL3,4)
3. compare and contrast statistical models in the context of a particular scientific question (BL4) (BL5)
4. implement statistical learning methods.

Unit	Contents	Teaching Hours (Total 45)
Unit-I	**Fundamentals:** Basics of Statistical Learning Theory, Accessing Model Accuracy	02
Unit-II	**Overview of Supervised Learning:** Introduction, Simple Approaches to Prediction, Local Methods in High Dimensions, Supervised Learning and Function Approximation, Model Selection and the Bias–Variance Tradeoff	07
Unit-III	**Linear Methods for Regression and Classification:** Linear Regression Models and Least Squares, Subset Selection, Shrinkage Methods, Linear Discriminant Analysis, Logistic Regression, Separating Hyperplanes, Piecewise Polynomials and Splines, Regularization, General Linear Modelling	09
Unit-IV	**Kernel Smoothing Methods:** One-Dimensional Kernel Smoothers, Selecting the Width of the Kernel, Local Likelihood and other Methods, Kernel Density Estimation and Classification, Radial Basis Functions and Kernels, Mixture Models for Density Estimation and Classification, Computational Considerations	07
Unit-V	**Model Assessment, Selection, Inference and Averaging:** Bias, Variance and Model Complexity, Cross-Validation, Bootstrap and Bagging, EM Algorithm	05

Unit-VI	**Additive Models, Trees, SVM and Nearest Neighbours:** Generalized Additive Models, Tree-based Methods, Boosting Methods, Random Forests, SVM and Kernels, k-Nearest-Neighbour Classifiers, Adaptive Nearest-Neighbour Methods	08
Unit-VII	**Unsupervised Learning:** K-means, Self-Organizing Maps, Non-negative Matrix Factorization, Independent Component Analysis.	07

Self-Study:

The self-study contents will be declared at the commencement of the semester. Around 10% of the questions will be asked from self-study content.

Suggested Readings/ References:

1. James, G., Witten, D., Hastie, T., & Tibshirani, R., An introduction to statistical learning, Springer
2. Berk, R. A., Statistical learning from a regression perspective, Springer
3. Urdan, T. C., Statistics in plain English, Routledge
4. Haslwanter, T., An Introduction to Statistics with Python, Springer International Publishing.

Suggested List of Experiments:

Sr.	Name of Experiments/Exercises	Hours
1	Implement polynomial regression and observe the tradeoff between bias and variance in the regression	02
2	Implement Ridge and Lasso regression to reduce overfitting. Demonstrate underfitting and overfitting with and without Ridge and Lasso Regression	02
3	Implement logistic regression and evaluate the performance using ROC curve. Calculate AUC score and demonstrate precision-recall tradeoff	04
4	Implement decision tree and Random Forest for classification and accuracy comparison. Also perform feature importance analysis	04
5	Perform unsupervised learning for clustering and visualization. Use PCA for high dimensional data and interpret cluster centroids	04
6	Train SVM model with different kernels and analyze hyperparameter (C, gamma) effects using GridSearchCV	02
7	Implement and predict future values in a time series dataset using ARIMA model. Decompose a time series into trend, seasonality and noise. Evaluate forecasting accuracy using RMSE	04
8	Implement a Gaussian Mixture Model to a detect anomalies. Use normal and anomalous data sets to deal with anomalies. Compare performance with k-Means clustering	02
9	Apply Bayesian learning to update beliefs based on new data. Compare Bayesian regression with classical regression	04
10	Train and evaluate a feedforward neural network. Tune hyperparameters (learning rate, number of layers). Compare performance with logistic regression.	02