# JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR
(Established by Govt. of A.P., ACT No.30 of 2008)
ANANTHAPURAMU – 515 002 (A.P) INDIA

## MASTER OF COMPUTER APPLICATIONS

| Course Code | DATA SCIENCE WITH PYTHON | L | T | P | C |
|---|---|---|---|---|---|
| 21P00202 | | 4 | 0 | 0 | 4 |
| | Semester | II | | | |

**Course Objectives:**

Ideally for a student to understand Data Science, he/she should have exposure to the following. This will give a basic feel about Data Science. In the following, the topics highlighted in light blue is minimum needed and those highlighted in yellow will help to get a feel about the subject.
Overall it covers the following:

- Basics of probability
- Basics of statistics
- Pattern Recognition
- Machine Learning
- Introduction on Deep Neural Networks.

**Course Outcomes (CO):** Student will be able to

- Obtain, clean/process, and transform data
- Analyze and interpret data using an ethically responsible approach
- Use appropriate models of analysis, assess the quality of input, derive insight from results, and investigate potential issues
- Apply computing theory, languages, and algorithms, as well as mathematical and statistical models, and the principles of optimization to appropriately formulate and use data analyses
- Formulate and use appropriate models of data analysis to solve hidden solutions to business-related challenges
- Perform well in a group

**UNIT – I** — Lecture Hrs:

Descriptive Statistics: Measures of central tendency—mean, median, mode, harmonic mean and geometric mean; Measures of dispersion – mean deviation from mean, standard deviation and variance. Central moments. Linear and rank correlation. Covariance and correlation; Statistics and sampling distributions; Hypothesis testing of means, proportions, variances and correlations Definition of random variable and probability, (problems depending on counting – taught in MFCS), discrete probability distributions: Bernoulli, Binomial, Poisson; Continuous probability distributions: Gaussian, Exponential, Chisquare. Definition of Bayesian probability.

**UNIT - II** — Lecture Hrs:

Exploratory Data Analysis (EDA), Data Science life cycle, Descriptive Statistics, Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA.
Data Visualization: Scatter plot, bar chart, histogram, boxplot, heat maps etc.

**UNIT - III** — Lecture Hrs:

Patterns, features, patter representation, curse of dimensionality, dimensionality reduction. Supervised and Unsupervised learning. Classification—linear and non-linear. Bayesian, Perceptron, Nearest neighbour classifier, Support vector machine, use of kernels, Logistic regression, Naïve-bayes, decision trees and random forests; boosting and bagging. Clustering---partitional and hierarchical; k-means clustering.
Regression. Least squares. Evaluation metrics: RMSE, MAE and Coefficient of Determination (R- square) Cost functions, training and testing a classifier. Cross-validation. Class-imbalance – ways of handling, Exploratory data analysis (EDA), evaluation metrics— Precision, Recall, RoC, AUC; Confusion matrix, Classification accuracy

**UNIT - IV** — Lecture Hrs:

Multilayer perceptron. Back propagation. Loss functions. Epochs and Batch sizes. Hyper parameter tuning. Applications to classification, regression and unsupervised learning. Overview(introduction to the terms) of RNN, CNN and LSTM.

**UNIT - V**

Applications to text, images, videos: recommender systems, image classification, Social network graphs

**Textbooks:**


- Cathy O'Neil, Rachel Schutt, Doing Data Science, Straight Talk from the Frontline. O'Reilly, 2013.
- Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2007.
- Ian Goodfellow, YoshuaBengio, Aaron Courville, Deep Learning, MIT Press, 2016
- Trevor Hastie, Robert Tibshirani,Jerome Friedman,The Elements of Statistical Learning, Springer 2009.
- Erwin kreyszig, Advanced Engineering Mathematics, 10th Edition, John Wiley & Sons,2011.