GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

Prerequisites: Basics of computer science including algorithms, data structure, Probability theory and Bayesian Concept Learning *

1. Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE (E)	PA (M)	ESE (V)	PA (I)	
3	-	2	4	70	30	30	20	150

1. Course Outcomes:

Course Outcome Component	Course Outcome (Learner will be able to)
CO-1	Understanding Various Machine Learning Algorithms & Their Area of Applications
CO-2	Explain the Process of Formulating & Solving Real World Problem using Machine Technology
CO-3	Design and Implement various Decision Making Problems using Machine Learning

2. Course Duration: The course duration is of **40 sessions of 60 minutes each.**

3. Course Contents:

Unit No:	Contents	No. of Sessions	70 Marks (External Exam)
I	• **Introduction to Machine Learning:** What is human learning? What is Machine Learning? Human learning versus machine learning, Types of machine learning, Applications of machine learning, Tools for machine learning.	4	7
II	• **Preparing to Model , Feature Engineering:** Machine Learning activities, Basic Types of data in Machine Learning, Structures of data, Data Quality and Remediation, Data Pre-Processing, Introduction to Feature Engineering, Feature Transformation, Feature Subset Selection	10	18
III	• **Modelling and Evaluation:** Selecting a Model, Training a Model, Model Representation and Interpretability, Evaluating Performance of a Model, Improving Performance of a Model	8	14
IV	• **Supervised Learning: Classification and Regression:** Example of Supervised Learning, Classification Model, Classification Learning Steps, Common Classification	10	17

GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

With effective from academic year 2020-21

	Algorithms Example of Regression, Common regression Algorithms			
V	• **Unsupervised Learning:** Unsupervised Learning versus Supervised Learning, Applications of Unsupervised Learning, Clustering, Finding Pattern using Association Rules	8	14	

* Additional Sessions may be taken to revise Prerequisites Contents.

4. Pedagogy:
- ICT enabled Classroom teaching
- Case study
- Practical to be implemented using python library
  (Sample practical list is attached as Appendix-A)
- Interactive class room discussions

5. Evaluation:
Students shall be evaluated on the following components:

	Internal Evaluation	(Total - 20 Marks)
A	• Continuous Evaluation Component	10 marks
	• Class Presence & Participation	10 marks
B	**Mid-Semester examination**	**(30 Marks)**
C	**End –Semester Examination(Theory)**	**(70 Marks)**
D	**End –Semester Examination(Practical/Viva)**	**(30 Marks)**

6. Reference Books:

No.	Author	Name of the Book	Publisher
1	Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das **	Machine Learning	Pearson Education
2	Tom M Mitchell	Machine Learning	McGraw Hill
3	Anuradha Srinivarasaraghavan, Vincy Joseph	Machine Learning	Wiley India
4	Peter Harrington	Machine Learning in Action	DreamTech
5	Henrik Brink, Joseph Richards, Mark Fetherolf	Real-World Machine Learning	DreamTech
6	Christopher Bishop	Pattern Recognition and Machine Learning	Springer
7	Jiawei Han and Michelline Kamber	Data Mining: Concepts and Techniques	Morgan Kaufmann

** The first Reference book is to be considered as Main Book as per Syllabus Contents

Appendix-A

Sample Practical List

1. Write a python code to implement **decision tree** for below given dataset. Identify the root node and all subpart or children of node and draw the tree.

Item no	Age	Income	Student	Credit Rating	Buys-Computer

GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

With effective from academic year 2020-21

1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Middle	Low	Yes	Excellent	No
7	Senior	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle	Medium	No	Excellent	Yes
13	Middle	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

2. Write a python code to implement **K-nearest neighbourhood** program for the given dataset.

Item no	Temp	Humidity	Wind Speed	Play
1	85	85	12	No
2	80	90	9	No
3	83	86	4	Yes
4	70	96	3	Yes
5	68	80	5	Yes
6	65	70	20	No
7	64	65	2	Yes
8	72	95	12	No
9	69	70	5	Yes
10	75	80	2	Yes
11	75	70	3	Yes
12	72	90	4	Yes
13	81	75	5	Yes
14	71	91	15	No

3. Write a python code to implement Apriori algorithm, apply join and prune method and find frequent item set

Sr No.	Item no	Name
1	T1	Bread, butter, milk, soda
2	T2	Coke, egg, milk
3	T3	Bread, butter, egg
4	T4	Break, coke, jam
5	T5	Bread, butter
6	T6	Potato chips, soda

GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

With effective from academic year 2020-21

7	T7	Coke, fruit, juice
8	T8	Bread, coke, milk
9	T9	Coke, soda, jam, milk
10	T10	Bread, butter, egg, milk, soda
11	T11	Bread, milk
12	T12	Bread, jam

4. Write a python code to apply **Naive Bayesian and Logistic Regression** algorithm to classify that whether a person can buy computer or not based on given test data:

Item no	Age	Income	Student	Credit Rating	Buys-Computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Middle	Low	Yes	Excellent	No
7	Senior	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle	Medium	No	Excellent	Yes
13	Middle	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

Test Data
Age: Youth Income: LOW Student: No Credit Rating: Fair Buy Computer -??

5. Python code for **Preparing to Model** (Appendix B.2, Page No. 365 – 373) and **Feature Engineering** (Appendix B.4, Page No. 378 – 385).

6. Apply k-means clustering approach with k = 2 to the following dataset.

X	Y	Z
-0.154	0.376	0.099
-0.103	0.476	-0.027
0.228	0.036	-0.251
0.330	0.013	-0.251
-0.114	0.482	0.014
0.295	0.084	-0.297
0.262	0.042	-0.304

GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

With effective
from academic
year 2020-21

-0.051	0.416	-0.306

7. Implement a python program that takes interest rate (x), finds the equation that best fits the data and is able to forecast out median home price for given interest rate using the data given below. (Use linear regression)

Interest rate (%)(x)	Median home price (y)
10.3	$183,800
10.3	$183,200
10.1	$174,900
9.3	$173,500
8.4	$172,900
7.3	$173,200
8.4	$173,200
7.9	$169,700
7.6	$174,500
7.6	$177,900
6.9	$188,100
7.4	$203,200
8.1	$230,200
7	$258,200
6.5	$309,800
5.8	$329,800

8. Apply following supervised machine learning algorithms for **All** classification problems as shown under:

Algorithm	Data set for Classification Problem
Decision Tree	Iris Dataset (from sklearn.dataset import load_iris)
Logistic Regression	Wine Dataset (from sklearn.dataset import load_wine)
K-Nearest Neighborhood	Breast Cancer Wisconsin Dataset (from sklearn.dataset import load_breast_cancer)
Naïve Bayes	Titanic Dataset (https://www.openml.org/d/42438)
SVM	Bank marketing Dataset (https://www.openml.org/d/1461)

Obtain model accuracy, generate classification report, and show the performance of the model graphically through roc_auc curve.

9. Build a prediction model using regression technique for (1) Boston house-prices (from sklearn.dataset import load_boston) (2) Diabetes (from sklearn.dataset import load diabetes) datasets. Also, evaluate the model.

10. Implement support vector machine approach to predictive modelling for (1) Boston house-prices (from sklearn.dataset import load_boston) (2) Diabetes (from sklearn.dataset import load diabetes) datasets. Also, evaluate the model.

11. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Titanic dataset (https://www.openml.org/d/42438) to cluster data by

GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Master of Computer Applications, 3rd Semester
Subject Name: Machine Learning
Subject Code: 639402

With effective
from academic
year 2020-21

removing the classlabel.Use elbow method to find the optimal number of clusters. Also, measure the quality of clustering via silhouette coefficient.

12. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Breast Tumour dataset to cluster data (use Breast Tumour dataset) by removing the classlabel.

13. Implement unsupervised machine learning algorithm (Clustering – Hierarchical) in python on Titanic dataset to cluster data (use Titanicdataset).

14. Implement Apriori algorithm in python to find rules which explain association between different products for given transactions at a retail store. (The data is available at[https://drive.google.com/file/d/1NUXoptUlHY8z4KcFKpFA6sQN5KnWzk3p/view?us p=sharing](https://drive.google.com/file/d/1NUXoptUlHY8z4KcFKpFA6sQN5KnWzk3p/view?us%20p=sharing))

Desirable Practical Lists

1. Classification and Prediction algorithms on UCI dataset using Python's scikit-learn library. Perform Exploratory Data Analysis, implement data visualization techniques, pre-process the data, build a model by applying suitable algorithm, evaluate the model by generating report, and show the performance of the model graphically through roc_auc curve.

2. For the sentiment analysis dataset given in link: [https://drive.google.com/file/d/1x6H7_KJjkbDrpgZFS7I2wjsZsILeSJ4S/view?usp=sharing,](https://drive.google.com/file/d/1x6H7_KJjkbDrpgZFS7I2wjsZsILeSJ4S/view?usp=sharing) implement the following in python,

    a) Clean and pre-process the dataset by removing URL, removing HTML tags, handling negation words which are split into two parts, converting the words to lower cases, removing all non-letter characters

    b) Split the dataset into training and testing set

    c) Implement feature extraction technique (to convert textual data to the numeric form)

    d) Build the classification model using Logistic Regression that classifies if a given sentiment text is positive or negative

    e) Obtain the accuracy score of the built model.

3. Implement a content-based recommender system in python that recommends movies that are similar to a particular movie using movielens-20m-dataset available at https://kaggle.com.

The practical exercises should be performed in python.

References:

1) Peter Harrington, "Machine Learning in Action",DreamTech
2) Michael Bowles, "Machine Learning in Python",Wiley
3) Gavin Hackeling, Mastering Machine Learning with scikit-learn,Packt
4) Giuseppe Bonaccorso, Machine Learning Algorithms - Second Edition,Packt