**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: B.Tech CSDS(311)/B.Tech-CSBS

**Course: Machine Learning**

**Experiment No.01**

PART A

(PART A : TO BE REFFERED BY STUDENTS)

**A.1 Aim:** Introduction to Machine Learning and Pandas library

**Task 1:** Create a group of two/three students and identify two/three papers based on Machine Learning Applications. The papers must be starting from 2018 onwards (Google Scholar\IEEE). It is mandatory to have papers on at least three different applications. The summary should be strictly in your own words.

1. Paper Title
2. ML application
3. Category of ML application (Supervised or Unsupervised)
4. Your reasoning for category of ML application
5. Algorithms used
6. Key concepts/ short summary in your own words

**Task 2:** Perform Exploratory data analysis on Indian cuisine dataset and write the inferences for each question.

1. Read the indianfood1.csv file into a DataFrame.
2. Explore size, shape, data types of each column in the dataset.
3. How many total Indian dishes are there?
4. Using Describe function, view the basic statistics of all columns. What Inference you can make out form that?
5. Are there any missing values in the dataset? If Yes, replace the missing values with the NaN values.
6. How many numeric features and categorical features are there in the dataset?
7. Display the number of unique values in each column.
8. Add a new column in the dataset to calculate the total time taken to make every dish.
9. Add a new column in the dataset that will count the number of ingredients from the ingredients column for each dish.

**A.2 Prerequisite:**

Python Programming, Pandas library

**A.3 Outcome:**

**After successful completion of this experiment students will be able to:**

* 1. Differentiate applications of supervised and unsupervised learning
  2. Read different types of data files(csv, excel, text file etc.)
  3. Obtain metadata of given dataset

**A.4 Theory:**

**Machine Learning:**

* Definition by Tom Mitchell (1998):
  + Machine Learning is the study of algorithms that improve their performance P at some task T with experience E.
  + A well-defined learning task is given by <P,T,E>
  + For example: Task (T): identifying correct shape
  + Experience ( E): images of various shapes as an input to the algorithm
  + Performance (P): number of correctly identified shapes out of all the shapes

**Supervised Machine Learning:**

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.

**Unsupervised Machine Learning:**

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA is to learn what our data can tell us. It generally starts out with a high level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modeling choices, such as by helping us decide which features to use.

**Pandas Library:**

key features and functionalities of the Pandas library:

1. DataFrame: The core data structure in Pandas is the DataFrame, which is a two-dimensional, tabular data structure resembling a spreadsheet or SQL table. It organizes data into rows and columns, and you can think of it as a dictionary of Series objects.
2. Series: A Series is a one-dimensional labeled array in Pandas, and it is the building block of a DataFrame. It is similar to a NumPy array but has additional functionality and a labeled index, allowing for more flexible and intuitive data manipulation.
3. Reading and Writing Data: Pandas supports reading and writing data from various file formats, including CSV, Excel, SQL databases, and more. The read\_csv(), read\_excel(), read\_sql(), and related functions make it easy to import data into a DataFrame.
4. Data Cleaning: Pandas provides various methods to handle missing data, duplicate rows, and data manipulation tasks. You can use functions like dropna(), fillna(), drop\_duplicates(), and more for data cleaning.
5. Data Selection and Slicing: Pandas allows you to access, slice, and filter data efficiently using labels, row indices, and conditional selections. You can use indexing, boolean masks, and various selection methods like loc[], iloc[], and boolean indexing.
6. Grouping and Aggregation: Pandas offers powerful tools for grouping data based on specific columns and performing aggregate operations like sum, mean, count, etc., on grouped data using groupby() and agg() functions.
7. Merging and Joining Data: You can merge multiple DataFrames based on common columns using functions like merge() and concat(), enabling you to combine data from different sources.
8. Time Series Functionality: Pandas has robust support for time series data, providing features like date/time parsing, resampling, time zone handling, and more.
9. Data Visualization: While Pandas itself does not handle data visualization, it integrates well with popular data visualization libraries like Matplotlib and Seaborn, making it easy to create insightful plots and charts from DataFrame data.

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical.)***

|  |  |
| --- | --- |
| Roll No.: C026 | Name: Anirbaan Ghatak |
| Class : B | Batch : B1 |
| Date of Experiment: 25/07/2023 | Date of Submission: 02/08/2023 |
| Grade : |  |

**B.1 Task1**

Research Paper 1:

1. Paper Title:

“Machine Learning and Deep Learning Applications-A Vision”

1. ML applications:
   1. Machine learning-based face recognition technology is used to recognize extremists in the crowded places from the visitors at congress centres, airports, and various other important events
   2. Handwriting recognition application makes the work easy for organizations where handwritten documents are large. For example, universities, exam centres, police, etc. It is a process of scanning and digitizing documents in a few minutes.
   3. Diagnosing patients, identifying who is at most risk, better understand viruses, predict the spread of the disease, Map from where the viruses come, discovering existing drugs that can help, developing drugs at the fastest speed, Predicting the next pandemic is the crucial task done by machine learning.
2. Category of ML application (Supervised or Unsupervised):

Supervised

1. Your reasoning for category of ML application
2. Face Recognition to recognize extremists uses supervised learning as it would help to categorize it more accurately as if the face recognition gives raises a flag which turns out to be false it would be a very disturbing affair.
3. Handwriting recognition typically uses supervised learning techniques. In supervised learning, the model is trained on a labeled dataset, where each input in this case, an image of handwritten text is associated with the corresponding correct output the actual text that the image represents
4. Diagnosing patients uses supervised learning, the model is trained on a labeled dataset, where each patient's data includes a set of input features (e.g., medical history, symptoms, test results) and the corresponding correct diagnosis or outcome
5. Algorithms used
   1. Back Propagation Algorithm
   2. Random Forest
   3. KNN, SVM, Naïve Bayes
6. Key concepts/ short summary in your own words

The paper provides an overview of machine learning and deep learning applications in various domains, with a focus on their relevance during the COVID-19 pandemic. It discusses different approaches in machine learning, including supervised, unsupervised, and reinforcement learning. The paper highlights the evolution of machine learning and its applications in computer vision, face recognition, handwritten recognition, fraud detection, and predicting the spread of COVID-19. It also explores the concept of deep learning, its advantages in feature extraction and scalability, and its challenges. The paper further discusses the adoption of machine learning and deep learning by companies during the pandemic, showcasing examples of AI and deep learning techniques used for prediction, analysis, and communication with customers. Specific applications of deep learning, such as speech recognition, image searching, and language processing, are explored

Research Paper 2:

1. Paper Title

Machine learning-based approach: global trends, research directions, and regulatory standpoints

1. ML application
   1. ML techniques to build cybersecurity models useful for detecting and protecting data, with minimal human intervention
   2. ML techniques for predictive modeling and the design of nanoproducts. ML can boost and reshape the de novo design of nanodelivery systems, thus generating new challenges for the next generation of smart drugs
   3. ML algorithms in farm management systems provide insightful advice and information on crop management, yield prediction, the identification of possible diseases and weed species, livestock management and welfare, (5) water and soil management, the level of soil moisture, seeding and harvesting dates, and the phenostages of crops.
2. Category of ML application (Supervised or Unsupervised)

Both are used in this case, supervised and unsupervised.

1. Your reasoning for category of ML application
   1. Cybersecurity uses both the methods supervised to allow the machine to learn about the current prevention techniques and then unsupervised to devise new ways to find different methods to tackle same or different attacks.
   2. Nanotechnology uses unsupervised learning as there isn’t much to go on with supervised learning and needs new, different and abrasive methods to help in the research and development of nano tech
   3. Supervised Learning is used for Farming as the soil moisture, seed health, etc are to be of specific point or nature for them to grow.
2. Algorithms used
   1. Naïve Bayes
   2. Support Vector Machines
   3. Logistic Regression
   4. K-means clustering
   5. Principle analysis
   6. Autoencoders
3. Key concepts/ short summary in your own words

This paper provides an overview of machine learning (ML) trends, research directions, and regulatory considerations. It highlights the rapid growth of ML in various fields, including healthcare, cybersecurity, data governance, and nanotechnology. The paper discusses different types of ML techniques, such as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. It also explores real-world applications of ML and addresses challenges and regulatory issues in managing ML technology. The concept of federated learning is discussed, along with its applications in different domains. The paper emphasizes the importance of data and learning algorithms in achieving successful ML techniques and highlights the global interest in AI and ML. It concludes that ML-based solutions have promising potential worldwide but emphasizes the need for careful data management.

**B.2 Task 2**

import pandas as pd

import numpy as np

#Loading the CSV file into a DataFrame:

df = pd.read\_csv('indian\_food.csv')

# Printing the size, shape, and data types of the DataFrame:

print(f'size: {df.size}')

print(f'shape: {df.shape}')

print(f'data types of each column in the dataset:\n{df.dtypes}')

#Getting the total number of Indian dishes

x, y = df.shape

print("Total no of Indian dishes: ", x)

# Replacing empty spaces with NaN

df.replace(' ', np.nan, inplace=True)

#Selecting and printing numerical columns

df.select\_dtypes(include=['float64', 'int64']).columns

# Printing the number of unique values in each column

for col in df.columns:

    print(col, df[col].nunique())

# Adding a new column for the total time

df['total time'] = df['prep\_time'] + df['cook\_time']

print(df['total time'].head(10))

# Finding the number of ingredients for each recipe

df['num\_ingredients'] = df['ingredients'].apply(len)

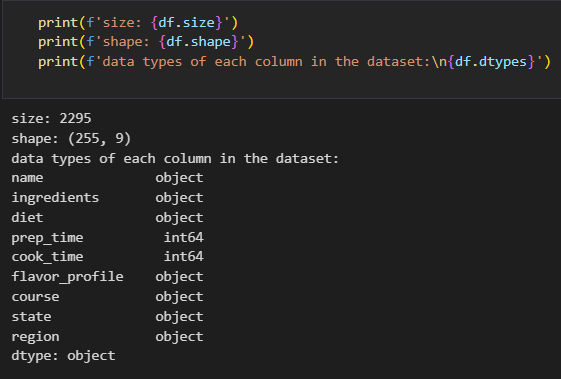
print(df['num\_ingredients'].head(10))

**Output**

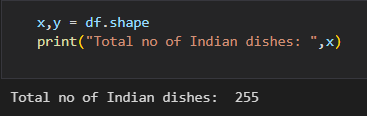
1.



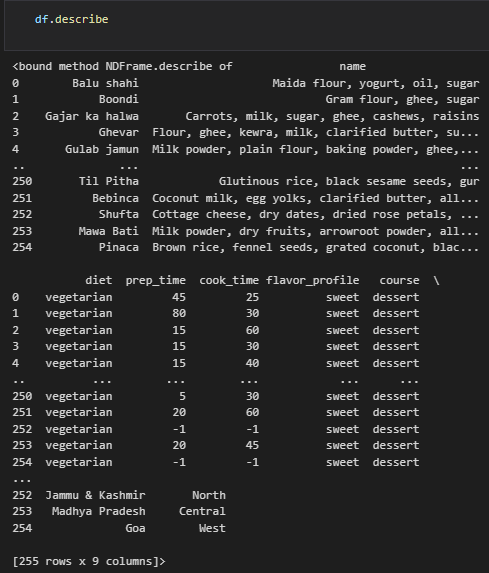
2.



3.

­

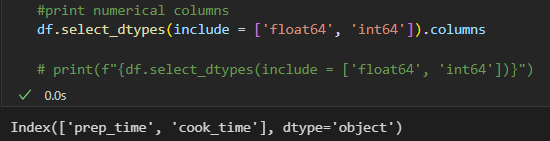
4.



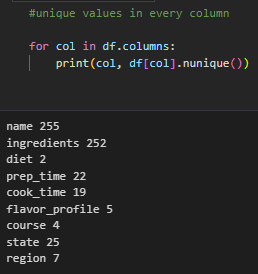
5.



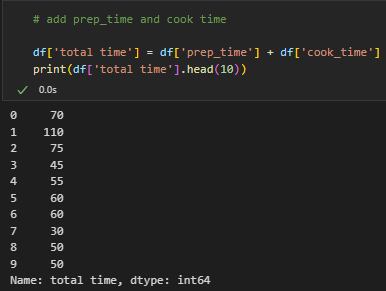
6.



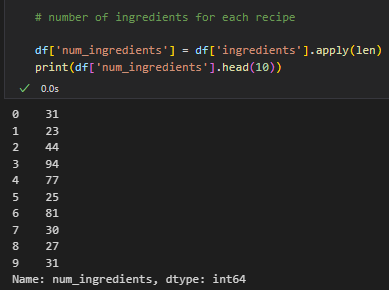
7.



8.



9.



**B.4 Conclusion:**

Revised the concepts of Supervised and Unsupervised Learning to better under the research paper that were to be search for. After finding appropriate research papers which fulfilled the criteria, a thorough read of it was done to provide a summary of it. Task 2 required the use of Pandas Library to perform the tasks mentioned, which was made to give real life experience of handling with data used for machine learning.