Alexandria University
Faculty of Computer and Data Science
General Department

Course Title: Data Mining
2022-2023

Datamining Final Project

Dear survivors, I hope you are doing well.

This is the specification for the final project that will be interpreted in this document for Data Mining course 2022/2023.

Introduction:

In this data-centric world, data mining projects hold great importance in everyday life. It provides us with a reliable source of resolving tough problems and different issues in this challenging world. Some of the benefits are:

- With the help of new and legacy systems, data mining helps in making well-informed decisions.
- It offers cost-effective solutions compared to other applications designed with other technologies.
- It helps data scientists to deal with huge amounts of data and scrutinize the essential data out of it.
- It makes businesses make profitable production and operational adjustments according to the demand.

How do you create a data mining project?

To create a data mining project, follow these steps:

- Understand business and project's objective
- Understand the problem deeply and collect data from proper sources.
- Cluster the essential data to resolve the business problem.
- Prepare the model using algorithms to ascertain data patterns.
- Evaluate the data according to the business goal or to find a remedy for the problem.
- Last, deploy the solution and get the results to make decisions.

What are the different tasks associated with data mining?

The following activities are performed for data mining.

- Classification
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery
- Regression
- Deviation Detection

What tools are used in data mining?

Top tools used in data mining are

- Rapid Miner
- Orange
- Kaggle
- Rattle
- Oracle Data Mining
- IBM SPSS Modeler
- Knime
- Python
- Weka
- Teradata

Project Description:

You will use the libraries of python and Rapid Miner tools for this project, and you should choose one of this Project to implement it:

1- Smart Health Disease Prediction Using Naive Bayes.
2- Online Fake Logo Detection System.
3- Color Detection.
4- Handwritten Digit Recognition.
5- Anime Recommendation System.
6- Evaluating and Analyzing Global Terrorism Data.
7- Prediction of Adult Income based on Census Data.
8- Music Genres Classification using KNN System.
9- Diabetes Prediction.
10- Crime Rate Prediction Using K Means.
11- Students Performance.
12- The price elasticity of all cafe items.
13- San Francisco Salaries.

14- Netflix movie ratings prediction.
15- Customers' preferences.

---

1. Smart Health Disease Prediction Using Naive Bayes:

Project Idea:

Nowadays, medical care is something that anyone might need immediately, but unavailable due to various reasons. Smart health disease prediction is an end user support system that allows users to get guidance immediately with the help of an online intelligent health system. The system holds complete information about symptoms and the diseases associated with it. The system analyzes diseases associated with the symptoms for the patient and advises them for X-ray, blood test or CT scan as requested by the system. Users can also directly get in touch with the specialist doctors for any ailment and share their reports. It is not just one time, rather a proper login detail is shared for future use.

Dataset:

https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning

---

2. Online Fake Logo Detection System.

Project idea:

Each year, thousands of brands lose a huge portion of their sales due to unauthorized knock off brands and their counterfeits. These counterfeit products are made of inferior quality and hence damage the credibility of the brand. Moreover, consumers feel cheated with their hard-earned money while shelling it out for just a mere counterfeit. Online fake logo detection system will distinguish between original product and forgeries for the consumers. Along with helping users to fight against forged products, it also helps brands to combat piracy.

Dataset:

Real or Fake logo | Kaggle

---

3. Color Detection:

Project idea:

There are around 16 million colors according to different RGB color values, but a human mind can only remember quite a few. It is common that after seeing the color, you are still not able to name the color. In this data mining project, you are going to build an amazing app which is going to help in recognizing color from any image. All you need is a labeled data of available colors and then the program runs to evaluate which color resembles most with the selected color value and helps in detecting colors easily. You can use the Python programming language.

Dataset:

color classification | Kaggle

---

4. Handwritten Digit Recognition.

Project Idea:

One of the best data mining projects is the Handwritten Digit recognition project among the data scientists and all the machine learning enthusiasts. In this project, machine learning algorithms are used to distinguish and classify images of the digits written by hand. With the help of computer vision AI model, machine learning techniques and Convolutional Neural Networks, this project can be created which will have a nice graphical user interface to write or draw on the canvas and for the output a model is good to predict the digit. Python's Scikit-learn model using algorithms such as K-Nearest Neighbors and a Support Vector Classifier will be apt for the project or using neural network.

Dataset:

MNIST Data for Digit Recognition | Kaggle

---

5. Anime Recommendation System.

Project Idea:

The Anime Recommendation system is one of the best projects as it includes a data set containing information regarding user preference from 73,516 users on 12,294 anime. Every user in the database will be able to add anime to the list and share ratings compiling a data set with those ratings. The anime recommendation system project helps in creating a system that produces efficient data based on the user viewing history and sharing rating.

Dataset:

Anime Recommendations Database | Kaggle

---

6. Evaluating and Analyzing Global Terrorism Data.

Project idea:

Terrorism has mushroomed due to its deep roots in certain locations of the world. With increase in its activities, it is important to stop its spread or analyze the global terrorism data to identify the terrorist activities. Internet plays a major role in spreading terrorism by way of videos and speeches among youth to join the terrorist organizations. This project will help in detecting, evaluating, and analyzing global terrorism data and flag them for human review. Data mining helps in scanning and mining from all the unorganized and unstructured pages or data available that promotes terrorism and flag them.

Dataset:

Global Terrorism Database | Kaggle

---

7. Prediction of Adult Income based on Census Data:

Project Idea:

The following project is the classification project to predict the income level of an individual that exceeds 50K based on the census data available at the repository. The dataset that is used in the projects are variables such as age, type of work, working hours, sex and many more. It helps in understanding the standard of living of the city, and the benefit of setting up a business or bank loan eligibility. Also, it helps in understanding the real estate preferences by average income of the people residing in the area. In this project, you will also be able to figure out the type of tourist places that people from other countries would like to travel to.

Dataset:

Adult Census Income | Kaggle

---

8. Music Genres Classification using KNN System.

Project Idea:

One way of categorizing and organizing music is based on the genre, which is identified by some characteristics of the music such as rhythmic structure, harmonic content and instrumentation. Being able to automatically classify and provide tags to the music present in a user's library,

based on genre. A music genre is a conventional category that predicts the genre of music belonging to tradition or a set of conventions. Categorizing music files according to their genre is a challenging task in music information retrieval.

Our Music Genres Classification System will detect the music from the audio file. Once the music is detected, the system will further continue to classification. As a result, the system will display the music genre. So, for this system, there is a predefined set of music genres that the system will classify.

Dataset:

Music Genre Classification | Kaggle

---

9. Diabetes Prediction:

Project Idea:

Diabetes is one of deadliest infections on the planet. It isn't just an ailment yet additionally a maker of various types of maladies like heart assault, visual deficiency, kidney infections, and so on. The typical recognizing process is that patients need to visit an indicative focus, counsel their specialist, and sit tight for a day or more to get their reports. Also, every time they need to get their conclusion report, they need to squander their cash futile. Be that as it may, with the ascent of Machine Learning approaches we can discover an answer for this issue, we have built up a framework utilizing information mining which can anticipate whether the patient has diabetes or not. Moreover, foreseeing the illness early prompts treating the patients previously it winds up basic. Information mining can remove concealed learning from a colossal measure of diabetes-related information. Therefore, it has a critical part in diabetes examine, now like never before. The point of this exploration is to build up a framework which can anticipate the diabetic hazard level of a patient with a higher exactness. This exploration has concentrated on building up a framework in light of three order techniques to be specific, Decision Tree, Naïve Bayes, and Support Vector Machine calculations.

Dataset:

Diabetes Prediction | Kaggle

---

10. Crime Rate Prediction Using K Means:

Project idea:

The crime rate is increasing now-a-days in many countries. In today's world with such a higher crime rate and brutal crime happening, there must be some protection against this crime. Here we introduced a system by which the crime rate can be reduced. Crime data must be fed into the system. We introduced data mining algorithms to predict crime. K-means algorithm plays an important role in analyzing and predicting crimes. K-means algorithm will cluster co-offenders,

collaboration and dissolution of organized crime groups, identifying various relevant crime patterns, hidden links, link prediction and statistical analysis of crime data. This system will prevent crime occurring in society. Crime data is analyzed which is stored in the database. Data mining algorithm will extract information and patterns from database. System will group crime. Clustering will be done based on places where crime occurred, gang who involved in crime and the timing crime took place. This will help to predict crime which will occur in future. Admin will enter crime details into the system which is required for prediction. Admin can view criminal historical data. Crime incident prediction depends mainly on the historical crime record and various geospatial and demographic information.

Dataset:

SF Crime Rate Prediction | Kaggle

---

11. Students Performance:

Project idea:

It will not be difficult for most of us to appreciate that a class in any school never has students of the same kind. Each student has an individual personality that defines their behavior and interests. Not all of them are good at academics. It is thus an exciting task to work on the dataset of a class and analyze student performances.

Dataset:

Students Performance in Exams | Kaggle

---

12. The price elasticity of all cafe items:

Project idea:

Deciding the items and their prices on a menu card is not an easy task for cafe owners. They have to constantly analyze their customers' choices to set the optimum prices of their food items on the menu.

Using the dataset mentioned above, you can verify a few fundamental economic trends in the dataset as a first step. These trends will include analyzing price trends and sales of all the items, sales on special holidays and weekends, and more such trends. You can draw more insights by visualizing the dataset through the seaborn library of the Python Programming Language. Another metric that you must evaluate for this project is the Price Elasticity of all cafe items.

Dataset:

Shared with you - OneDrive (live.com)

---

13. San Francisco Salaries:

Project Idea:

For this project, you can use the San Francisco Salaries Dataset to understand the income inequality in San Francisco city. In addition, you can also analyze the factors responsible for the promotions of certain employees. It would be easy to use the python programing language for this project and visualize the datasets, to look at the distribution of the salaries.

Dataset:

Salaries for San Francisco Employee | Kaggle

---

14. Netflix movie ratings prediction:

Project Idea:

This project is an example of performing data mining techniques on a dataset of Netflix movies and TV shows using Python libraries and machine learning techniques. The project explores the data using descriptive statistics and visualizations and uses machine learning models to predict movie ratings. The project demonstrates the power of data mining and analysis in understanding trends and making predictions in the entertainment industry.

Dataset:

Netflix Dataset | Kaggle

---

15. Customers' preferences:

Project idea:

For a company, analyzing its customers' preferences is very important. Most companies have now started mining customers' data to understand their customers' choices and behaviour better. This approach helps them recommend appropriate products to their customers and inventory management of their warehouses.

Dataset:

FoodMart Dataset | Kaggle

Conclusion:

To cut a long story short, data mining is the process of analyzing huge chunks of data to discover business intelligence which helps in solving problems, seizing new opportunities, and mitigating long term risks. The process of discovering useful patterns and relationships in large volumes of data helps in understanding a problem deeply and tactics to deal with it diligently. It is widely used in research, medical, business and security to turn large data into useful information. Get started from the above list of projects from beginner to advanced and sharpen your skills. These data mining projects with source code will help in learning new abilities.

Project Details:

Please read the following guide carefully:

- Cheaters 😡 will get negative grades for the project (we will have a cheaters list) Delivering an incomplete project 😔 will be much better than cheating, the lowest possible grade without cheating will be more than the highest possible grade with cheating.
- Each group should consist of 5.

Project Requirements:

- Data Preprocessing (Cleaning Data, Normalization)
- Split your data set into train data and test data (If the data is not split)
- You will use a suitable algorithm for your dataset. (Classification, Regression, Clustering, Association, etc...).
- The programming language you will be using is python and its libraries (Pandas, Numby, Matplot, Seaborn, Scikitlearn, etc...).
- After prediction, you will work on Evaluation metric.
- PowerPoint presentation to represent why you chose this project.
- You will choose the suitable visualization for your project by using Matplot library from python. (Histogram, pie chart, box plot, scatterplot, etc...).
- Make a README file to explain your code and don’t write your code in this file.

README file:

- A README is a text file that introduces and explains a project. It contains information that is commonly required to understand what the project is about.
- It's an easy way to answer questions that your audience will likely have regarding how to install and use your project and also how to collaborate with you.
- Definitely before you show a project to other people or make it public. You might want to get into the habit of making it the first file you create in a new project.

Here are some guides for making a README file in this website:

How to Write a Good README File for Your GitHub Project (freecodecamp.org)

Do not hesitate to contact us via email if you face any problem.

Good luck ♥