Project 3 Term project/open topic (100 points)

Please follow the exact deadlines, no extension can be granted.

Phase 1 (20 points) – team up (2 person in each group) and find a data science task to work on, some suggestions are listed on 2nd page of this description.

Due date 4/18

Things to submit: A project proposal that includes:

1. the title
2. team member names
3. description of the project
4. data set that will be used
5. techniques/platforms that will be used (this can be changed later)

Phase 2 (60 points) – design and implementation

Due date 5/2

Things to submit:

1. All the source codes and programs developed in the project
2. A report that includes:
    a. the detailed description of the procedures used to solve the problem
    b. the techniques, algorithms, strategies that are used
    c. the result
    d. conclusion and possible future work
    e. a table indicates the work completed by each individual

Phase 3 (20 points) – presentation (video recording)

Due date 5/9

Things to submit:

A video recording link that explains your solution to the proposed project. (The video can be uploaded to youtube or other platforms, **please make sure it can be accessed by anyone that has the link**)

Extra requirement for graduate students:

At least two algorithms/strategies need to be used to solve the proposed problem and a comparison of corresponding results needs to be included in the report.

Possible topics

Sentiment analysis of product reviews

This involves analyzing a data set and creating visualizations to better understand the data. For instance, a project idea may be to examine user evaluations of products on Amazon using natural language processing (NLP) methods to ascertain the general mood toward such things. To accomplish this, a sizable collection of product reviews from Amazon can be gathered by using web scraping methods or an Amazon product API.

Once the data has been gathered, it can be preprocessed by having stop words, punctuation and other noise removed. The polarity of the review, or whether the sentiment indicated in it is favorable, negative or neutral, can then be determined by applying a sentiment analysis algorithm to the preprocessed language. In order to comprehend the general opinion of the product, the results might be represented using graphs or other data visualization tools.

Predicting house prices

This project involves building a machine learning model to predict house prices based on various factors such as location, square footage, and the number of bedrooms.

Using a machine learning model that uses housing market data, such as location, the number of bedrooms and bathrooms, square footage and previous sales data, to estimate the sale price of a particular house is one example of a data science project connected to predicting house prices.

The model could be trained on a data set of past house sales and tested on a separate data set to evaluate its accuracy. The ultimate objective would be to offer perceptions and forecasts that might help real estate brokers, buyers and sellers make wise choices regarding price and buying/selling tactics.

Customer segmentation

A customer segmentation project involves using clustering algorithms to group customers based on their purchasing behavior, demographics and other factors.

A data science project related to customer segmentation could involve analyzing customer data from a retail company, such as transaction history, demographics and behavioral patterns. The goal would be to identify distinct customer segments using clustering techniques to group customers with similar characteristics together and identify the factors that differentiate each group.

This analysis could provide insights into customer behavior, preferences and needs, which could be used to develop targeted marketing campaigns, product recommendations and personalized customer experiences. By increasing customer satisfaction, loyalty and profitability, the retail company can benefit from the results of this project.

Fraud detection

This project involves building a machine learning model to detect fraudulent transactions in a data set. Using machine learning algorithms to examine financial transaction data and spot patterns of fraudulent activity is an example of a data science project related to fraud detection.

The ultimate objective is to create a reliable fraud detection model that can assist financial institutions in preventing fraudulent transactions and safeguarding the accounts of their consumers.

Image classification

This project involves building a deep learning model to classify images into different categories. An image classification data science project could involve building a deep learning model to classify images into different categories based on their visual features. The model could be trained on a large data set of labeled images and then tested on a separate data set to evaluate its accuracy.

The end goal would be to provide an automated image classification system that can be used in various applications, such as object recognition, medical imaging and self-driving cars.

Recommendation system

This project involves building a recommendation system to suggest products or content to users based on their past behavior and preferences.

A recommendation system project could involve analyzing Netflix user data, such as viewing history, ratings and search queries, to make personalized movie and TV show recommendations. The goal is to provide users with a more personalized and relevant experience on the platform, which could increase engagement and retention.

Other resources:

Kaggle competition

Kaggle dataset

Corgis – a civic dataset

Recommender Systems and Personalization Datasets