Amrita Vishwa Vidyapeetham, Coimbatore
Amrita School of Computing
Department of Computer Science and Engineering

19CSE305 Machine Learning
V Sem B.Tech. CSE (2020-24 Batch)

CAPSTONE PROJECT PROBLEM STATEMENTS

Sl. No.	Problem Statement	Faculty
1	**Title**: Music recommendation **Description**: The purpose of a recommendation system like this would be to give customised material by correctly understanding what the user wants. This means that computers will have to think like humans, analysing each user's previous selections to forecast what they want in the future. As we have seen, if a computer is required to emulate human behaviour, Machine Learning techniques must be used. As a result, in this project, we will use Machine Learning techniques to create the ideal music recommendation system. **Dataset Available:** 1. WSDM – KKBox’s Music Recommendation: KKBOX provides a training data set consisting information of the first observable listening event for each unique user-song pair within a specific time duration. 2. Last.FM: This dataset contains social networking, tagging, and music artist listening information from a set of 2k users from Last.fm online music system.	MA
2	**Title**: Stock Price Predictions **Description**: Predicting stock prices is a difficult endeavour since they are affected by a variety of factors, including but not limited to geopolitics, the global economy, a company's financial reports and performance, and so on. There are two basic methods for forecasting stock prices: For stock prediction, technical analysis uses metrics such as closing and opening price, volume traded, adjacent close values, and so on, whereas qualitative analysis looks at external factors such as company profile, market situation, political and economic factors, textual information in news, social media, and even blogs by the economic analyst.	MA

	Dataset Available: 1. Huge Stock Market Dataset: The dataset is a collection of the daily prices and volumes of all US stocks and ETFs. Get the dataset here. 2. Daily News for Stock Market Prediction: The dataset is a collection of historical news headlines from Reddit WorldNews Channel and stock data. Get the data here.	
3	**Title**: Sports Prediction **Description:** Sports prediction is often regarded as a classification issue, with one class to be predicted (win, lose, or draw). A wide variety of elements, such as club history, match outcomes, and player statistics, must be considered in sports prediction to enable diverse stakeholders comprehend the probabilities of winning or losing. **Dataset Available:** 1. ATP World Tour tennis data: This dataset contains tennis data from the ATP World Tour website. 2. FIFA 19 Dataset: FIFA 19 complete player dataset is a collection of detailed attributes for every player registered in the latest edition of FIFA 19 database.	MA
4	**Title: Movie Genre Classification** **Description:** Dataset contains 5 list of movies for Genre - Action, Sci-fi, Comedy, Animation, Adventure which has been taken from IMDb site. Fields/columns - {Movie title: name of the movie; Released year - year it got released; IMDb rating: Movie rating given by IMDb} This dataset can be used for exploratory data analysis (EDA) and data cleaning or for model building. **Dataset Available:** 1. 5 files – one for Each Genre is available in Kaggle 2. Popular Movies Datasets - 58098 Movies from Kaggle. This is a movie rating dataset with a file relating a greater number of Genre classes. It can also be used for the given project tile.	CB

5	**Title: Social network Advertisement Impact analysis** **Description:** Social networks are populated with Advertisements in the aim that users will involve in a purchase just by seeing the ad. The efficiency of the Advertisement is analyzed through how better it gains the attention of the social network user. **Dataset Available:** 1. A categorical dataset to determine whether a user purchased a particular product. The dataset has details of user, age, salary and purchase status after being exposed to an Advertisement.	CB
6	**Title: Steel Industry Energy Consumption Dataset** The information gathered is from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea. It produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown. **Dataset Available:** 1. UCI Machine Learning Repository	CB
7	**Title:** AI-driven Sentiment Analyzer **Description**: Millions of user-generated postings and information are flooding social media. The majority of users use social media to express their personal ideas, thoughts, and feelings through photos and text. The largest issue, however, is precisely interpreting the thoughts or sentiments underlying such user-generated messages. Modern social media companies such as Snapchat, Facebook, Linked In, Twitter, and others, as well as online meal delivery businesses, are investing heavily in projects to better comprehend human attitudes and moods. These companies would be able to detect user behaviour more quickly if they analysed the phrases and visuals to grasp the feelings. Based on such data, businesses may provide better customer service, hence boosting customer happiness. Sentiment analysis projects need a thorough grasp of areas such as text analysis, natural language processing (NLP), and computational linguistics. Huggingface and TensorFlow machine learning frameworks come in helpful, as can supervised algorithms such as neural networks, Random Forest, decision trees, Support Vector Machines (SVM), and logistical regression. OpenCV and SimpleITK libraries aid	MA

	with picture segmentation, registration, and object identification. **Dataset**: 1. https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews 2. https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe	
8	**Title : Stroke Prediction Dataset** According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient. **Dataset available: Kaggle** Attribute Information 1) id: unique identifier 2) gender: "Male", "Female" or "Other" 3) age: age of the patient 4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension 5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 6) ever_married: "No" or "Yes" 7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed" 8) Residence_type: "Rural" or "Urban" 9) avg_glucose_level: average glucose level in blood 10) bmi: body mass index 11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"* 12) stroke: 1 if the patient had a stroke or 0 if not *Note: "Unknown" in smoking_status means that the information is unavailable for this patient	AT

9	**Title:Airline Passenger Satisfaction** The dataset contains an airline passenger satisfaction survey. What factors are highly correlated to a satisfied (or dissatisfied) passenger? Predict passenger satisfaction. Gender: Gender of the passengers (Female, Male) Customer Type: The customer type (Loyal customer, disloyal customer) Age: The actual age of the passengers Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel) Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus) Flight distance: The flight distance of this journey Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5) Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient Ease of Online booking: Satisfaction level of online booking Gate location: Satisfaction level of Gate location Food and drink: Satisfaction level of Food and drink Online boarding: Satisfaction level of online boarding Seat comfort: Satisfaction level of Seat comfort Inflight entertainment: Satisfaction level of inflight entertainment On-board service: Satisfaction level of On-board service Leg room service: Satisfaction level of Leg room service Baggage handling: Satisfaction level of baggage handling Check-in service: Satisfaction level of Check-in service Inflight service: Satisfaction level of inflight service Cleanliness: Satisfaction level of Cleanliness Departure Delay in Minutes: Minutes delayed when departure Arrival Delay in Minutes: Minutes delayed when Arrival Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction) Note that this data set was modified from this dataset by John D here. It has been cleaned up for the purposes of classification.	AT
10	**Title: News Category identification** Identify the type of news based on headlines and short descriptions. This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost. This is one of the biggest news datasets and can serve as a benchmark for a variety of computational linguistic tasks. HuffPost stopped maintaining an extensive archive of news articles sometime after this dataset was first collected in 2018, so it is not possible to collect such a dataset in the present day. Due to changes in	AT

	the website, there are about 200k headlines between 2012 and May 2018 and 10k headlines between May 2018 and 2022. **Dataset**: Kaggle Content Each record in the dataset consists of the following attributes: • category: category in which the article was published. • headline: the headline of the news article. • authors: list of authors who contributed to the article. • link: link to the original news article. • short_description: Abstract of the news article. • date: publication date of the article.	
11	**Title: Forest cover prediction** Predicting Forest Cover Types with the Machine Learning https://archive.ics.uci.edu/ml/datasets/covertypev To predict seven different cover types in four different wilderness areas of the Roosevelt National Forest of Northern Colorado with the best accuracy. **Attribute Information** 1. Name 2. Elevation 3. Aspect 4. Slope 5. Horizontal_Distance_To_Hydrology 6. Vertical_Distance_To_Hydrology 7. Horizontal_Distance_To_Roadways 8. Hillshade_9am 9. Hillshade_Noon 10. Hillshade_3pm 11. Horizontal_Distance_To_Fire_Points 12. Wilderness_Area (4 binary columns) 13. Soil_Type (40 binary columns) 14. Cover_Type (7 types)	VD
12	**Title: Predict Ads Click** Indicating whether a particular internet user clicked on an advertisement. https://github.com/shubham13p/Ad-Click-Prediction/blob/master/advertising.csv **Attribute Information** 1. 'Daily Time Spent on Site': consumer time on site in minutes 2. 'Age': customer age in years 3. 'Area Income': Avg. Income of geographical area of consumer 4. 'Daily Internet Usage': Avg. minutes a day consumer is on the internet	VD

	5. ‘Ad Topic Line’: Headline of the advertisement 6. ‘City’: City of consumer 7. ‘Male’: Whether the consumer was male 8. ‘Country’: Country of consumer 9. ‘Timestamp’: Time at which consumer clicked on Ad or closed window 10. ‘Clicked on Ad’: 0 or 1 indicated clicking on Ad	
13	**Title: Glass Identification** Predicting the type of Glass using ML algorithms. https://archive.ics.uci.edu/ml/datasets/glass+identification **Attribute Information** 1. d number: 1 to 214 2. RI: refractive index 3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10) 4. Mg: Magnesium 5. Al: Aluminum 6. Si: Silicon 7. K: Potassium 8. Ca: Calcium 9. Ba: Barium 10. Fe: Iron 11. Type of glass: (class attribute)	VD
14	**Title: Customer Churn Prediction Analysis Using Ensemble Techniques in Machine Learning** **Description**: Customers are a company's most valuable asset, and maintaining customers is critical for any organisation looking to increase revenue and develop long-term meaningful relationships with customers. Furthermore, the cost of obtaining a new client is five times that of keeping an existing customer. Customer Churn/Attrition is one of the most well-known business difficulties in which consumers or subscribers discontinue doing business with a service or a firm. Ideally, they will no longer be a paying customer. A client is considered to have been churned if a certain length of time has passed since the consumer last interacted with the company. Identifying whether or not a client will churn and offering relevant information aimed at customer retention are crucial to lowering churn. Our brains cannot anticipate customer turnover for millions of clients; here is where machine learning may assist. **Dataset**: https://www.kaggle.com/code/kerneler/starter-wa-fn-usec-telco-customer-05c825b4-5/data	MA

15	**Title**: Sales Forecasting using Walmart Dataset **Description**: One of the most prominent applications of machine learning is sales forecasting, which involves detecting characteristics that influence product sales and anticipating future sales volume. This machine learning experiment takes use of the Walmart dataset, which contains sales data for 98 goods from 45 different locations. Weekly sales by store and category are included in the dataset. This machine learning project's purpose is to anticipate sales for each department in each outlet to assist them in making better data-driven decisions for channel optimization and inventory planning. The challenging aspect of working with the Walmart dataset is that it contains selected markdown events that affect sales and should be taken into consideration. **Dataset**: https://www.kaggle.com/datasets/yasserh/walmart-dataset	MA