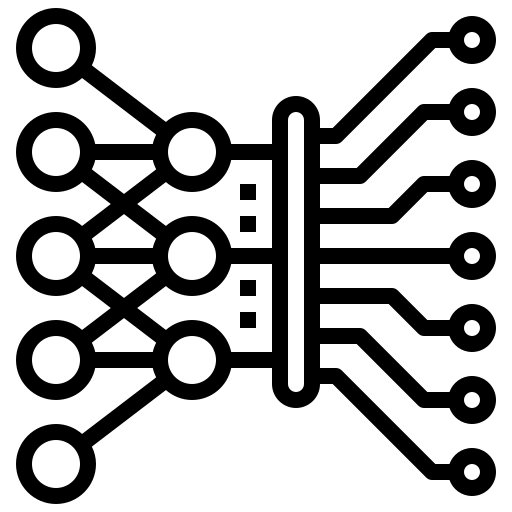
****

**School of Engineering & Technology**

**Asian Institute of Technology**

**AT82.03 - Machine Learning**

****

**Date: 10 October 2023**

**Group Project <Group 11>**

**Submitted To: Submitted By:**

# **Dr. Chaklam Silpasuwanchai Mr. Gholamreza Izadi**

# **Mr. Tairo Kageyama**

# **Mr. Rojan Manandhar**

# **Mr. Bakyt Tursaliev**

# **Mr. Panithi Sirisatjapipat**

# **Mr. Nitesh Ghimire**

# **Table 1. Grocery data**

| Description | Comprehensive grocery product info from Real Canadian Superstore, Canada. |
| --- | --- |
| Potential Applications | Price Comparison, Market Basket Analysis, Inventory Management, Data Visualization. |
| Related Industry | E-commerce and retail; Unique focus on Real Canadian Superstore for tailored insights |
| Impact | Customers: Informed shopping; Retailers: Efficient operations and competitive edge |
| Value Proposition | Informed Decision-Making, Efficient Retail Ops, Enhanced Customer Exp, Competitive Adv., Research & Insights. |

<https://www.kaggle.com/datasets/maximsakhan/rc-superstore-grocery-data/data>

# 

# **Table 2. Supermarket data**

| Description | Consumer behavior data from Hunter's e-grocery, a lifestyle brand in 10 counties |
| --- | --- |
| Potential Applications | K-Means Clustering & PCA for customer segmentation, Apriori & Causal ML for product recommendation |
| Related Industry | Similar applications in e-commerce; Unique focus on diverse customer behaviors across countries |
| Impact | Customers benefit from personalized recommendations; Retailers optimize inventory and marketing. |
| Value Proposition | Informed Decision-Making, Improved Customer Experience, Adaptability to Market Events. |

<https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023>

<https://www.kaggle.com/datasets/yapwh1208/supermarket-sales-data>

# 

# 

# **Table 3. Nutritional facts for most common foods**

| Description | Nutritional content of 300+ foods, including Calories, Fats, Proteins, and more. |
| --- | --- |
| Potential Applications | Unsupervised learning for Dietary planning, Nutrition apps, Healthcare guidance, Food industry product development. |
| Related Industry | Used in nutrition and healthcare; Unique for diverse food categories and nutrients. |
| Impact | Individuals make healthier choices, Healthcare professionals advise better, Food industry benefits |
| Value Proposition | Informed eating, Versatile applications, Valuable nutritional data source. |

<https://www.kaggle.com/datasets/niharika41298/nutrition-details-for-most-common-foods/data>

# 

# **Table 4. Million song data**

| Description | Spotify Million Song Dataset with song names, artist names, links, and lyrics. |
| --- | --- |
| Potential Applications | Song name predictions from the lyrics, Song recommendation, Song classification, Song clustering based on lyrics and metadata. |
| Related Industry | Common in music streaming and recommendation services; Unique for a large song dataset. |
| Impact | Music listeners discover new songs, Music platforms enhance recommendations. |
| Value Proposition | Personalized music discovery, Versatile applications for music industry and listeners. |

<https://www.kaggle.com/datasets/thaisneubauer/million-song-dataset-studies>

<https://www.kaggle.com/datasets/undefinenull/million-song-dataset-spotify-lastfm>

**Table 5. Individual using the internet world wide**

| Description | The global usage of the internet is a phenomenon that has revolutionized the way people communicate, access information, conduct business. Series ID, Series Code, Series Name, Series Parent, Series Unit, Entity ID, Entity Iso, Entity Name, Data Value Data Year, Data Source, Description. |
| --- | --- |
| Potential Applications | Communications, Information Access, E-commerce, Entertainment, Education |
| Related Industry | Technology, Telecommunication, E-commerce, Media & Entertainment |
| Impact | Global Connectivity, Information Acess, Economic Growth, Education and Skill Development |
| Value Proposition | The value proposition of global internet usage lies in its ability to empower individuals, drive economic growth, foster innovation, and enhance quality of life. It offers opportunities for education, communication, entrepreneurship, and access to a vast array of services and information. |

<https://datahub.itu.int/data/?i=11624>

**Table 6. Access To Electricity World Wide (1990-2020)**

| Description | Dataset: Access To Electricity World Wide with Country Name, Country Code, Indicator Name, Indicator Code, Year |
| --- | --- |
| Potential Applications | Household Lighting and appliance, Industrial and Commercial Aciivites, Human Deveropment Index, Education. |
| Related Industry | Common use in energy sector, power generation, distribution and renewawable energy sources. Technology and infrastructure, energy transmission. |
| Impact | Common in Economic growth, Education and Healthcare, Poverty Alleviation, Quality of Life |
| Value Proposition | The value proposition of addressing global access to electricity lies in its potential to drive economic growth, improve social well-being, and advance sustainable development. |

<https://ourworldindata.org/energy-access>

**Table 7. Employment ratio between male and female**

| Description | The employment ratio between males and females worldwide represents the proportion of men and women in the labor force, providing insights into gender disparities in the workplace. Entity, Code, Year, Employment to ratio men, Employment to ration Women, Population, Continent. |
| --- | --- |
| Potential Applications | Gender Equality Assessment, Policy Development, Wokforce Planing, Economy Analysis. |
| Related Industry | Labor and employment, Human Resource, Education and Trainign |
| Impact | Gender Equality, Economic Growth, Innovation and Cretitivty |
| Value Proposition | The value proposition of addressing gender disparities in employment lies in achieving greater gender equality, economic prosperity, and social well-being |

<https://ourworldindata.org/grapher/ratio-of-female-to-male-labor-force-participation-rates-ilo-wdi>

**Table 8. Urban Population Gowth (1960-2022)**

| Description | Urban population growth worldwide refers to the phenomenon of an increasing number of people residing in urban areas within countries across the globe. Country Name, Country Code, Indicator Name, Indicator Code, Year |
| --- | --- |
| Potential Applications | Urban Planning, Real Estate Development, Transportation, Economic Developement. |
| Related Industry | Urban Planning and Develpment, Transportation and Mobility, Retail and Commerece |
| Impact | Economic Oppoturnity, Infrastructure Challenges, Environmental Consideration. |
| Value Proposition | The value proposition of understanding and managing urban population growth lies in achieving sustainable urban development and improving the quality of life for urban residents. |

<https://data.worldbank.org/indicator/SP.URB.GROW>

**Table 9. US road accidents data**

| Description | countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. |
| --- | --- |
| Potential Applications | Predicting accident severity and likelihood by some environmental and location features (classification problem). |
| Related Industry | Transportation, Insurance, infrastructure |
| Impact | Reduced accidents and traffic congestion, Improved safety measures, Enhanced insurance risk assessment, Efficient urban development |
| Value Proposition | By the final model, the likelihood of accident and its severity can be predicted. |

<https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

**Table 10. Cancer data**

| Description | variables about death rate and some social and medical situation of patients for 2010 – 2016 period. |
| --- | --- |
| Potential Applications | classification problem of cancer patient based on some economic and demographic features |
| Related Industry | Healthcare, Research, sociology |
| Impact | Enhanced understanding of cancer socio-economic causes, Informed healthcare policies, Improved safety measures, Enhanced insurance risk assessment, Efficient urban development |
| Value Proposition | improve cancer care |

<https://data.world/nrippner/ols-regression-challenge>

**Table 11. Road accident data**

| Description | detailed information on road accidents reported over multiple years. The dataset encompasses various attributes related to accident status, vehicle and casualty references, demographics, and severity of casualties. |
| --- | --- |
| Potential Applications | Accident causality analysis and prevention strategies, Road safety prediction and improvement recommendations, Targeted public safety interventions, Infrastructure planning and maintenance optimization |
| Related Industry | Transportation, Public Safety, Urban Planning |
| Impact | Safer roads and reduced accidents, Informed infrastructure planning, Effective public safety policies, Insights into accident causest |
| Value Proposition | Enhancing road safety |

<https://www.kaggle.com/datasets/juhibhojani/road-accidents-data-2022>

**Table 12. e-commerce behavior data**

| Description | behavior data for 7 months (from October 2019 to April 2020) from a large multi-category online store.  Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users. |
| --- | --- |
| Potential Applications | Customer segmentation and personalized product recommendations, Fraud detection and prevention, Customer churn prediction and retention strategies, Sales forecasting and inventory optimization |
| Related Industry | Ecommerce, Marketing |
| Impact | Improved customer experience and engagement, Reduced fraud and increased security, Higher customer retention and revenue,- Efficient inventory management |
| Value Proposition | improving market knowledge |

# <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>

**Table 13. Twitter (X) trending words**

| Description | Both trending topics, and people that use them on Twitter during 4 years. |
| --- | --- |
| Potential Applications | Guess what people are interesting and its transition |
| Related Industry | Social media, marketing, data analyses |
| Impact | It is helpful for marketing and is able to know the world's needs and interests. |
| Value Proposition | From the perspective of marketing, it can provide what people are interested in. |

<https://www.kaggle.com/datasets/hwassner/trending-topics/data>

**Table14. Video games sales dataset**

| Description | A dataset of video game sales from 2013 to 2020 is included. it includes information on platform, region, genre, and sales amount. |
| --- | --- |
| Potential Applications | Can be used for analysis of the video game industry, marketing research, and investment decision-making. |
| Related Industry | Video game industry, marketing, and investment industries. |
| Impact | Can identify new trends and opportunities for video game industries. |
| Value Proposition | Can provide customer needs and trendings to video game industries. |

<https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset>

**Table 15. Data science job salaries**

| Description | A dataset of data science job salaries from 2020 to 2022, includes information on job title, experience, skills, and location. |
| --- | --- |
| Potential Applications | Can be used for analysis of data science job salary levels, career planning, and salary negotiation. |
| Related Industry | data science, human resources, and salary management industries. |
| Impact | Helpful for our future decision and planning. |
| Value Proposition | Provides valuable information about data science job salaries and can be used by data scientists, job seekers, and businesses to understand data science job salaries and make better decisions. |

<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>

**Table 16. Speed dating experiment**

| Description | A dataset of speed dating participants from a speed dating event. The dataset includes information on participants' age, gender, occupation, income, and hobbies, information on participants' ratings of each date partner's attractiveness and desire to meet again. |
| --- | --- |
| Potential Applications | Can be used for analysis of the factors that contribute to the success of speed dating, development of matching algorithms, and improvement of dating apps. |
| Related Industry | Dating, matching, and dating app industries. |
| Impact | Increase people’s happiness. |
| Value Proposition | Can be used by researchers, businesses, and individuals to make gf or bf efficiently. |

<https://www.kaggle.com/datasets/annavictoria/speed-dating-experiment>

# **Table 17. Medical Cost Personal Datasets**

| Description | Medical insurance dataset containing details of users for a medical Insurance company: age, sex, bmi, index of body weight, children / Number of dependents, smoker, region, health insurance charges. Medical Datasets are in the public domain but simply needed some cleaning up and recording to match the format in the book. |
| --- | --- |
| Potential Applications | It can be used to explore and analyze medical insurance matters or health condition trends, patterns, and relationships with other related areas |
| Related Industry | Primary health care system, medical insurance, personal health management |
| Impact | This data is useful for these people and is useful to make predictions of the insurance cost they will have to pay |
| Value Proposition | People are always confused about their medical insurance and don't know the cost of insurance at different ages and conditions. |

<https://www.kaggle.com/datasets/rajgupta2019/medical-insurance-dataset>

# **Table 18. Asian Development Bank - Procurement data**

| Description | This dataset is Procurement Plan by Origin of Goods and Services financed from ADB funds |
| --- | --- |
| Potential Applications | It can be used to explore and analyze procurement plan in term of foreign funding and trends, patterns, and relationships with economic development investment spheres |
| Related Industry | Government sector, Line ministries of countries and independent assessment entities |
| Impact | Procurement by origin presents the details of procurement by origin of goods and services. Origin of goods is the place where the goods have been mined, grown, cultivated produced, manufactured, or processed (or through manufacture, processing, or assembly, another commercially recognized article results that differs substantially in its basic characteristics from its imported components). |
| Value Proposition | Procurement by origin presents the details of procurement by origin of goods and services. Overall context of works, consulting and other services, nationality, or places of incorporation and service providers. |

# <https://www.kaggle.com/datasets/zsinghrahulk/asian-development-bank-procurement-data>

# 

# **Table 19. Singapore waste management**

| Description | The dataset contains data from waste data management causing pollution of the environment, it is divided between two timelines, from 2003 to 2017 and 2018 to 2020. It also has additional energy generation data per material type. |
| --- | --- |
| Potential Applications | The National Environment Agency (NEA), research and analysis problems in waste and recycling management |
| Related Industry | Government sector, Municipalities and public entities , NGO sector as well |
| Impact | Discover how much Singapore is saving energy per years by recycling. Singapore has limited land for building new incineration plants or landfills. How to motivate citizens by sharing the total energy that the combined recycling efforts have saved every year |
| Value Proposition | a new milestone of becoming a zero-waste nation and worry about the rising number of waste disposal. At the current pace the Semakau Landfill will run out of space by year 2035 which is an alarming situation for Singaporeans |

https://www.kaggle.com/datasets/kingabzpro/singapore-waste-management

# **Table 20. Water quality data**

| Description | Dataset: The dataset contains 200 rows, each representing a unique water quality measurement across all six parameters. |
| --- | --- |
| Potential Applications | The dataset suits various data science applications such as data visualization, machine learning, and statistical analysis. It can be used to explore and analyze water quality trends, patterns, and relationships |
| Related Industry | Health and Sanitation, Environmental Sustainability, Water management |
| Impact | Water quality is a crucial aspect of environmental management, and it is essential to measure various physical, chemical, and biological parameters to monitor it effectively. |
| Value Proposition | provides a representative snapshot of water quality and can be used for various research, education, and decision-making purposes |

# <https://www.kaggle.com/datasets/shreyanshverma27/water-quality-testing>

# **Table 21 . Global Wealth Inequality Database**

| Description | This is the amalgamation various data sources, including national accounts, survey data, fiscal records, and wealth rankings. Through this comprehensive approach, the Global Wealth Inequality Database enables precise tracking of the evolution of income and wealth levels, ranging from the lowest to the highest strata of society. What distinguishes this initiative is its systematic utilization of such data, permitting comparisons across countries and over extended timeframes. This innovative undertaking builds upon the research on long-term inequality trends crafted by an international consortium of scholars over the past 15 years. |
| --- | --- |
| Potential applications | Wealth Inequality Prediction  Poverty Identification  Fair Wage Analysis  Data-Driven Advocacy |
| Related Industry | * Nonprofit and Philanthropic * Organizations Government and Public Policy * International Organizations (INGOs) * Microfinance and Social Enterprise |
| Impact | * Identifying Root Causes * Policy Design * Monitoring Global Progress * Philanthropy and Aid Allocation |
| Value proposition | Potential to address one of the most pressing and pervasive societal challenges of the current era.  Offers a powerful toolset for tackling one of the most persistent and challenging global issuesCan drive informed decision-making, optimize resource allocation, expand financial inclusion, ultimately contributing to a more equitable and prosperous world. |

<https://www.kaggle.com/datasets/lasaljaywardena/global-wealth-inequality-database>

# **Table 22. Gun Violence Data**

| Description | The dataset contains a record of more than 260k gun violence incidents, with detailed information about each incident, available in CSV format. The data was downloaded from [gunviolencearchive.org](http://www.gunviolencearchive.org/).  The dataset contains 29 columns |
| --- | --- |
| Potential applications: | Study gun violence patterns and statistics and make informed predictions about future trends.  Using gun violence data to make informed predictions about future trends can have significant impacts on public policy, law enforcement strategies, and community safety efforts.   * Gun violence data can inform the development and modification of gun control policies and legislation. * Law enforcement agencies can use predictive models based on gun violence data to allocate resources more effectively. * Community organizations and public health agencies can use predictive models to identify individuals or groups at higher risk of involvement in gun violence. |
| Related Industry | No robust application have been developed in this field apart from some research and manual alert applications |
| Impact | * Law Enforcement Agencies * Government and Policymakers * Community Organizations * General Public |
| Value proposition | The value proposition of using gun violence data to make informed predictions about future trends lies in its potential to save lives, improve public safety, and guide more effective policies and interventions. |

<https://www.kaggle.com/datasets/jameslko/gun-violence-data>

# 

# **Table 23. Loan Default Prediction Dataset**

| Description | This dataset has been taken from Coursera's [Loan Default Prediction Challenge](https://www.coursera.org/projects/data-science-coding-challenge-loan-default-prediction) and will provide you the opportunity to tackle one of the most industry-relevant machine learning problems with a unique dataset that will put your modelling skills to the test. The dataset contains 255,347 rows and 18 columns in total. |
| --- | --- |
| Potential applications | Predictive modeling using a Loan Default Prediction Dataset can have several significant impacts on financial institutions, borrowers, and the overall lending industry.   * Loan default prediction models are crucial for financial institutions to assess and manage credit risk. * Accurate predictive modeling can lead to a reduction in the number of loans granted to individuals or businesses likely to default, resulting in fewer financial losses for the institution. * Minimizing loan defaults can help financial institutions stabilize their financial health. |
| Related Industry | * Banking * Credit Unions * Online Lending Platforms * Finance Companies |
| Impact | * Lenders and Financial Institutions * Borrowers * Investors * Financial Analysts and Researchers |
| Value proposition: | Provide valuable insights and benefits to various stakeholders involved in the lending industry   * Risk Mitigation * Enhanced Decision-Making * Customer-Centric Approaches * Reduced Loan Application Processing Time |

<https://www.kaggle.com/datasets/nikhil1e9/loan-default>

# 

# 

# 

# 

# **Table 24. Asia Pacific: Storm tracks 1956 to 2018**

| Description | This dataset provides a comprehensive record of tropical storm paths in the Asia Pacific region from 1956 to 2018. It includes detailed attributes for each storm such as the storm's name, advisory date and time, wind speed, pressure, type, and GPS coordinates for each advisory point.  The dataset has many source files including shape files and dbf files with over 227,000 entries and 12 columns. |
| --- | --- |
| Potential applications | * **Climate Research**: Analyze trends in storm frequency, intensity, and location to understand changes in climate over time. * **Risk Assessment**: Identify areas at high risk of storms for disaster planning and mitigation. |
| Related Industry | * Weather Forecasting Services * Navigation and Marine Industry * Environmental Research |
| Impact | * Public Safety * Disaster Preparedness and Response * Economic Impact * Tourism and Hospitality |
| Value proposition | Provides a wide range of benefits to governments, businesses, communities, and individuals.   * Reduced Property Damage * Lives Saved * Preparedness |

<https://www.kaggle.com/datasets/lasaljaywardena/asia-pacific-storm-dataset>

**Table 25. Summary**

| Selected data set | Speed dating experiment (Table 16) |
| --- | --- |
| ***This dataset looks most promising and feasible because:***  **- Rich Information:** The dataset contains a diverse range of information about participants, including age, gender, occupation, income, hobbies, and ratings of date partners' attractiveness and desire to meet again. This rich information allows for multifaceted analysis and insights.  **- Human Interaction Study:** It provides valuable insights into human behavior and preferences in the context of dating and social interactions, making it promising for research in psychology, sociology, and human behavior analysis.  **- Machine Learning Potential:** With machine learning techniques, this dataset can be used to develop predictive models for various outcomes, such as predicting the likelihood of a successful date or compatibility between individuals.  **- Recommendation Systems:** It can be used to build innovative recommendation systems for dating platforms, suggesting potential matches based on compatibility metrics derived from the dataset.  **- Social Insights:** Analyzing the dataset can reveal trends and patterns in dating preferences across different demographics, leading to a better understanding of societal norms and preferences.  **- Business Applications:** Dating apps and platforms can leverage this data to improve user experiences, enhance matchmaking algorithms, and create more effective marketing strategies.  **- Ethical Considerations:** However, it's important to note that this dataset should be handled with sensitivity to privacy and ethical concerns, ensuring that participant identities and sensitive information are protected.  ***Which datasets can be potentially combine to create a very unique application?***  The domain of this dataset is very special so may not combine with other datasets. But some found datasets such as road accidents, grocery and supermarket data, and world development indicators, internet usage by individuals and sex ratio could be combined to develop a better applications. | |
|
|
|