Data Mining (2CSDE71)

Name :- Harsh Pansuriya

Roll :- 21BCE176

Date :- 5th January, 2024

Prac. No :- 1

Aim :- Data Domain selection and Identification of Characteristics of selected Dataset of different Formats. Also write a report with following detail

a) Selection of data domain
b) Define the data domain
c) The data source
d) Objective
e) Define the selection of fields
f) Characteristic and behaviours (distribution and inference) of data for each selected field

---

- **Selection of Data Domain :-**
  - HR departments gather tons of info on their employees, like records, reviews, and even surveys on how happy they are. This info is like a hidden treasure chest, but data mining acts as the key to unlock it. By digging through this data, we can discover hidden patterns and understand what's really going on. Imagine knowing who might make a great leader, spotting any missing skills before they become a problem, or even creating perfect benefits packages to make everyone happy. By looking closely at all this data, we can build a stronger, happier, and more productive team.

- **Definition :-**
  - The HR data domain encompasses all data related to an organisation’s workforces, including employees, their employment history, compensation, benefits, performance, skills, training, and other relevant information.

- **Data Source :-**

Table 1-1

Sr. No.	Dataset Name	URL	Information
1	Human Resources Data Set	https://www.kaggle.com/datasets/rhuebner/human-resources-data-set	The CSV revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score.
2	Absenteeism Dataset	https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset	Absenteeism- is a major expense to most organizations. Getting a handle on it, predicting it and affecting it is important for organizations. This dataset provided for HR data scientists to practice on
3	Employee Absenteeism	https://www.kaggle.com/datasets/tonypriyanka2913/employee-absenteeism	XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism.
4	IBM HR Analytics Employee Attrition & Performance	https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset	The IBM dataset, crafted by data scientists, delves into factors influencing employee attrition. It features key attributes like education level and job satisfaction, using scales from 'Below College' to 'Doctor' and 'Low' to 'Very High.' This structured dataset enables detailed analysis, facilitating insights into workforce dynamics.
5	Turnover data set by Edward Babushkin	https://www.aihr.com/wp-content/uploads/2019/10/turnover-data-set.csv	The data set contains information on gender, age, wage type, way of travel, traffic (source of hire), and big five personality!
6	Job classification	https://www.aihr.com/wp-content/uploads/2019/10/job-classification-dataset-1.zip	Job classifications reflect both job families and pay grade related information. This is especially relevant when new jobs are created which need to fit in the existing job structure. Jobs have a number of distinct features which impact the job's classification. These include education level, experience, organisational impact, level of supervision, financial budget, and more.
7	Engagement survey	https://www.aihr.com/online-courses/statistics-in-hr/ Note :- This dataset is not available until you purchase the course	In this Statistics in HR course they use an engagement data set with 85 individuals who all filled in an engagement survey. The data set contains variables like performance rating, function group, but also innovation behavior, multi-dimensional engagement scores, personal initiative, career management behavior, mobility behavior (i.e. the likelihood of leaving the company), organizational and professional commitment, and more.

Table 1-2

Pros	Cons	Knowledge
The pros of the provided dataset include its diversity, allowing for the prediction of suboptimal performance in production staff using various dependent variables. Additionally, the recruitment_cost.csv sheet offers insights into sourcing channel effectiveness and termination predictions.	The abundance of information poses a challenge, requiring a focused research question to avoid getting lost in the extensive data.	Dr. Carla Patalano provided the baseline idea for creating this synthetic data set, which has been used now by over 200 Human Resource Management students at the college. Students in the course learn data visualization techniques with Tableau Desktop and use this data set to complete a series of assignments.
The dataset is advantageous for identifying absence patterns, enabling targeted interventions. Dependent on "AbsentHour," it provides opportunities to explore associations with age and length of service.	The simplicity of the dataset may limit the depth of analysis, and data cleaning, including removing individuals under 18 or above 65, is necessary.And it's totally fake data.	This data set is suitable for identifying pockets of absence in the organization. These pockets may require interventions. 'AbsentHour' will be used as a dependent variable. In addition, age and length of service may also be associated with absence
The dataset is valuable for predicting absence, allowing analyses of BMI, season, workload, and other factors.	The challenge lies in structuring data with multiple records per employee before analysis. Longitudinal research opportunities exist.	The company has shared it dataset and requested to have an answer on the following areas: 1. What changes company should bring to reduce the number of absenteeism? 2. How much losses every month can we project in 2011 if same trend of absenteeism continues?
The dataset provides rich analysis possibilities, including decision trees or logistic regression for predictor identification.	Caution is advised in using logistic regression for attrition prediction, as discussed by Pasha Robert. Simpler tests like ANOVA or Chi-squared can reveal group differences.	Uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'. This is a fictional data set created by IBM data scientists.
The dataset offers an exciting opportunity for real-world analysis, particularly in predicting employee tenure using survival analysis.	Challenges may arise due to potential translation discrepancies in terms such as 'disagreeable', 'self-control', 'anxiety', and 'openness.'	Predict the survival rate of company.
The dataset from Sundmark offers opportunities for job classification using Linear Discriminant Analysis (LDA), aiding in categorizing new jobs within the existing structure. With 66 specifications covering 11 paygrades, it encompasses a comprehensive set of factors.	Potential challenges may arise in the complexity of interpreting and applying LDA results effectively.	Knowing these factors for different jobs enables a job analyst to classify jobs into groups – which are connected to pay scales and benefit packages.
The dataset provides a straightforward challenge for students, guiding them through t-tests, ANOVA, and multiple linear regression analyses in SPSS and R.	The extensive 30-minute lessons for each answer may pose a time-intensive learning curve.	The engagement dataset, featuring variables like performance rating, function group, innovation behavior, and mobility likelihood, offers opportunities for predictive analytics. Knowledge mining may focus on forecasting employee attrition based on mobility behavior, exploring correlations between performance and engagement factors, and identifying patterns in innovation and personal initiative to inform strategic HR decisions.

- **Objectives :-**

- To analyse HR data to gain insights into employee trends, patterns, and behaviours.
- To inform strategic decision-making in areas such as recruitment, retention, talent development, performance management, and employee engagement.
- To identify areas for improvement and optimisation within HR processes.
- To measure the effectiveness of HR initiatives and programs.

• Define the selection of field :-

In the HR dataset from the fictitious courier company XYZ, crucial fields include names, DOBs, age, gender, marital status, date of hire, termination reasons, department, active/terminated status, position title, pay rate, manager name, and performance score. The IBM dataset emphasises education levels and job satisfaction, while the engagement dataset features variables like performance rating, function group, innovation behaviour, and mobility likelihood. Job classification data incorporates education level, experience, organisational impact, supervision level, and financial budget.

• Characteristic and behaviours of Data :-

In the above datasets, variables like age, gender, and performance score display varied distributions, impacting absenteeism prediction.

Example :-
Older employees tend do be more absent then younger one.

The IBM dataset's education levels and job satisfaction exhibit diverse patterns, influencing attrition.

Engagement dataset variables, such as performance rating and innovation behaviour, showcase distinctive distributions, guiding insights into workforce dynamics.

Job classification data's characteristics involve education, experience, organisational impact, supervision level, and financial budget, each influencing job classifications uniquely.

Example :- Higher educated and more experienced people are generally do jobs like manager, CEO, etc. And not much experienced people are ground level employees.