CIDS Assignment-1

K. Mahathi
21BSCIA3125
CSIT-A

1) Define Data Science.

Data Science is an area that manages, manipulates, extracts and interprets knowledge from tremendous amount of data. It is a multidisciplinary field of study with goal to address the challenges in Big data.

2) What is datafication?

Datafication is a process of taking all aspects of life and turning them into data. It simply means turning many physical aspects of life into computerized data.

Ex:- i) Google's augmented-reality glasses datafy the gaze.
2) Twitter datafies stray thoughts.
3) LinkedIn datafies professional networks.

3) What is Population?

The total possible outcomes of an experiment is called population. It could be any no. of objects or units such as tweets, photographs or stars.
It is represented by N.

4) Differentiate Feature Generation and Feature Selection.

Feature Selection	Feature Generation (or) Feature Extraction
1) Feature selection selects a subset of relevant features from the original set of features	It extracts a new set of features that are more informative and compact.
2) Reduces the dimensionality of the feature space and simplifies the model	Captures the essential information from the original features and represents it in a lower-dimensional feature space.
3) Can be categorized into filter, wrapper and embedded methods.	Can be categorized into linear and non-linear methods.

4)	Requires domain knowledge and feature engineering.	can be applied to raw data without feature engineering.
5)	It can improve the model's interpretability and reduce overfitting	It can improve model's performance and handle non-linear relationships
6)	May lose some information and introduce bias if the wrong features are selected	May introduce some noise and redundancy if the extracted features are not informative.

5) What is adjusted R^2 coefficient and explain its use?

The adjusted R-squared coefficient is a statistical measure used to assess the goodness of fit of a regression model. It is a modified version of R^2 coefficient that takes into account no. of predictors in the model.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

P = no. of predictor variables

N = No. of records

R = R-squared value of dataset.

Use of Adjusted R^2

When more no. of predictor variables are added, R^2 will generally increase which misleads the impression of model fitting. The adjusted R^2 controls this by penalizing the addition of uninformative attributes.

6) Explain different data science applications?

Applications Of Data Science

1) In Search Engines

The most useful application of Data Science is Search Engines like Google, Yahoo, Safari etc. Data science is used to get search engines faster.

Eg: Top most visited web links are shown first.

2) In Transport-

Data science is also entered in real-time such as transport field like Driverless Cars. With the help of driverless cars it is easy to reduce the no. of accidents.

3) In Finance

Data science plays a key role in financial industries. Financial industries always have an issue of fraud and risk of losses. Thus, financial industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company.

Also, financial industries uses Data science analytics tools in order to predict the future. It allows companies to predict customer lifetime value and their stock market moves.

4) In E-Commerce

E-Commerce websites like amazon, flipkart etc. uses data science to make a better user experience with personalized recommendations.

5) In Health Care

In healthcare, data science is used for:

i) Detecting Tumor

ii) Drug discoveries

iii) Medical Image Analysis

iv) Virtual Medical Bots
v) Genetics and Genomics
vi) Predictive Modeling for Diagnosis etc.

6) Image Recognition

Data Science is also used in Image Recognition. It gives Suggestions tagging who is in the picture. When an image is Recognized, the data analysis is done on One's facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile the facebook suggests us auto-tagging.

7) Targeting Recommendation

It is the most important application of datascience. Whatever the user searches on the Internet, he/she will see numerous posts everywhere.

8) Airline Routing Planning

With the help of Data science, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or to take a halt in between.

9) Data Science in Gaming

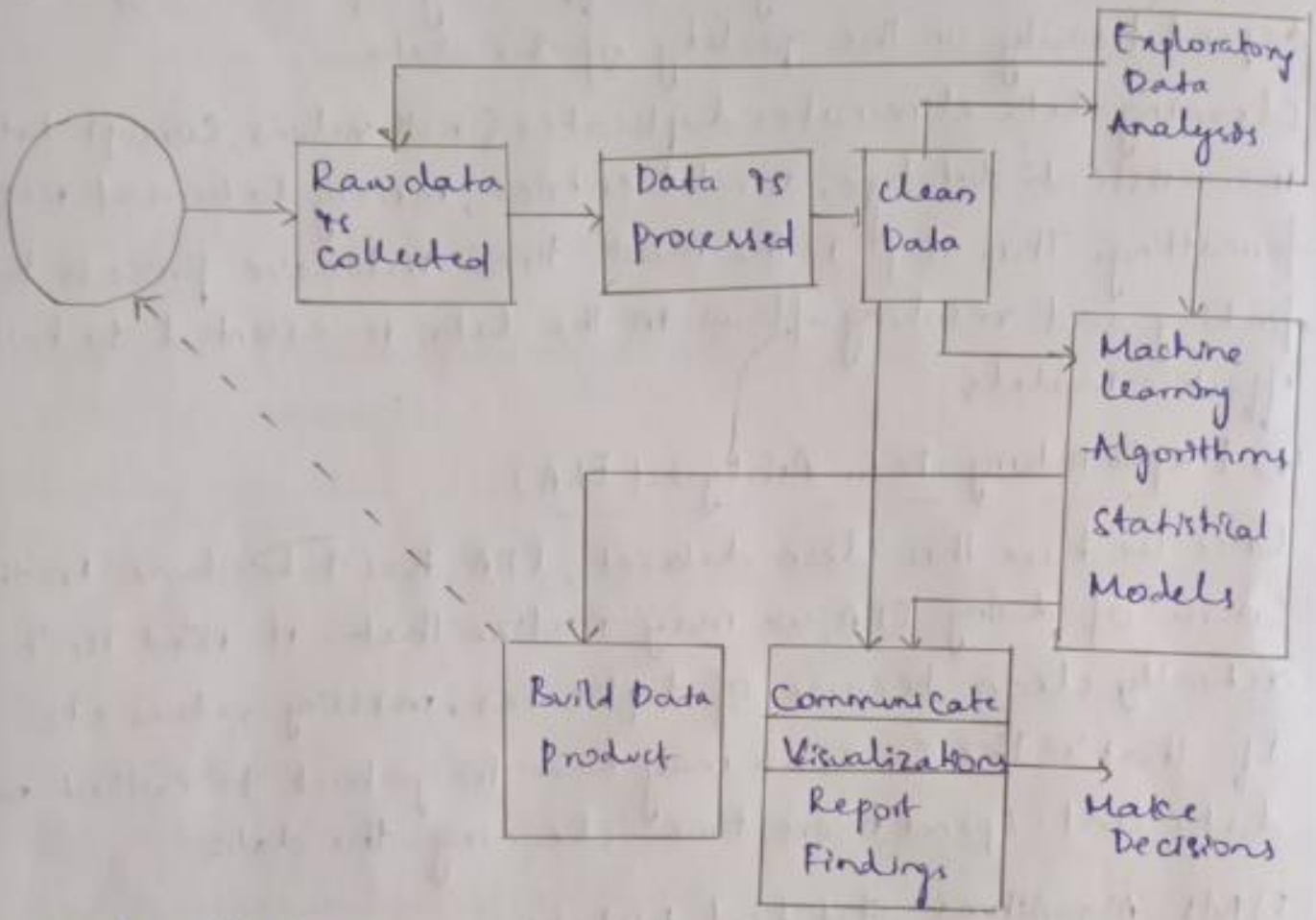
In most of the games where a user will play with an opponent i.e, a computer opponent, data science concepts are used with machine learning where with the help of past data the computer will improve its performance. There are many games like chess, EA sports etc will use data science concepts.

10) In Delivery Logistics

Various Logistics Companies like DHL, Fedex etc make use of Data Science. Data Science helps these companies to find the best route for the shipment of their Products, the best

time suited for delivery, the best mode of transport to reach the destination etc.

7) Explain the process of Data Science with a neat diagram.



1) Data collection

Gather relevant data from various sources, which may include databases, APIs, spreadsheets or web scraping. Ensure data quality, quantity and completeness.

2) Data Processing

Perform a preliminary exploration of the raw data to gain an understanding of its structure, format and potential issues. This can help identifying the scope of preprocessing required. Validate the collected data to ensure it meets the expected format and quality. Document the metadata associated with the collected data. Create backups of the raw data to prevent data loss incase of accidental changes or errors during processing.

3) Clean data

Most of the data we collect during the collection phase will be unstructured, irrelevant and unfiltered. Bad data produces bad results, so the accuracy and efficiency of the analysis will depend heavily on the quality of the data.

Cleaning data eliminates duplicates & null values, corrupt data, inconsistent datatype, invalid entries, missing data and improper formatting. This step is the most time intensive process but finding and resolving flaws in the data is essential to build effective models

4) Exploratory Data Analysis (EDA)

Once we have this clean dataset, EDA has to be done. In the course of doing EDA, we may realize that it wasn't actually clean because of duplicates, missing values etc. If that's the case, we may have to go back to collect more data and spend more time cleaning the data.

5) ML Algorithms Statistical Models

We design our model to use some algorithm like k-nearest neighbour (k-NN), Linear regression, Naive Bayes etc.

The model we choose depends on the type of problem we are trying to solve. It could be a classification problem, a prediction problem or a basic description problem.

6) Visualization

We then can interpret, visualize, report or communicate our results. This could take the form of reporting the results upto our boss or coworkers or publishing a paper in a journal & going out and giving academic talks about it

~~b) Take a different approach~~

1) Build data product

In a different approach, our objective might be to develop a test "data product". This could be something like a spam filter, a method for ranking searching results or a system that gives recommendations. The unique aspect of data science, as opposed to stats, in that this product is put to use in the real world where people interact with it. This interaction generates more data which in turn forms a loop of feedback.

8) What is feature generation and feature selection? Explain with an example.

Feature generation

Making a list of things what we want for our project to do is called feature generation or extraction.

In other words, feature generation is the process of constructing new features from the existing ones.

The goal of feature generation is to derive new combinations and representations of our data that might be useful to the machine learning model.

Ex:- Chasing Dragons App

Consider the above app where users pay a monthly subscription fee to use it. The more users you have, the more money you make. Suppose you realize that only 10% of new users ever come back after 1st month. So there are 2 options to increase revenue:

1) find a way to increase retention rate of existing users
2) Acquire new users.

Generally it costs less to keep an existing user around than to market and advertise to new users. But setting aside that particular cost-benefit analysis of retention, focus on user

retention situation by building a model that predicts whether or not a new user will come back next month based on their behavior this month. We could build such a model in order to understand retention situation, but instead focus on building algorithm that is highly accurate at predicting

Feature Selection

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant or noisy features
Selecting the best features helps the model to perform well.
Ex: Consider a dataset with information about customer behavior, including features like age, income, no. of purchases and browsing history. Feature selection techniques can be used to identify which of these features have the most influence on on predicting whether a customer will make a purchase.
For instance, if age and income are the most important predictors, we might select only those 2 features discarding the others

9) What is feature selection? Explain the role of information gain in feature selection

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant or noisy features.

→ Information gain calculates the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the information gain of each variable in the context of the target variable.

To find which attribute provides information about a particular event X, we use information gain denoted by $IG(X,a)$ where a is the attribute which we want to find the information.

$$IG(X,a) = H(X) - H(X|a)$$

$H(X)$: entropy for event X

$H(X|a)$: conditional entropy for event X wrt attribute a

High entropy means the data is more disordered and unpredictable, while low entropy indicates a more organized and predictable dataset.

Information Gain assesses how much splitting the dataset by a particular feature reduces its entropy.

11) a) What is Logistic regression? Explain in detail.

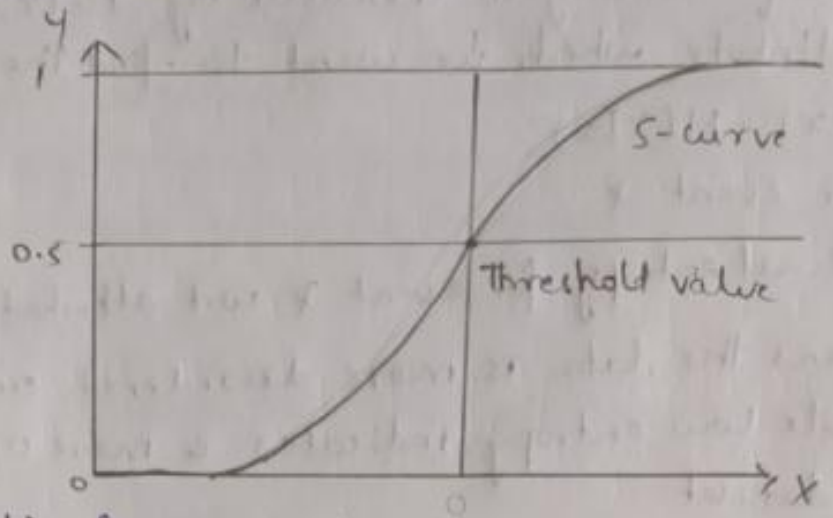
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

→ The outcome must be a categorical value. It can be either yes or no, 0 or 1, true or false etc. but instead of giving the exact value, it gives the probabilistic values which lie between 0 and 1.

→ Logistic Regression is much similar to Linear regression except that Linear regression is used for solving regression problems whereas Logistic regression is used for solving the classification problems.

→ In Logistic regression, instead of fitting a regression line, we fit an 'S' shaped logistic function which predicts 2 maximum values (0 or 1)

Logistic Function



Logistic Regression Equation

It can be obtained from Linear regression Equation.

Equation of straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

In Logistic regression y can be between 0 and 1, so divide the above equation by (1-y)

$\frac{y}{1-y}$; 0 for y=0 and ∞ for y=1

We need range in b/w $(-\infty, \infty)$, take Logarithm of the equation

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

b) What is Multiple Linear Regression? Explain in detail.

It is an extension of straight line regression which involves more than 1 predictor variable. An example of multiple linear regression with 2 predictor variables A_1 and A_2 and 1 response variable is as given below.

$$y = w_0 + w_1x_1 + w_2x_2$$

x_1 and x_2 - values of attributes A_1 and A_2 respectively in X

To solve w_0, w_1 and w_2 use the method of least square whose values are given below

$$w_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$w_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \quad a = w_0 \quad b_1 = w_1 \quad b_2 = w_2$

10) Find the regression coefficients for the following data.

Hours worked	Pre Exam Marks	Score
1	15	80
2	28	95
1	14	76
1	13	70
2	22	81
2	28	97
3	30	98
3	29	95
1	14	74
2	25	85

Consider mean hours worked $\bar{x}$ and mean score $\bar{y}$

$$\bar{x} = \frac{1+2+1+1+2+2+3+3+1+2}{10}$$

$$\bar{x} = 1.8$$

$$\bar{y} = \frac{80+95+76+70+81+97+98+95+74+85}{10}$$

$$\bar{y} = 861.1$$

The regression coefficients can be estimated by using the following equations: $w_0 = \bar{y} - w_1\bar{x}$ —①

$$w_1 = \frac{\sum_{i=1}^{101} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{101} (x_i - \bar{x})^2}$$

For Hours worked

$$w_1 = \frac{(1-1.8)(80-861.1) + (2-1.8)(95-861.1) + (1-1.8)(76-861.1) + \text{---} + (2-1.8)(85-861.1)}{(1-1.8)^2 + (2-1.8)^2 + (1-1.8)^2 + \text{---} + (2-1.8)^2}$$

$w_1 \approx 6.74$

$w_0 = \bar{y} - w_1 \bar{x} = 861.1 - 6.74 \times 1.8 = 850.9$

For pre-exam marks

$\bar{x} = \frac{15+28+14+13+22+28+30+29+14+25}{10}$

$\bar{x} = 21.8$

$\bar{y} = \frac{80+95+76+70+81+91+98+95+74+85}{10} = 861.1$

$$w_1 = \frac{(15-21.8)(80-861.1) + (28-21.8)(95-861.0 + \text{---} + (25-21.8)(85-861.1)}{(15-21.8)^2 + (28-21.8)^2 + \text{---} + (25-21.8)^2}$$

$w_1 \approx 4.49$

$w_0 = \bar{y} - w_1 \bar{x} = 861.1 - 4.49 \times 21.8 = 768.41$

Therefore, for hours worked

slope $w_1 = 6.74$, intercept $w_0 = 850.9$

for pre exam marks

slope $w_1 = 4.49$, intercept $w_0 = 768.41$