Ashoka Horizons Proram
Assignment #2 HW
Applied Data Science
Student: Advit Agrawal
Instructor: Rintu Kutum, Gautam Ahuja

Part 1: Probability and Statistics

1) Probability is a measure of the likelihood of an event taking place. A probability of 0,means that it is an impossible event. A probability of 0.5, means that there is a 50% chance of the event taking place and a probability of 1.0, means that there is a 100% chance of the event taking place and it is the only possible outcome
2) The total possible outcomes on a standard 6 sided die are:
   6
   5
   4
   3
   2
   1

   The only favourbale outcome is:
   3

   Probability = no. of favourbale outcomes/total possible outcomes
   Therefore the probability = 1/6
3) The three main measures of central tendency are:
   a) Mean
   b) Mode
   c) Median
4) The primary purpose of descriptive statistics is to arrange data in a way that is easy to understand and contextualise. For example, if we use descriptive statistics to analyse the academic performance of a class, we can find out the class average, trends in particular subjects etc. This helps teachers identify which students may need extra support, recognise patterns of improvement or decline, and revise the curriculum and teaching style accordingly.
5) Range is the difference between the maximum and minimum values in a dataset. In the example of the test scores provided in the assignment, the range would be the maximum value - the minimum value, which is 100-60=40.
6) Variance is measured in terms of the square of the original units. For example, if we are measuring the amount of plastic consumed by all the states of India and the unit is metric tonnes, then the variance would be in metric tonnes squared. This makes it harder to interpret in real-world terms. Standard deviation, on the other hand, is the square root of the variance and is easier to interpret because it is expressed in the same units as the original data. In the example above, the standard deviation would be

expressed in metric tonnes, making it more meaningful and practical for understanding the spread of data.

7) Understanding probability is crucial when working with machine learning models, especially during the hypothesis testing phase, as explained in the slides. Probability helps quantify the likelihood of observed outcomes based on data. In hypothesis testing, it is used to determine whether a result is statistically significant or could have occurred by random chance. For example, in a drug trial, if 86% of patients report improvement after taking a drug, this probability (0.86) is used to assess the effectiveness of the drug. The model uses this to decide whether the improvement is likely due to the drug or just a coincidence.
8) Since we are trying to describe the central tendency of a dataset, the median would be a better option because in the case of mean, the average of all the values is being calculated. If there is an excessively large or small outlier in the dataset, that would skew our understanding of the data. For example, if these were the house prices in a city
    a) [45 lakh, 50 lakh, 52 lakh, 48 lakh, 49 lakh, 47 lakh, 10 crore]

If we used the mean, the outlier 10 crore, would result in an average of around 2 crore, which is not showing the full picture. However, the median would be ardoun 50 lakhs, which is much more accurate in representing the central tendency of house prices.

9) Data exploration helps us identify trends and patterns in a given dataset, which can help determine likely outcomes. It is based on these patterns that machine learning models are trained, especially in supervised learning, where data exploration helps reveal which features (inputs) are most related to the target variable (output). This understanding allows the model to learn accurate relationships from the labeled data and make better predictions.
10) In the IBM Watson case at Memorial Sloan Kettering, large datasets such as patient histories, treatment outcomes, clinical trial results, and medical research articles were essential. They provided a wide base of real-world information that allowed Watson to learn how different cancers respond to different treatments in various situations.The methods used were just as important. Statistical analysis helped identify patterns in treatment effectiveness, while machine learning allowed the system to improve its recommendations over time as it processed more data (large amounts of data was pivotal in this case). Natural language processing enabled Watson to read and understand unstructured information from doctors' notes and research papers.
11) The standard deviation might be very large due to outliers in the housing price dataset. If one house is priced very high or low, the mean will change accordingly and increase the standard deviation. If we only looked at the mean, the average house price could be very high, even though most houses have lower prices, therefore skewing our interpretation
12) The volcano plot shows which genes changed the most and how reliable those changes are. The x-axis shows how much a gene increased or decreased (up- or down-regulated), and the y-axis shows how statistically significant that change is. Genes

far from the center and high up changed a lot and are highly significant. Up-regulated means more active; down-regulated means less active.

(I asked chatgpt to explain this to me, while I kind of understood, I'm still unclear)

13) Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.
14) Supervised Learning, unsupervised Learning, Reinforcement Learning
15) In supervised learning, classification involves predicting a category or label, such as identifying handwritten digits using LeNet-1 in 1989. Regression involves predicting a continuous numerical value, such as estimating the price of a house or forecasting tomorrow's temperature.
16) The main goal of unsupervised learning is to discover hidden structures, patterns or relationships in the data. We can group similar data points together which allows us to discover new relationships which we may not have otherwise found.
17) PCA stands for Principal Component Analysis and can combine various factors in a dataset, ranking them from most important to least important, therefore simplifying data.
    a) https://www.youtube.com/watch?v=FD4DeN81ODY
18) There are 3 main factors when it comes to any sort of programming, the rules(program), data and the output. In traditional programming, we give the program and the data and the computer produces the output, but in machine learning, we give the input and the output, btu the computer comes up with its own programs/rules.
19) The core idea of "Learning from Examples" is that we use large amounts of data so that the Machine Learning algorithm can come upw ith its own definitions of what a particular image will be classified into. If we take the cat recognition analogy, we find that its very difficult to define the features of a cat because they can be very very different. This is why we train ML algorithms using pictures of different kinds of cats so that they can identify complex patterns which are very difficult to define. This is why ML algorithms are more accurate than simply defining a particular range or feature set.
20) In reinforcement learning, an agent is the one that takes decisions in an environment and identifies which decisions lead to better awards or outcomes. Through the variation of feedback based on the action taken, it can identify which actions lead to more positive feedback and work accordingly.
21) Two common algorithms for supervised learning are linear regression and logistic regression. One algorithm for unsupervised learning is K-Means, which is used for clustering.
22) Two workflow steps:
    a) Data Preprocessing: This is very important because the data fed to the ML algorithm determines its predicted outcome. If the data has a lot of outliers, erratic readings or missing entries, the results will not be as accurate. I have experienced this firsthand. I was making a glove that convert sign language to speech and used an ML model which took 100 samples of readings from the sensors on the glove to calculate an average value. Because of the inaccuracy of my sensors, the prediction was only 66%. This is a reason why data preprocessing is important.

b) https://www.youtube.com/watch?v=pYVScuY-GPk Feature engineering is important because it allows the ML algorithm toe learn patterns better. When you take the existing data that you have and combine it with your domain knowledge or come up with another data column from the existing data, it is easier for the ML algorithm to identify the pattern. For example, height and weight alone cant help a model predict health risks, but BMI can.

23) In this case, a positive is when the email is marked as spam. If the email is not spam but is marked as such, that would be a false positive. This may be provelmatic if the email contains important information about exam dates, fees submission etc. and would make spam detection useless if it happens to the users because of their fear of missing out on important emails. Therefore, they would now check the spam folder along with their main inbox.
24) AI is a broad field ofcomputer science focused on creating systems that can perform tasks that typically require human intelligence.
25) Deep Learning is a part of ML, ML is a part of AI
26) The types of AI are Artifical Narrow Intelligence, Artificial General Intelligence and Artifical Superintelligence. We only have ANI today.
27) Two key areas which are foundations of AI are Planning and Natural Language Processing
28) Thinking Humanly means designing AI to mimic how humans think while Acting Rationally means designing AI to make the best possible decisions, irrespective of how a human would think.
29) Natural Language Processing is a branch of AI that allows computers to actually interpret human language, for example AI detectors and chatbots
30) Generative AI is the type of AI that can create content like text, images, code etc. Traditional AI models usually only detect or classify patterns, but Generative AI can learn from those patterns and create its own versions. For example: Chatgpt
31) Hypothetically, if an AI detector was trained using 100 texts (all human-written), but 90 of those texts were written by native speakers and only 10 by foreign speakers, the model would learn that the writing patterns of native speakers represent "normal" human writing. As a result, when it encounters writing by a foreign speaker, possibly with simpler grammar, unusual phrasing, or non-standard structure, it might misclassify it as AI-generated, simply because it doesn't match the dominant pattern it saw during training.
32) Explainability and transparency in AI are important, especially in healthcare. If an AI analyses an image and flags it for some sort of disease, it is important for the medical community to know which particular features led it to make that decision (learning wise). Its also important because if the AI makes a decision based on some sort of flawed reading or logic, there could be devastating consequences for the patient.