1. Describe the following concepts:

* Machine Learning:

Ans: Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observation or data, e.g. direct experience, or instruction, in order to look for patterns in data and make better decisions in future based on the examples that we provide. The primary aim is to allow computers learn automatically without human intervention or assistance and adjust actions accordingly.

* Human Learns from past experiences, Computer learns from data (True/False?)

Ans: True

* What is Data? What is Goal? give two example

Ans: **Data:** Data is individual units of information.

e.g. weights, prices, costs, numbers of items sold

**Goal:** The goal of Machine Learning is to devise learning algorithms that do the learning automatically without human intervention or assistance.

e.g.

* face detection: find faces in images (or indicate if a face is present)
* spam filtering: identify email messages as spam or non-spam
* topic spotting: categorize news articles (say) as to whether they are about politics, sports, entertainment, etc.
* What is a Class in Decision Tree learning?

Ans: A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification". Each element of the domain of the classification is called a ***class***. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature.

Each leaf of the tree is labeled with a class or a probability distribution over the classes, signifying that the data set has been classified by the tree into either a specific class, or into a particular probability distribution (which, if the decision tree is well-constructed, is skewed towards certain subsets of classes).

* Naive Bayes is one of common Machine Learning algorithms that is often used for the purpose of text classification (True/False)?

Ans: True

* Information Theory?

Ans: Information Theory is the mathematical treatment of the concepts, parameters and rules governing the transmission of messages through communication systems.

* What is an Heuristic Algorithm?

Ans: A **heuristic algorithm** is one that is designed to solve a problem in a faster and more efficient fashion than traditional methods by sacrificing optimality, accuracy, precision, or completeness for speed. **Heuristic algorithms** often times used to solve NP-complete problems, a class of decision problems.

* All current tree building algorithms are heuristic algorithms (True/False), why?

Ans: True.

All trees follow a top-down greedy approach known as recursive binary splitting. We call it as ‘top-down’ because it begins from the top of tree when all the observations are available in a single region and successively splits the predictor space into two new branches down the tree. It is known as ‘greedy’ because, the algorithm cares (looks for best variable available) about only the current split, and not about future splits which will lead to a better tree.This splitting process is continued until a user defined stopping criteria is reached

* A Decision Tree Learning algorithm is a greedy divide-and-conquer algorithm? Yes/No, Why?

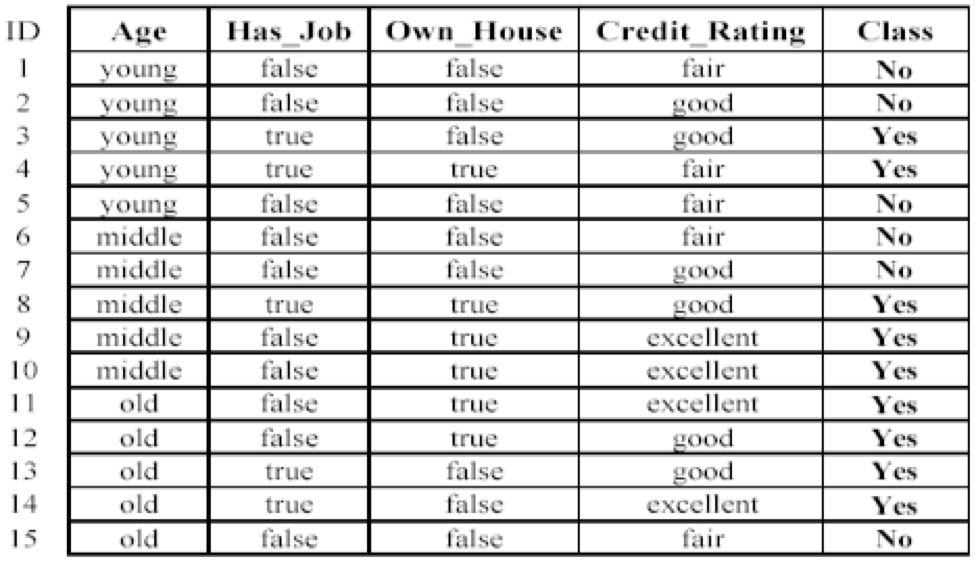
Ans: True

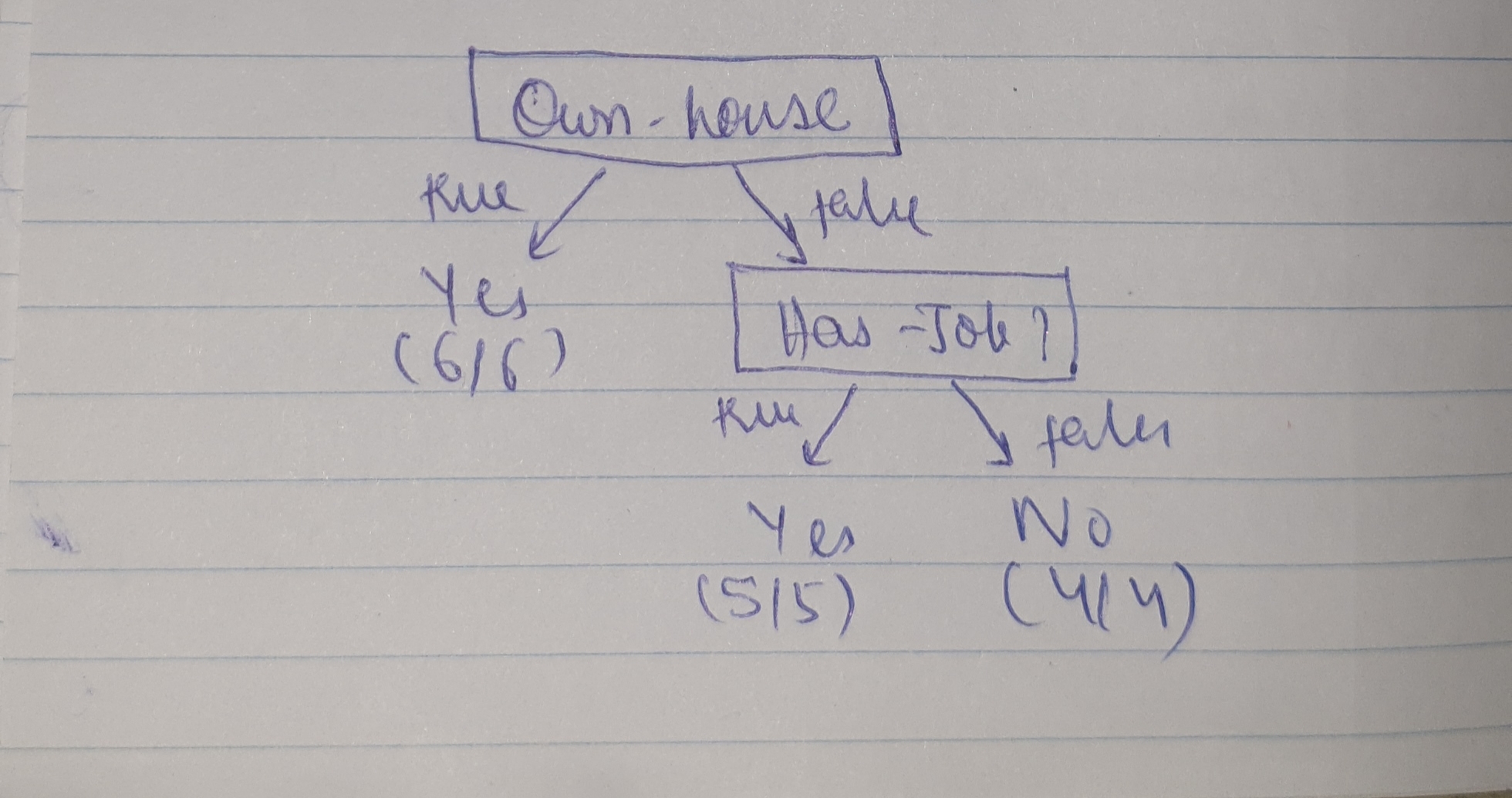
The basic algorithm used in decision trees is known as the ID3 (by Quinlan) algorithm. The ID3 algorithm builds decision trees using a top-down, greedy approach.

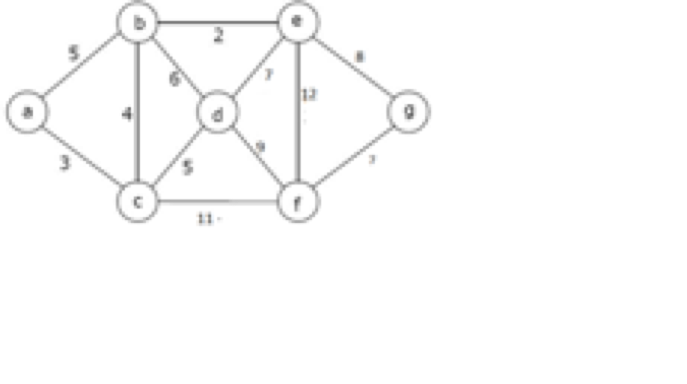
Briefly, the steps to the algorithm are: - Select the best attribute → A - Assign A as the decision attribute (test case) for the **node**. - For each value of A, create a new descendant of the **node**. - Sort the training examples to the appropriate descendant node leaf. - If examples are perfectly classified, then stop else iterate over the new leaf nodes.

Now, the next big question is how to choose the best attribute. For ID3, we think of the best attribute in terms of which attribute has the most information gain, a measure that expresses how well an attribute splits that data into groups based on classification.

2. Consider the following Loan data, build a Decision Tree



Ans:

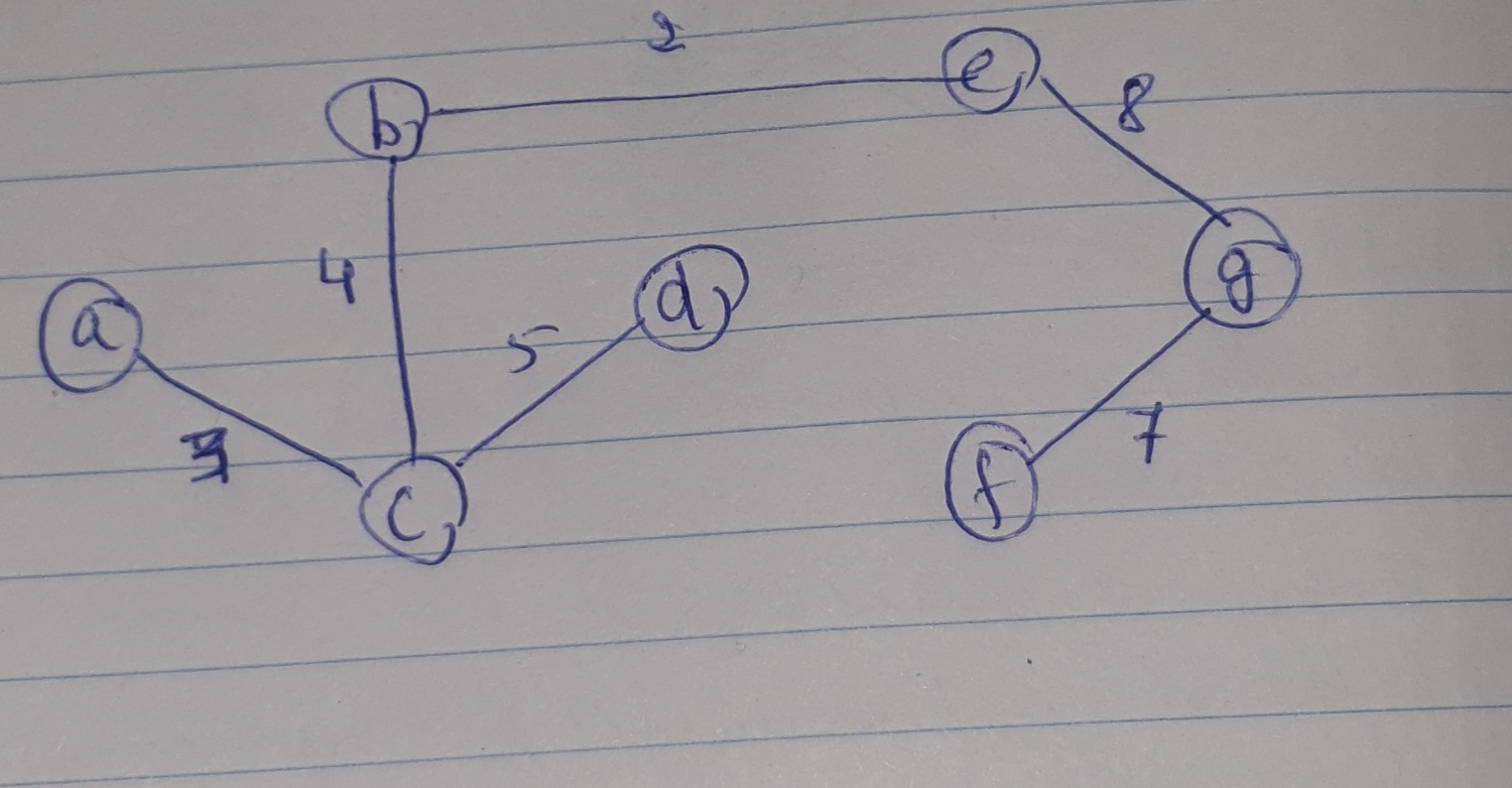
3. Solve the minimum spanning tree for the following graph:

* Krushkal’s Algorithm step by step

Ans: The graph contains 7 vertices and 12 edges. So, the minimum spanning tree formed will be having (7 – 1) = 6 edges.

Now pick all edges one by one from sorted list of edges

* Pick edge b-e, no cycle is formed, include it.
* Pick edge b-c, no cycle is formed, include it.
* Pick edge a-c, no cycle is formed, include it.
* Pick edge d-c, no cycle is formed, include it.
* Pick edge f-g, no cycle is formed, include it.
* Pick edge e-g, no cycle is formed, include it.



* Write Java code, compile and run

Ans: Code attached.

* Compare space and time complexity

Ans: **Time Complexity:** **O( |E| log|V| ),**in worst case we would have to do find or union operation for all the edges. *Find* and *Union* both are O( log|V| ) operation, which can be optimised to O(1) if we use path compression*,* in which case the time complexity of the *Kruskal’s Algorithm* will be O(|E|). If we consider the sorting of the edges then we would have another O(|E|log|E|).

**Space Complexity: O(|E| + |V|),**since Disjoint Set Data Structure takes O(|V|) space to keep track of the roots of all the vertices and another O(|E|) space to store all edges in sorted manner.

5. What is Cell, Gene, Chromosomes, DNA, Human Genome Project?

Ans:

Cell: The cell is the basic unit of all the organisms. It is basic building blocks of life and the smallest unit of life. A Cell consist of cytoplasm enclosed within a membrane, that contains many biomolecules such as proteins and nucleic acids.

Gene: A gene is a sequence in RNA or DNA that encodes the synthesis of a gene product, either protein or RNA. A gene is located on a chromosome and is the functional unit of inheritance controlling the transmission and expression of one or more traits.

Chromosome: A chromosome is a DNA molecule with part or all of the genetic material (genome) of an organism. Each chromosome is made up of DNA tightly coiled multiple times around proteins called histones that support its structure. Each chromosome has a constriction point called the centromere, which divides the chromosome into two arms. The short arm of the chromosome is called the p arm. The long arm of the chromosome is called the q arm.

DNA: Deoxyribonucleic acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms.

Human Genome Project: The Human Genome Project (HGP) was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint.

6. What are type DNA mutations, give example for each.

Ans: A gene mutation is an alteration in the DNA sequence that makes up a gene, such that the sequence differs from what is found in most people permanently.

Type of mutations:

Missense mutation

In this type of mutation, one DNA base pair changes that results in the substitution of one amino acid for another in the protein made by a gene.

Example- Sickle- cell anemia

Nonsense mutation

A nonsense mutation is a change in one DNA base pair. The altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.

Example- AUG GCC TGC AAA  
                AUG GCC TGA AAA

Insertion mutation

An insertion mutation changes the number of DNA bases in a gene by adding a piece of DNA. Hence, the protein made by the gene doesn’t function properly.

Example- CTGGAG  
          CTGGTGGAG

Deletion mutation

A deletion mutation changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighbouring genes. The deleted DNA alters the function of the resulting protein/proteins.

Example- CTGGAG  
                CT~~GG~~AG

Duplication mutation

A duplication mutation consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.

Example- CTAG  
               CTCTAG

Frameshift mutation

A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

Example- ACG AGG ACU GCA UAC  
              A CGA GGA CUG CAU ACC

Repeat expansion mutation

A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function improperly.

Example- CAT TCA CAG GTA  
          CAT TCA CAG CAG CAG GTA

7. In the article that I sent you:  
 a)What is Tumor DNA sequencing?  
 b) Which gene does the article identify as an example?, and the mutations in the identified gene causes what kind of problem?  
c) In this article what is the name of Gene and what is the root cause of cancer and how it is created?

Ans:

* Tumor DNA sequencing is the whole genome sequencing of a single, homogeneous or heterogeneous group of cancer cells. It is a biochemical laboratory method for the characterisation and identification of the DNA or RNA sequences of cancer cells.
* Mutation of EGFR gene causes the cells to divide rapidly and can be found in lung cancer cells.
* An EGFR mutation mentioned in this article refers to a mutation to the portion of the DNA in a lung cancer cell which carries the instructions for making EGFR (epidermal growth factor receptor) proteins. The root cause is DNA mutation in the EGFR gene called T790M. DNA sequencing test can be used to determine if the cause for a lung cancer is EGFR mutation.