**Northeastern University**

CS 5100  Foundations of Artificial Intelligence

                                                Homework and PA 5

                                    NAME: **PARSHVA TIMBADIA**

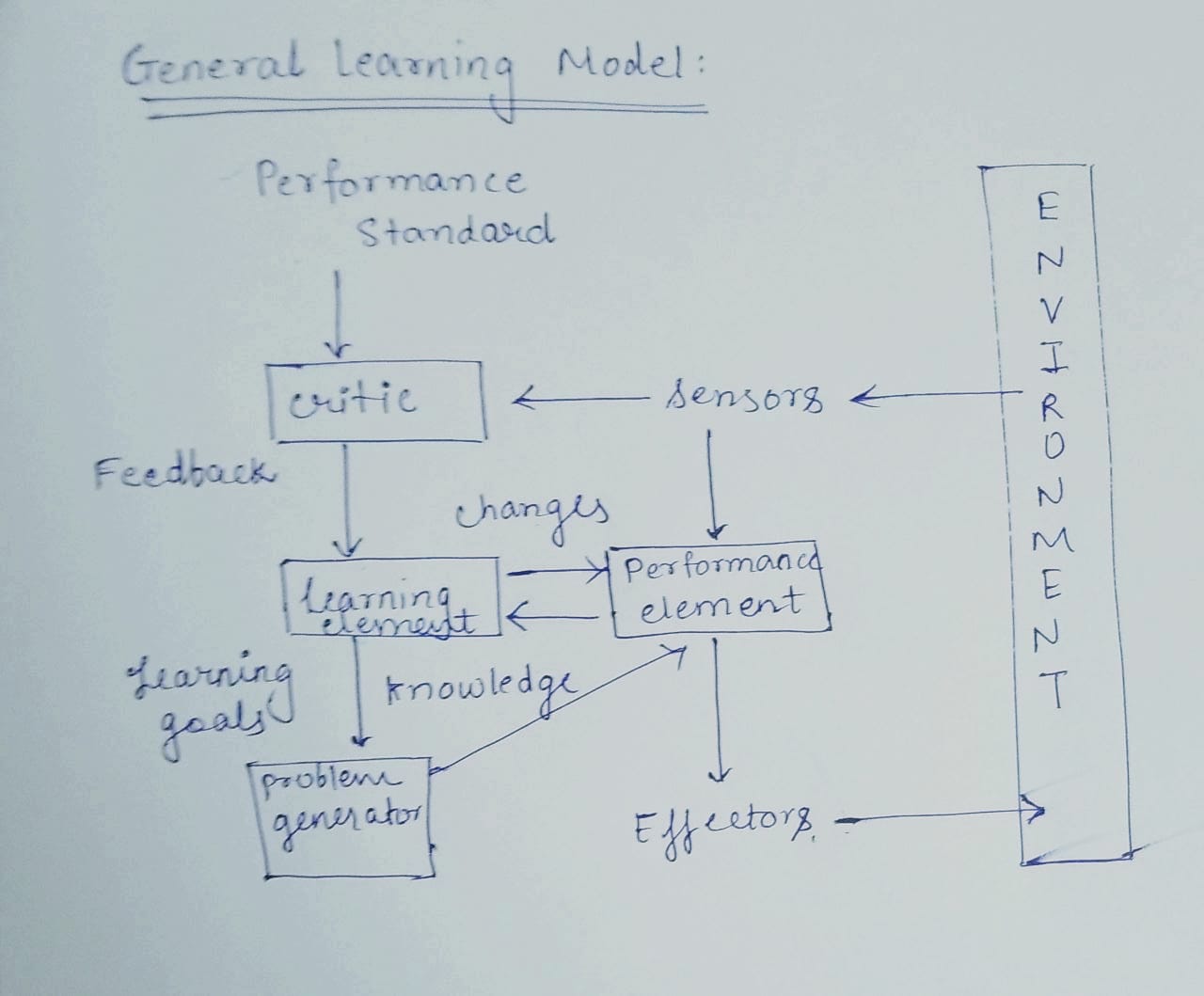
                                    EMAIL: [timbadia.p@northeastern.edu](mailto:timbadia.p@northeastern.edu" \t "https://mail.google.com/mail/u/0/" \l "advanced-search/from=parshwat%40gmail.com&query=parshwat%40gmail.com&isrefinement=true&fromdisplay=parshwat/_blank)

                                    NUID: **001091783**

**Q1>**

**a)**

Below is the diagram of the general learning model and we will now try to fit that into our given problem.



The given problem can be considered as a learning model since speaking and understanding language develops over time for the infant.

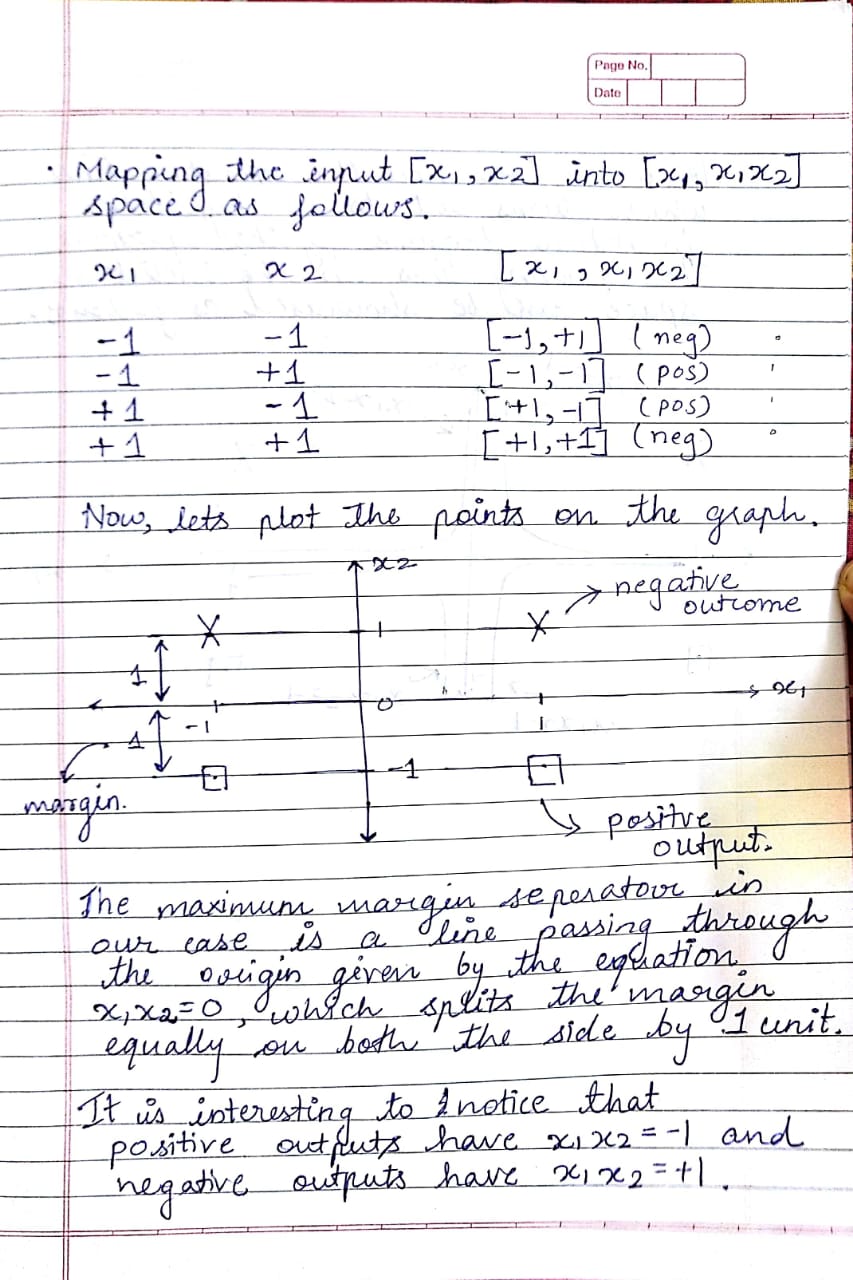
Usually, infants learn from the observationswhich includes watching and listening other person speaking in our case. So here the **environment** can be the **physical objects or the other humans** which can be perceived **(sensor)** with the help of **ears, eyes and hands.** The coordination of hearing and visual works the best for teaching the infant. For instance, while teaching a child to pronounce a particular wors, say apple, the physical object apple can provide much better understanding to the kid and also helps in memorizing the word. The **effectors** for the kid can be **mouth, tongue and lip movements.** The **performance measure** can be done by checking how efficiently the child Is **speaking the language** after few weeks of learning and training. The **critic provides the feedback** so in our case it could be something like **parents reaction.** For instance, when a child utters a word ‘’Mom” the happiness of the parent can be considered as the critic and **the learning element** is nothing but the **positive or negative review.** For instance, child speaking wrong word can create a disappointment reaction from the parents and the child will get the gist that he/she is doing something wrong.

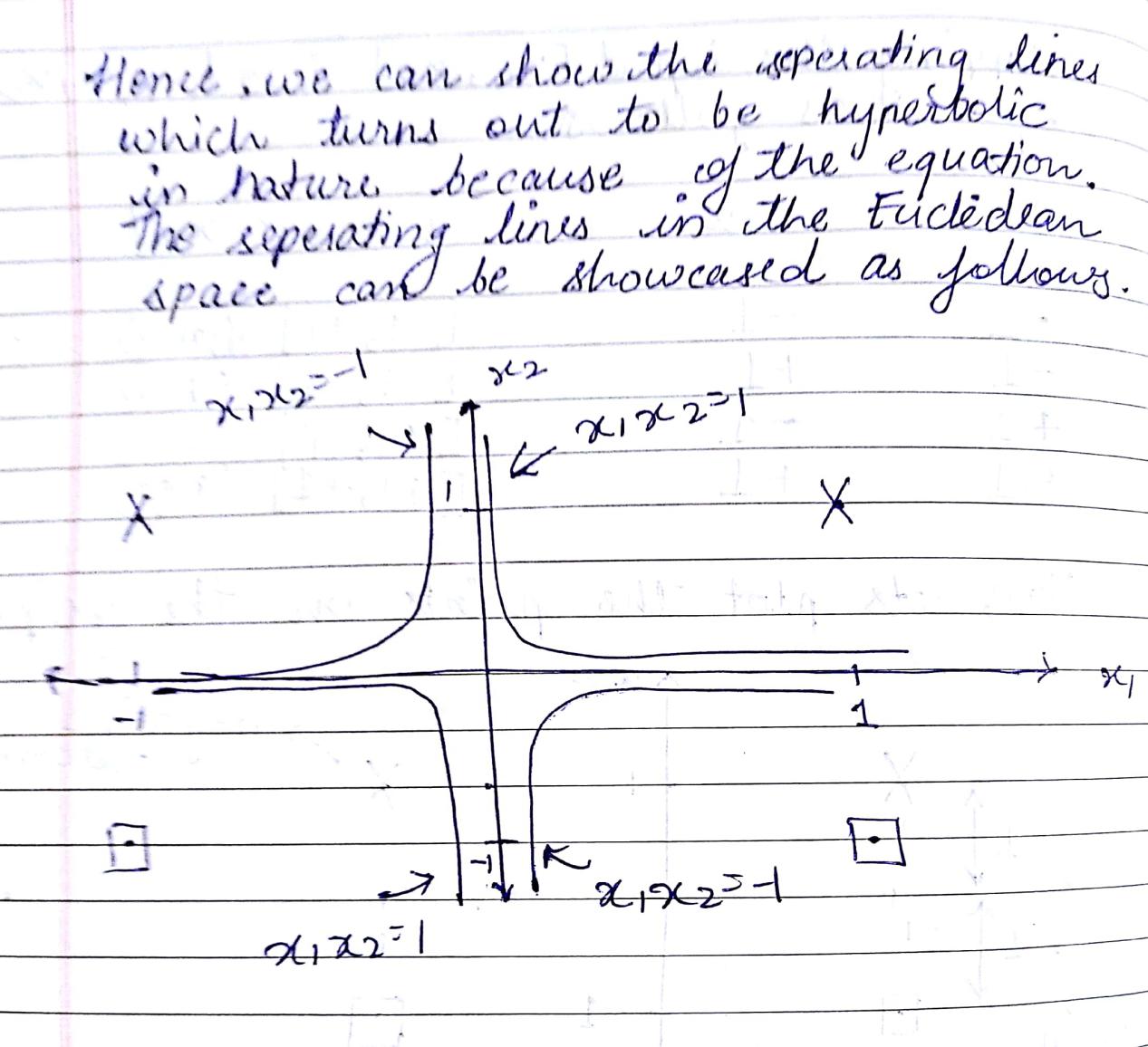
**b)**

The algorithm can not always represent the same tree, but since there are multiple ways of representing the tree there are chances of tree being logically equivalent. However, it totally depends on whether or not generation of the training set generate all the combinations possible as far as attributes are concerned.

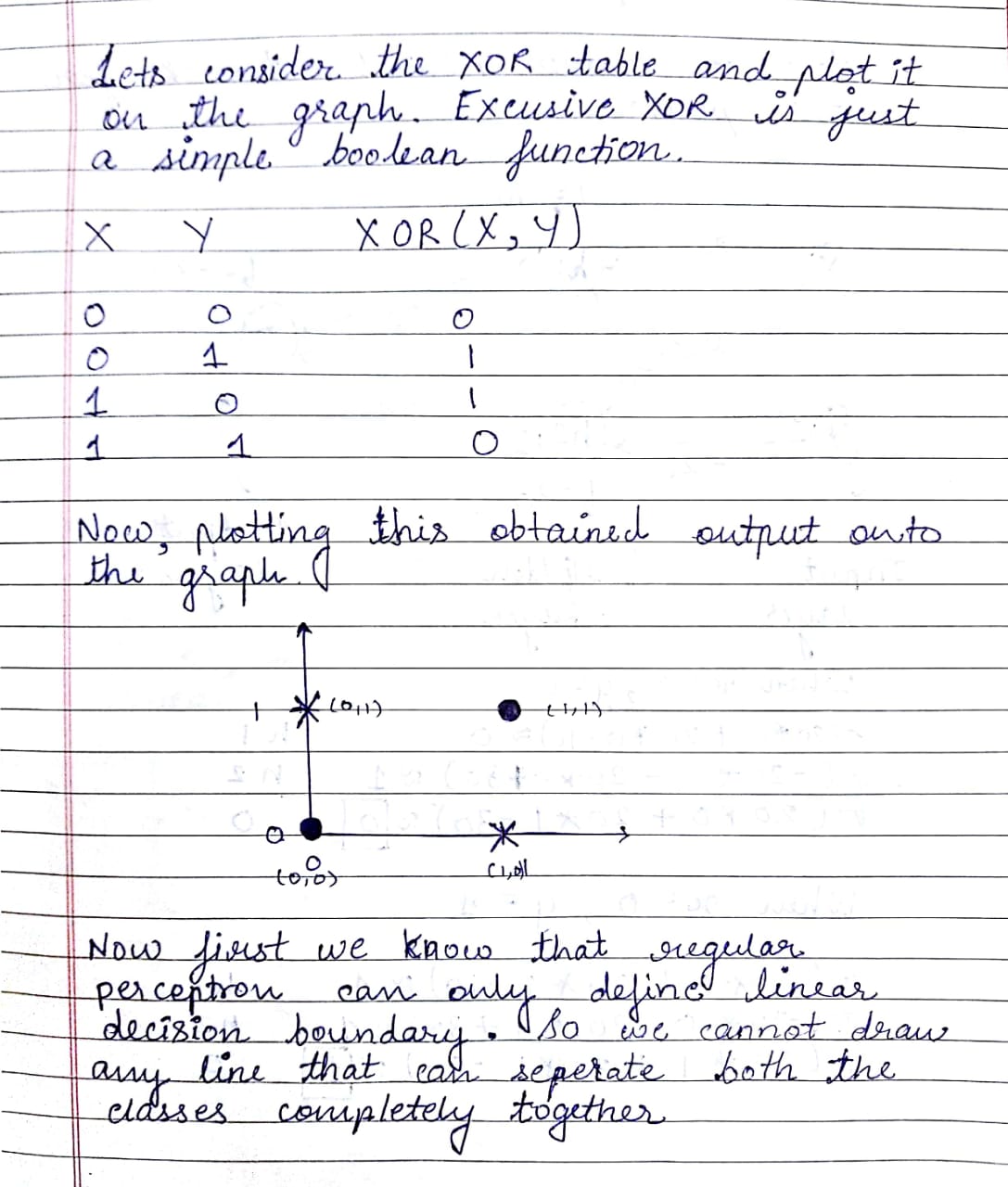
For instance, lets say a method selects the values for the attribute at random, the chances increases of having the same if not then at least logically when the training set is declared as Infinity.

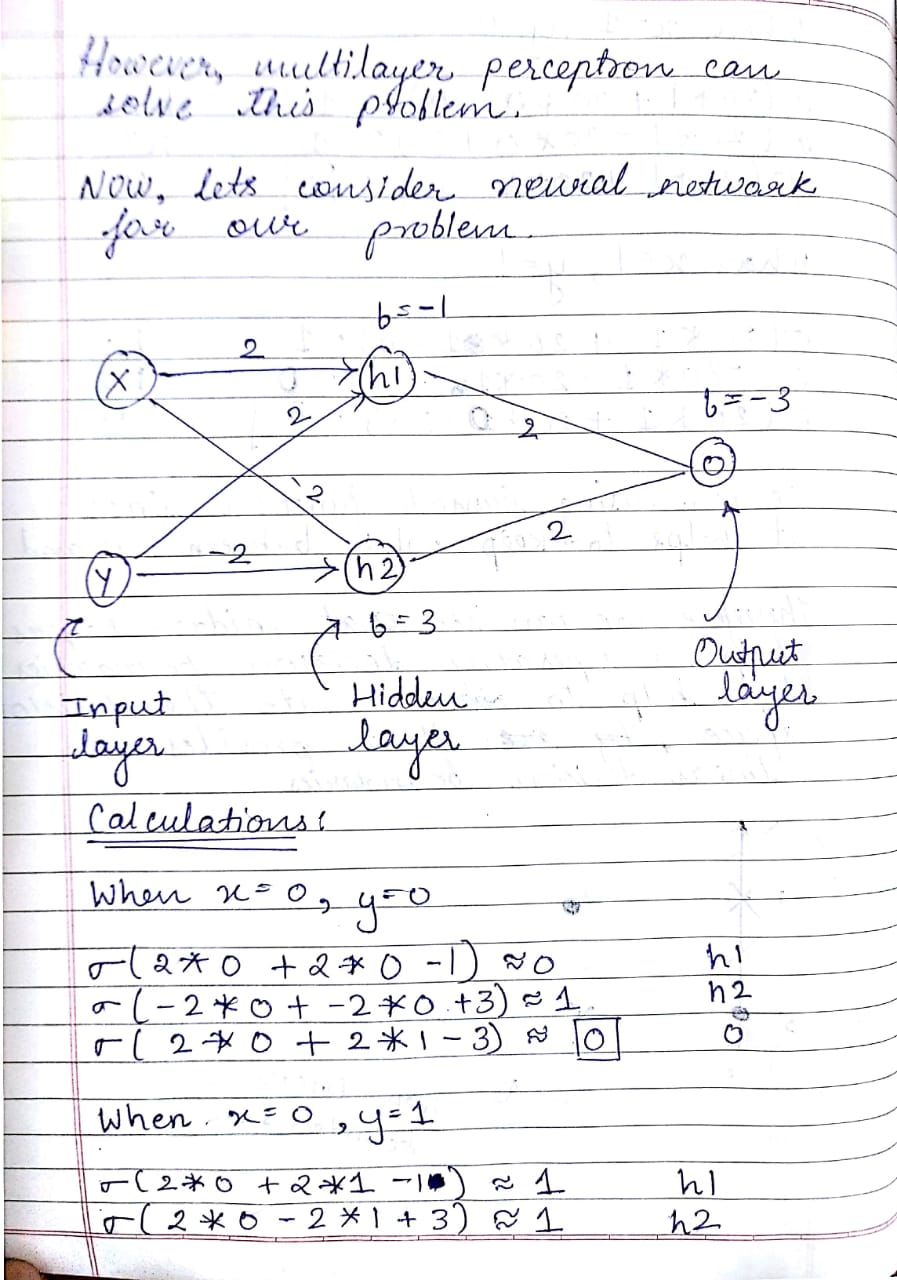
**c)**

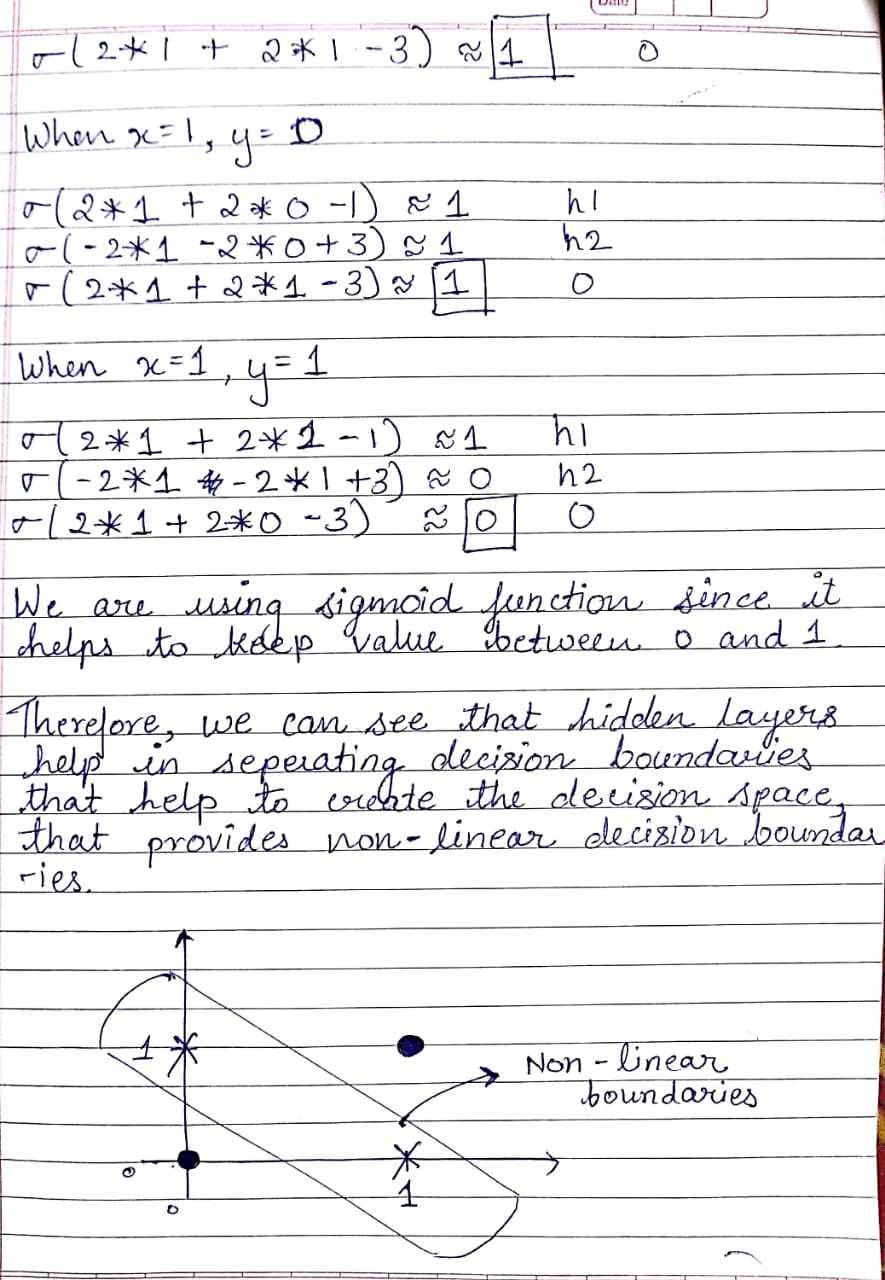




**d)**







**Q2>**

1. What are the **important features and how the features are correlated**? In simple terms what **factors** decide whether or not the customer is going to pay a loan or not? It is obvious that it is more likely dependent on Income, previous banking history, and so on rather than a person’s gender, school education.

l We use**the Random Forest Algorithm** in such a case since the algorithm helps in providing important and highly co-related features.

2. How can we**predict the probability** of not paying the loan for the given instance? Like we will have the training data set and based on it we can make a prediction in terms of probability, which algorithm suits here?

* First of all, as we have training data, we can go for **supervised learning,** and for prediction **logistic regressio**n would work well.

**3. Which** customer should we **target based on credit history?** This is a classification problem as it categories the customer into classes like Elite, Prime, Special, and General.

* According to me classification algorithms such as **Naive Bayes, K- nearest neighbor** would help to perform better in such cases.

4. What are the chances of a person getting**eligible for a loan, but won't repay**?

* **Multiple Regression algorithm** would fit the best.

**Q3>**

**A.**

1. **A flexible method will perform better in the given case.** It is because with the large sample size n available the flexible method will reduce the chances of over fitting and hence reduces the bias.
2. **A flexible method would perform worse in this case**. This is because with the small sample size a flexible model would lead to the over fitting.
3. **A flexible model would perform better in the given case.** This is because to accommodate the non-linearity flexible model would be more accurate than the inflexible one.
4. **A flexible model would perform worse** since the variance is extremely high the model would also take noise into the consideration. Noise is nothing but not useful information provided in the data set.

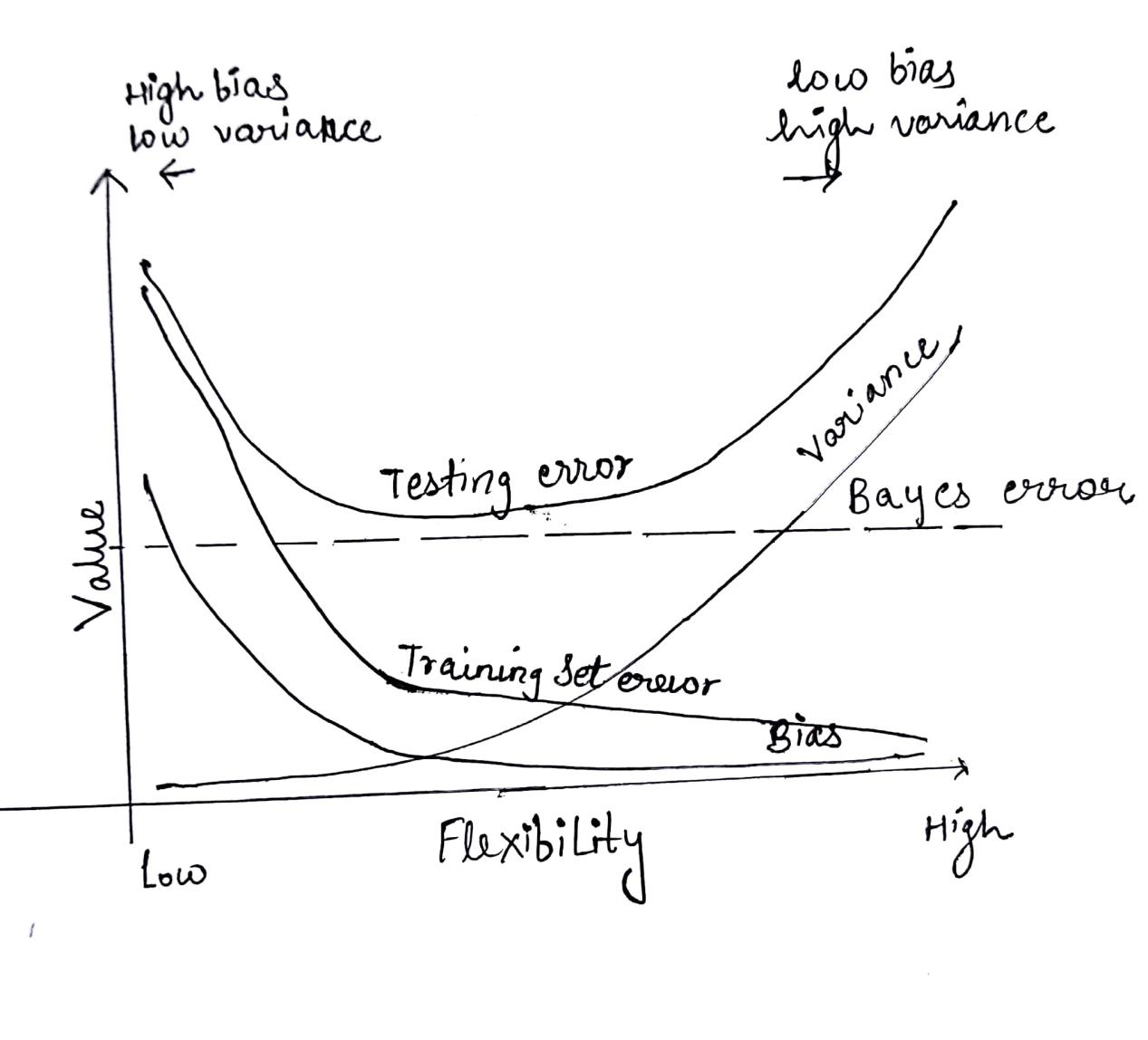
**B.**

**Regression** would be used in the above scenario since we need the outcome to be **Quantitative.** As far as inference or prediction is concerned, we are trying to **infer** the relation between the provided factors and its effect on the salary, **hence inference**. For the given problem, we have **n= 500 and p=3** since we have record profit, number of employees and industry for inferring.

**C.**

**a)**

Here is the graph for as per the question



**b)**

The **training error decreases with increase in the flexibility** of the model, since we are trying to fit the model on the given training set. So with the increase in flexibility the model fits the training data more precisely, reducing the training error.

The **testing error decrease at first, remains stable up to some point and then increases.** This happens because the as the flexibility increases the model **starts over fitting** the training data and making it difficult to test accurately based on over fitted data and hence testing error increases.

The **bayes error remains constant** as shown in the above diagram as it is dependent on the **data set** rather than that of the learning model.

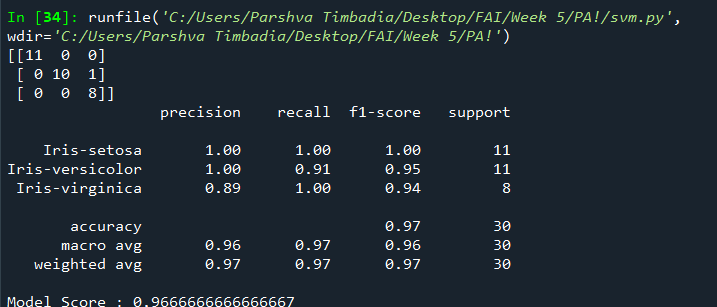
**Bias** is nothing but the average difference between the prediction of the model and the actual value. So when the model is **oversimplified** the **bias is more** and hence it **decreases with the increase in the flexibility.** Hence, we were able to see inverse relation with the bias and flexibility.

**Variance start to increase with the increase in flexibility.** Variance is nothing but the variability and sensitivity while considering the training data. So basically what happens is when the **flexibility is high**, the algorithm will keep on providing different models depending on the training set, there **by increase in the variance as tending towards over fitting.**

***PART B: SVM Screen Shots.***

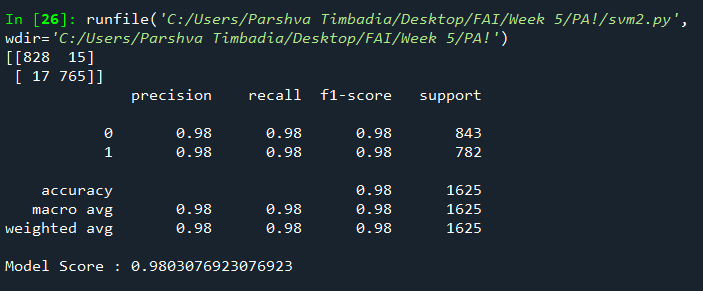
OUTOUT FOR IRIS DATA SET:

*File Name: svm.py*



OUTPUT FOR MUSHROOM DATA SET:

*File Name: svm2.py*



***Note:*** *While running the code, the code will provide different outputs every time you run the algorithm. This is because I have not included the random\_state in train and test split and hence it gets selected completely randomly. Also please read the multi-commented line with NOTE written in the code to get the better accuracy.*