Ziyan Guo

Machine Learning, Fall 2022: Project 2

Linear Regression：

**In this exercise, you will implement a linear regression model to predict the house price. For this**

**exercises use the dataset from the link below. Only use a single feature for you regression model and**

**explain your reasons for selecting that feature. Please explain the data setting and experimental setup**

**similar to Project 1.**

**explain your reasons for selecting that feature:**

I have chosen sqft\_living as the feature because the size of the house has the greatest impact on the price.

**data setting and experimental setup:**

I randomly divided the whole database into 20% test set and 80% sample set, using the (sklearn.model\_selection import train\_test\_split) it allow me Split arrays or matrices into random train and test subsets. In this way, the accuracy of the test can be guaranteed to the greatest extent.

Graphical user interface

Description automatically generated

Train data result.

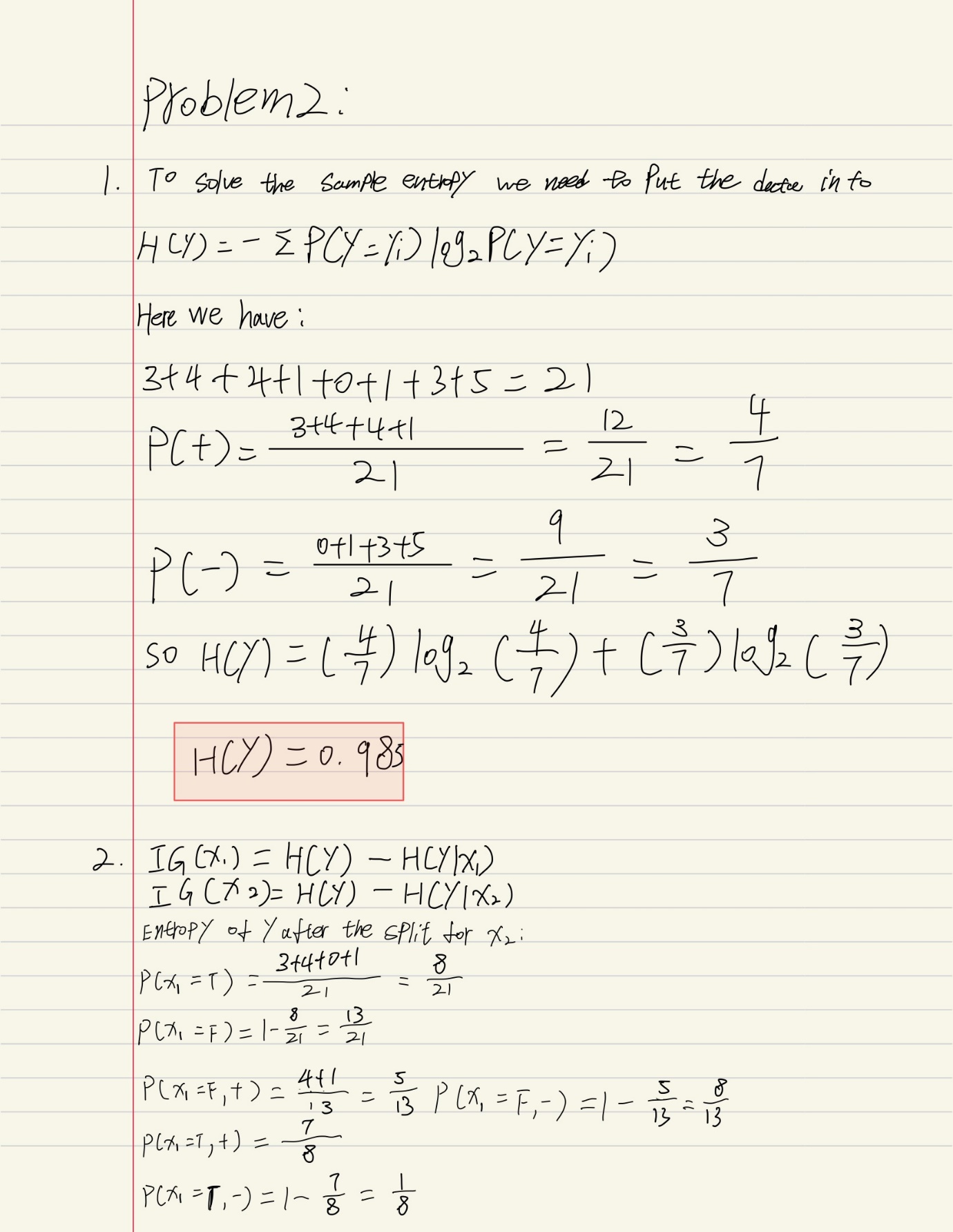
Graphical user interface

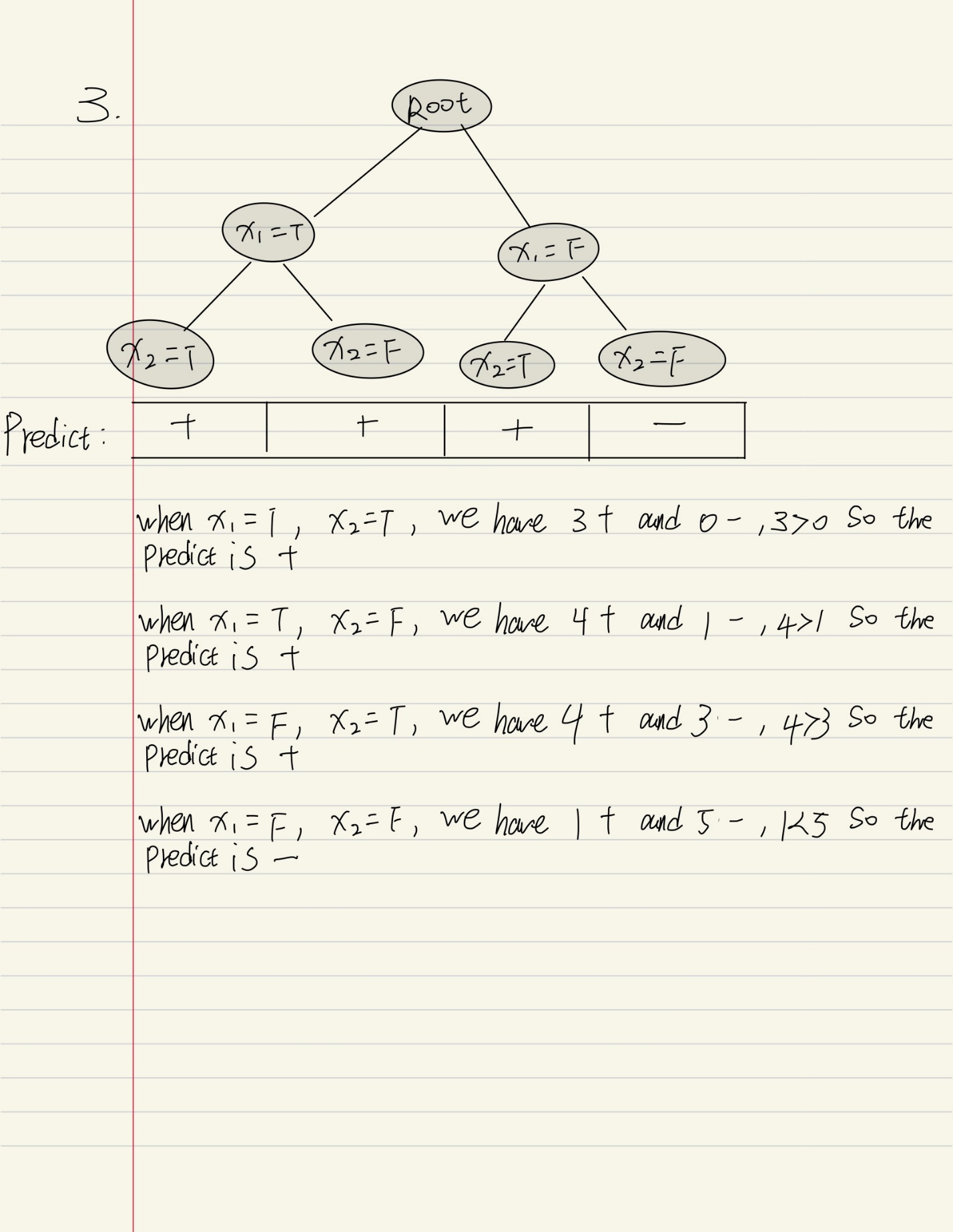
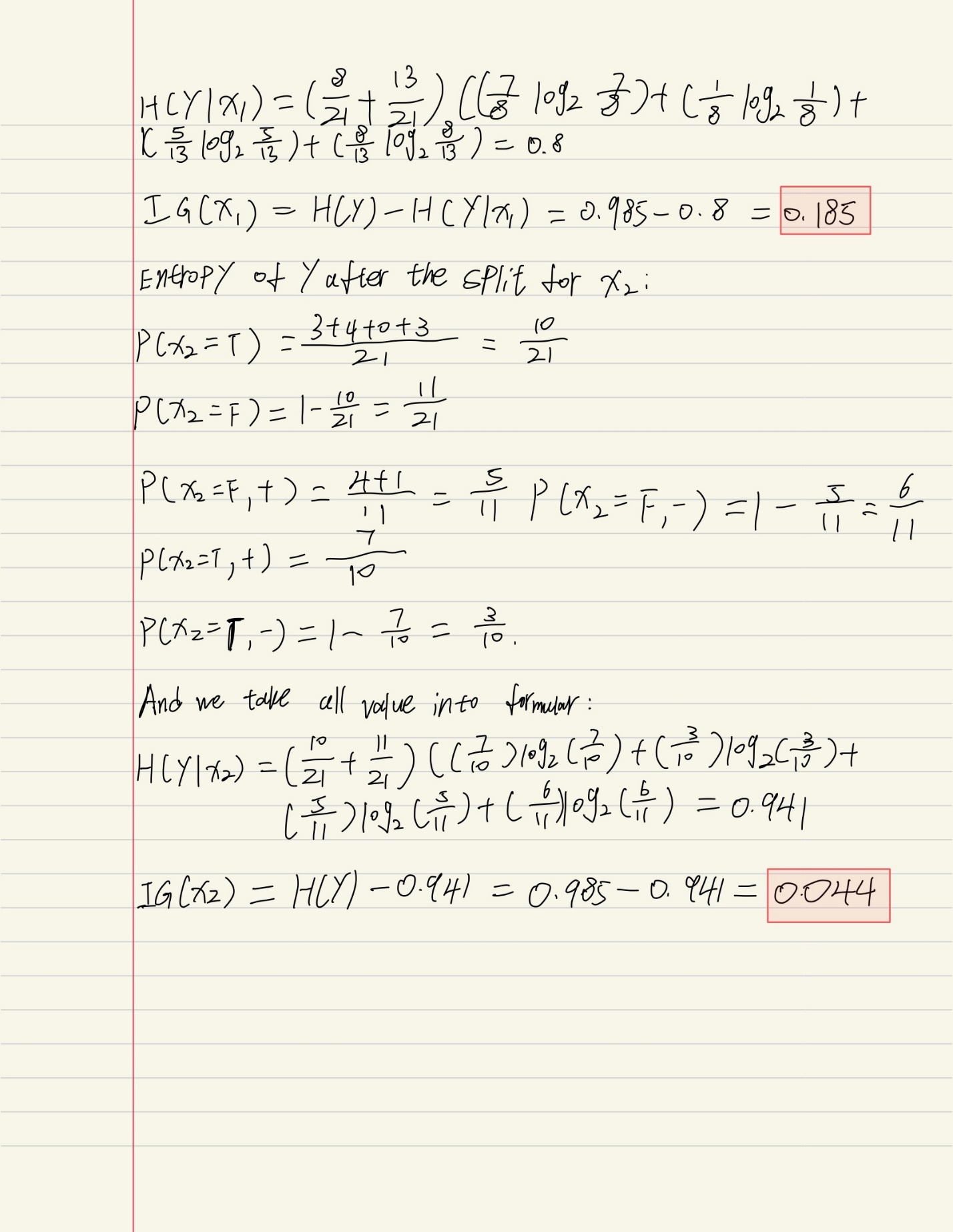
Description automatically generated

Test data result fit the train data.

Problem 2:

Decision Trees:





Problem 3：

Perceptron：

Text

Description automatically generated

**Graphical user interface

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**4. Support Vector Machine**

In this problem, you will repeat the format of Project 1 but using an SVM. On the Breast Cancer Wisconsin

(Diagnostic) Data Set. See associated link.

(30 points) Dataset details:

Breast Cancer Wisconsin (Diagnostic) Data Set：

About Dataset：

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Here is some scatter plots show a big picture of the data:

Chart, scatter chart

Description automatically generated

(This is "radius\_mean", "texture\_mean" scatter diagram can help understand the entire data structure)

I randomly divided the whole database into 20% test set and 80% sample set, using the (sklearn.model\_selection import train\_test\_split) it allow me Split arrays or matrices into random train and test subsets. In this way, the accuracy of the test can be guaranteed to the greatest extent.

(15 points) Algorithm Description: SVM is a very clear algorithm, so here describe any data pre-

processing, feature scaling, distance metrics, or otherwise that you did.

**Data preprocessing:**

**For the data of this data set, I kept the useful features, removed the feature columns that are not useful to me, such as ID, and put these data into the test and training sets for preprocessing.**

**Feature scaling:**

**I scaled the too large data in the eigenvalues, and limited the data between 0 and 1, which can reduce its influence on other eigenvalues and improve the overall accuracy.**

(45 points) Algorithm Results:

图形用户界面, 文本

描述已自动生成

Accuracy around 97.36%

文本

描述已自动生成

(10 points) Runtime: Describe the run-time of your algorithm and also share the actual "wall-clock" time that it took to compute your results.

You may find time counter on the end of the all the print. And

 the actual "wall-clock" time for Breast Cancer Dataset will be around 27 seconds.