## 1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)

[Crime Data from 2020 to Present | Los Angeles - Open Data Portal (lacity.org)](https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data)

Dataset is Crime Data from 2020 in the city of LA. This dataset is categorical and contains useful insights of crime happening all over LA.

**Statistics:**

| DR\_NO | TIME OCC | AREA |
| --- | --- | --- |
| count 8.834470e+05  mean 2.173593e+08  std 1.139567e+07  min 8.170000e+02  25% 2.103145e+08  50% 2.203220e+08  75% 2.302215e+08  max 2.499046e+08  missing values | count 883447.000000  mean 1336.738147  std 653.112632  min 1.000000  25% 900.000000  50% 1415.000000  75% 1900.000000  max 2359.000000 | count 883447.000000  mean 10.702601  std 6.101298  min 1.000000  25% 6.000000  50% 11.000000  75% 16.000000  max 21.000000 |

| Part 1-2 | Crm Cd | Vict Age |
| --- | --- | --- |
| count 883447.000000  mean 1.411699  std 0.492141  min 1.000000  25% 1.000000  50% 1.000000  75% 2.000000  max 2.000000 | count 883447.000000  mean 500.847664  std 207.620420  min 110.000000  25% 331.000000  50% 442.000000  75% 626.000000  max 956.000000 | count 883447.000000  mean 29.670338  std 21.823780  min -3.000000  25% 0.000000  50% 31.000000  75% 45.000000  max 120.000000 |

| Premis Cd | Weapon Used Cd | Crm Cd 1 |
| --- | --- | --- |
| count 883447.000000  mean 306.245656  std 217.198988  min 101.000000  25% 101.000000  50% 203.000000  75% 501.000000  max 971.000000 | count 883447.000000  mean 126.034277  std 187.661917  min 0.000000  25% 0.000000  50% 0.000000  75% 400.000000  max 516.000000 | count 883447.000000  mean 500.582506  std 207.416067  min 0.000000  25% 331.000000  50% 442.000000  75% 626.000000  max 956.000000 |

| Crm Cd 2 | Crm Cd 3 | Crm Cd 4 |
| --- | --- | --- |
| count 883447.000000  mean 69.766992  std 250.696349  min 0.000000  25% 0.000000  50% 0.000000  75% 0.000000  max 999.000000 | count 883447.000000  mean 2.423104  std 48.833437  min 0.000000  25% 0.000000  50% 0.000000  75% 0.000000  max 999.000000 | count 883447.000000  mean 0.069547  std 8.304651  min 0.000000  25% 0.000000  50% 0.000000  75% 0.000000  max 999.000000 |

| LAT | LON |
| --- | --- |
| count 883447.000000  mean 33.986423  std 1.725889  min 0.000000  25% 34.014300  50% 34.058500  75% 34.163500  max 34.334300 | count 883447.000000  mean -118.050971  std 5.983354  min -118.667600  25% -118.429700  50% -118.321500  75% -118.273900  max 0.000000 |

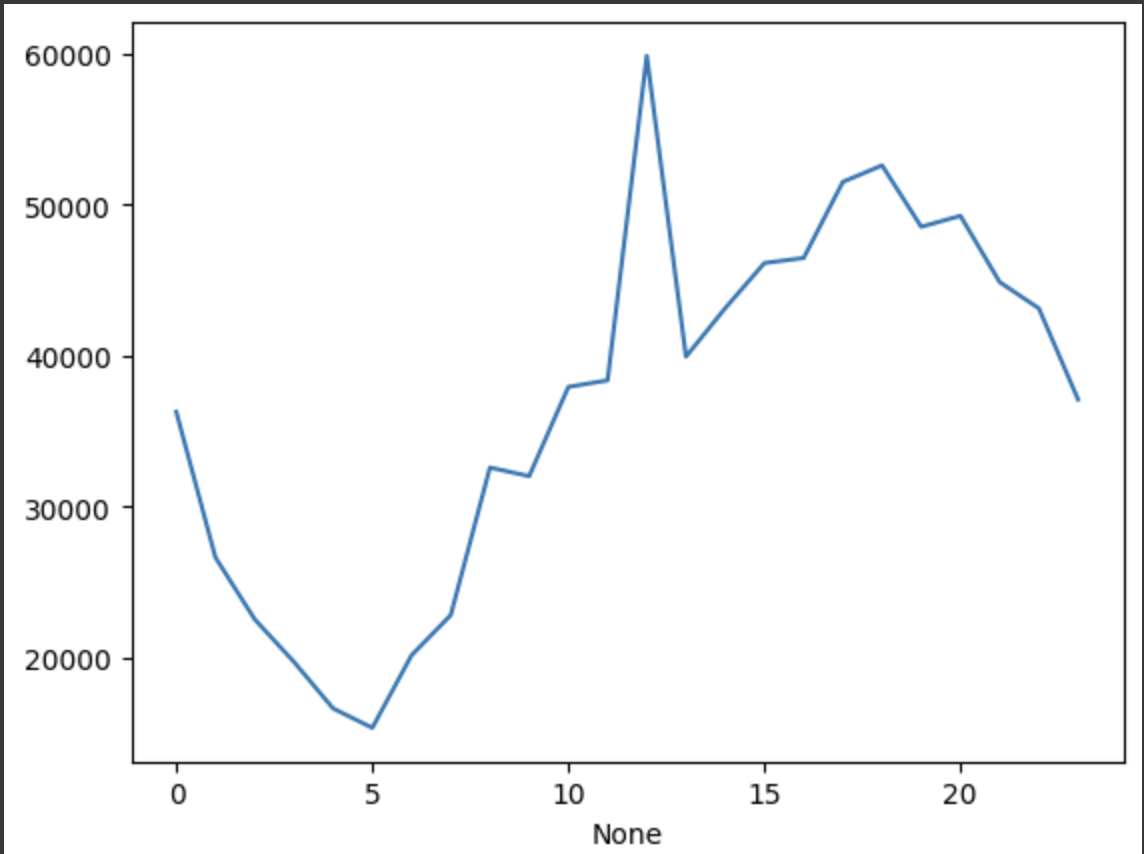
## 2. What kind of preprocessing techniques have you applied to this dataset?

### Processing Technique

1. Handle missing values
2. Convert DATE OCC in string format to MONTH OCC, DAY OF WEEK OCC
3. Identify values that are not common and weren’t described on the dataset official page. Merge all of them into ‘unknown’ value or drop them
4. For some columns that are missing because they weren’t present in the case like weapons not used, or no victim involved. Fill those with 0

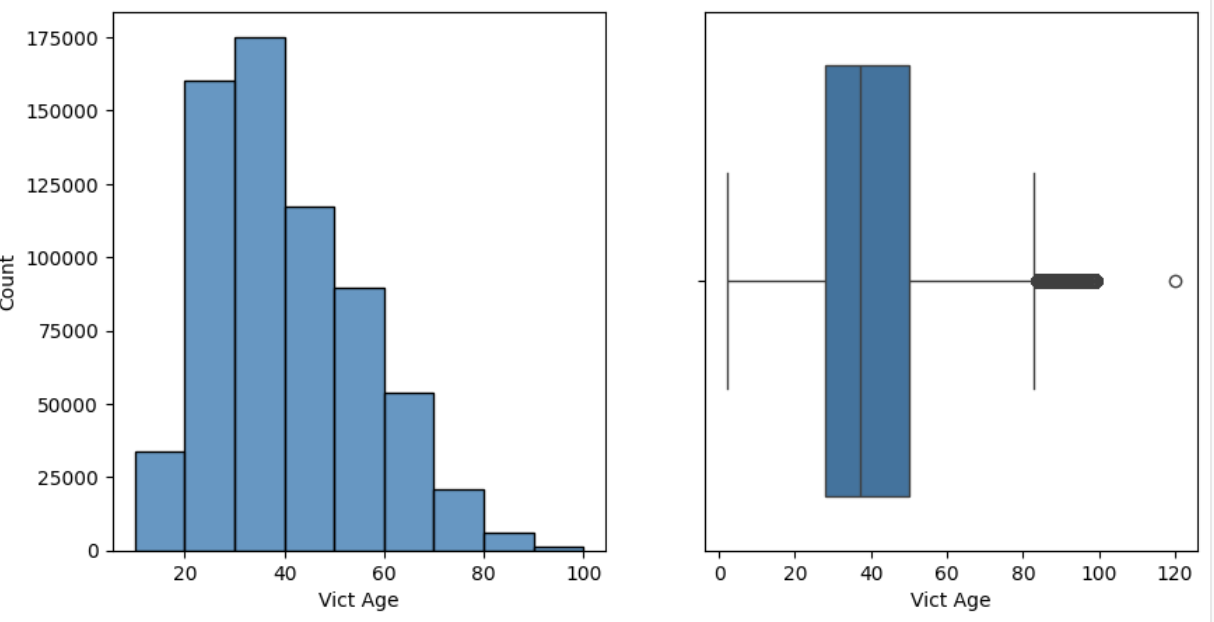
## 3. Provide at least 5 visualization graphs with a brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.

### Graph 1 Hour Occured



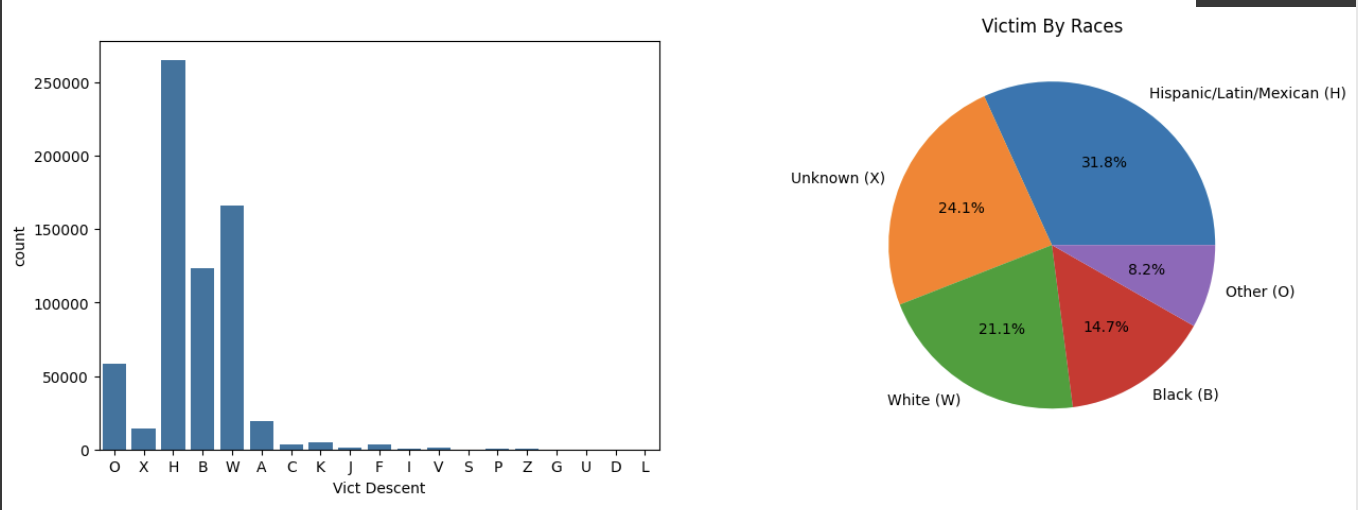
Hour frequency that crime occured. We can see it peaks around 19 PM and steadily goes down after 20PM

### graph 2 Victim Age



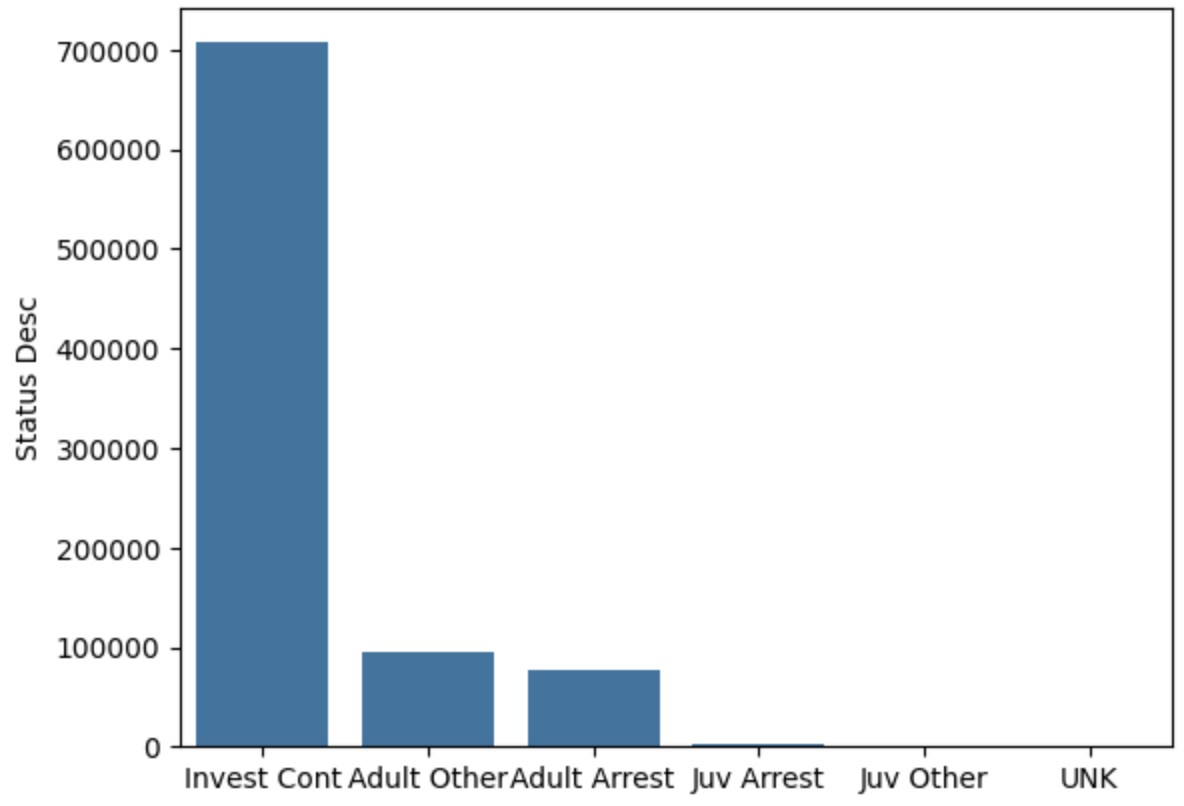
histogram of victim age. Most victims are in 20s or 30s years of age

### graph 3 Victim demographicc



This most common victm race is Hispanic people.

### Graph 4 Crime status



Most crime cases are in investigation.

### Graph 5 Month Occured

### 

Crime occurrence peak around summer months (from June to September)

## 4. Provide brief details and mathematical representation of the ML methods you have used. What are the key features? What are the advantages/disadvantages?

### 1.MLP using backpropagation to update weights

The predicted output follows this formula:

y = ɸ(

**key features:** MLP is a feed-forward neural network that uses backpropagation to train. It’s a simple NN with an adjustable number of hidden layers.

| **Advantages** | **Disadvantages** |
| --- | --- |
| MLP is able to learn complex relationships between features and the target and fit into non-linear dataset better than a simple linear classifier | Tunning MLP takes time because each dataset requires different parameters |
|  | MLP requires proper scaling or it won’t be optimal |

Accuracy trained on Crime Dataset: 60%

Loss value of MLP = 0.67100171

### 2. Decision Tree

A classification method that learns simple decision rules in the dataset to predict.

**Features:**

A decision tree is a simple and versatile classification model. The most important feature of a decision tree is its easy interpretability.

| **Advantages** | **Disadvantages** |
| --- | --- |
| No domain knowledge is required to train | easy to overfit |
| Effective with high-dimension data | required to train the whole model again if new data is available |

Accuracy trained on Crime Dataset: 62%

### 3. Random Forest

**Features:**

### Random Forest is an ensemble method that fits many decision trees into sub-samples of the dataset to improve robustness and accuracy

| **Advantages** | **Disadvantages** |
| --- | --- |
| Multiple decision trees are used to fit sub-samples of the dataset |  |
|  |  |

Accuracy trained on Crime Dataset: 61%

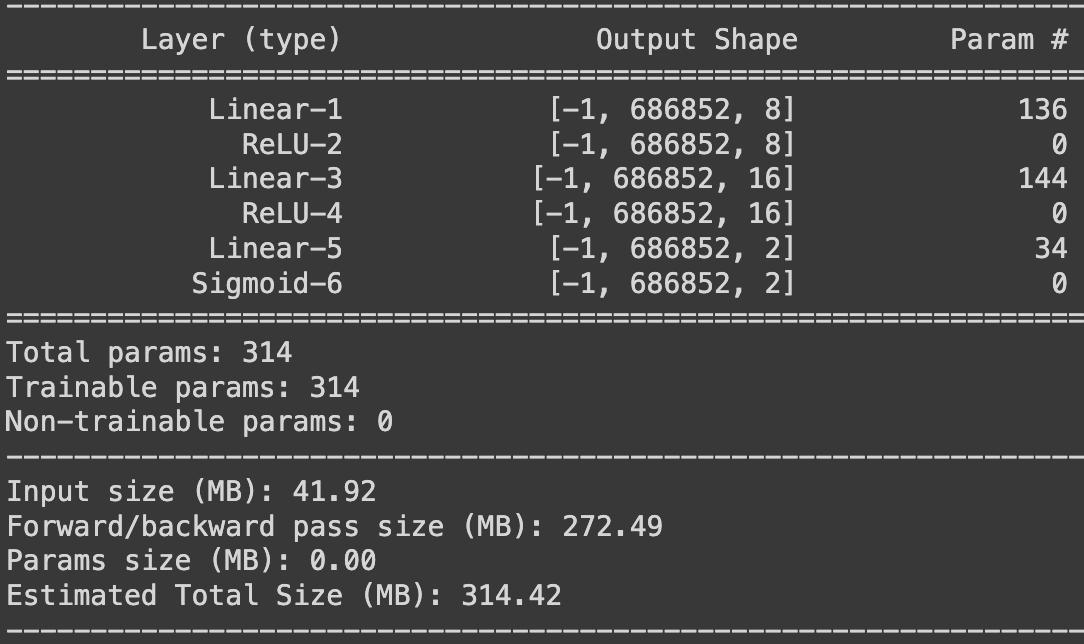
## 

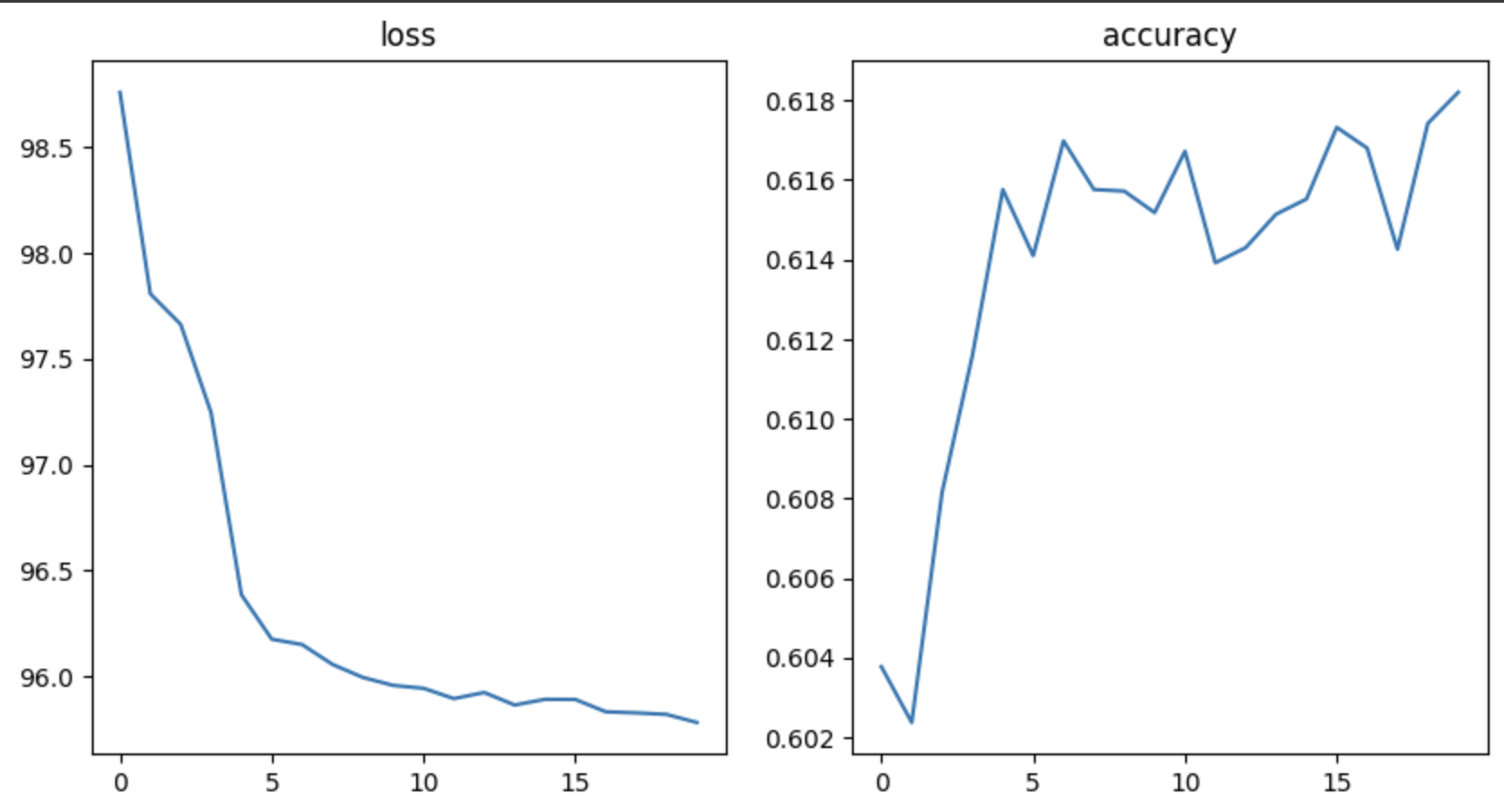
## 5. Provide your loss value and accuracy for all 3 methods.

## 6. Show the plot comparing the predictions vs the actual test data for all methods used.

## Analyze the results. You can consider accuracy/time/loss as some of the metrics to compare the methods.

## 7. Provide the neural network structure you have built to solve the problem defined in Part I. Show the plot. Analyze the results.





**Evaluation:**

After 20 training epochs. the NN model reached its minimum loss values and maximum accuracy.

## 

# Part II: Deep Learning Theoretical Part [20 points]

## 1. Perform a forward pass and estimate the predicted output (ŷ)

y\_pred = 0.5655

## 2. Estimate the MSE

MSE = 0.0021

## 3. Find the gradient using back-propagation

w1 = w1 - lr \* (-(y - y\_pred) \* w7 \* i1)

w2 = w2 - lr \* (-(y - y\_pred) \* w8 \* i1)

w3 = w3 - lr \* (-(y - y\_pred) \* w7 \* i2)

w4 = w4 - lr \* (-(y - y\_pred) \* w8 \* i2)

w5 = w5 - lr \* (-(y - y\_pred) \* w7)

w6 = w6 - lr \* (-(y - y\_pred) \* w8)

w7 = w7 - lr \* (-(y - y\_pred) \* h1)

w8 = w8 - lr \* (-(y - y\_pred) \* h2)

w9 = w9 - lr \* (-(y - y\_pred) \* (1)

## 4. Update the weights and the bias

w1 = -0.30096285

w2 = 0.799656125

w3 = 0.14931225

w4 = 0.199754375

w5 = 0.8986245

w6 = -0.14049125

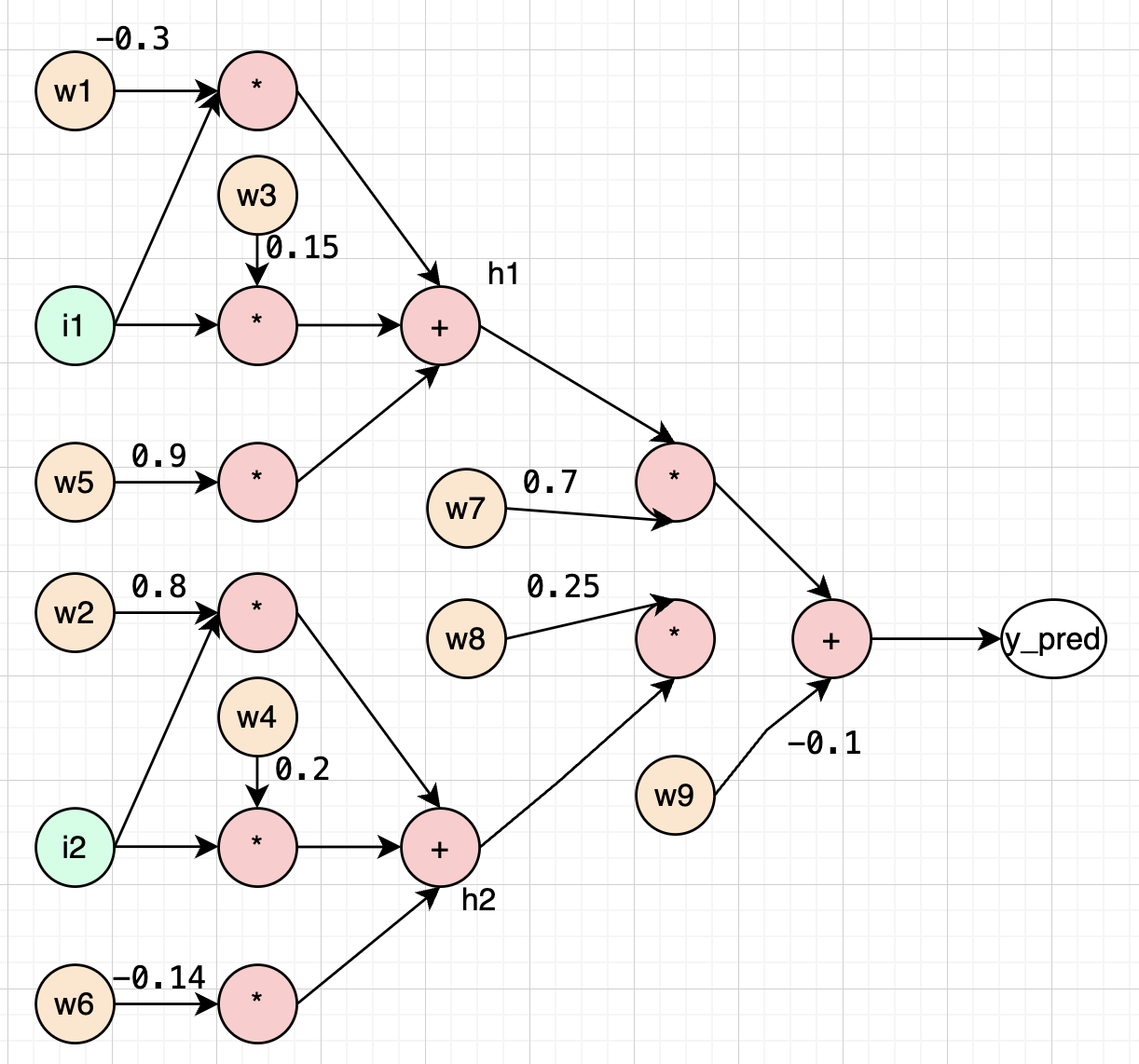
w7 = 0.698496775

w8 = 0.2489782

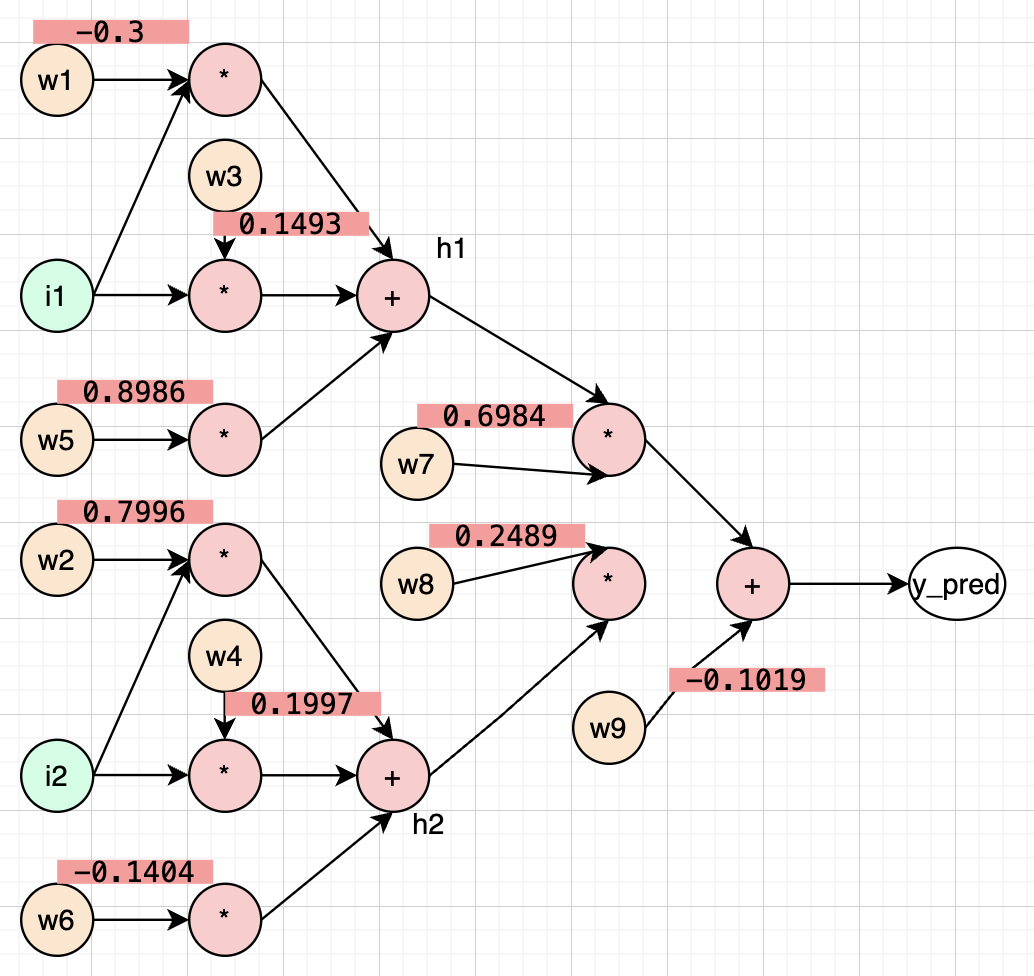
w9 = -0.101965

## 5. Draw a computation graph for the forward and backward pass

**Forward Pass**



**Backward Pass**



## 6. Perform a forward pass to estimate the predicted output using the updated weights.

new y\_pred 0.5599691153077132

## 7. Estimate the MSE and compare the results with Step 2.

new MSE = 0.0017981473953949012