# QUY TRÌNH PHÂN TÍCH VÀ
# GIẢI QUYẾT BÀI TOÁN TÍNH LƯƠNG NHÂN VIÊN

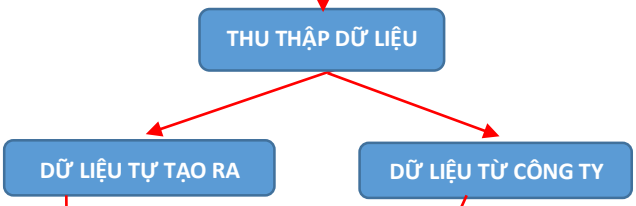
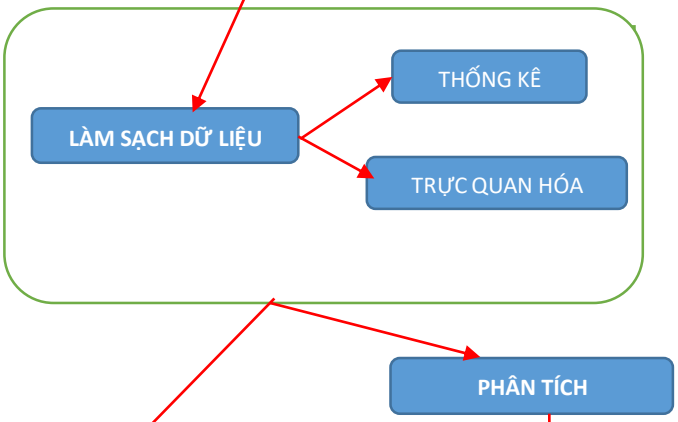
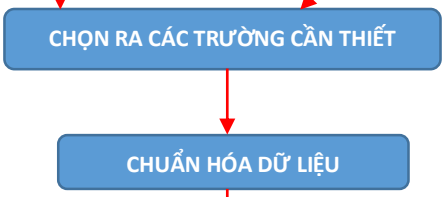
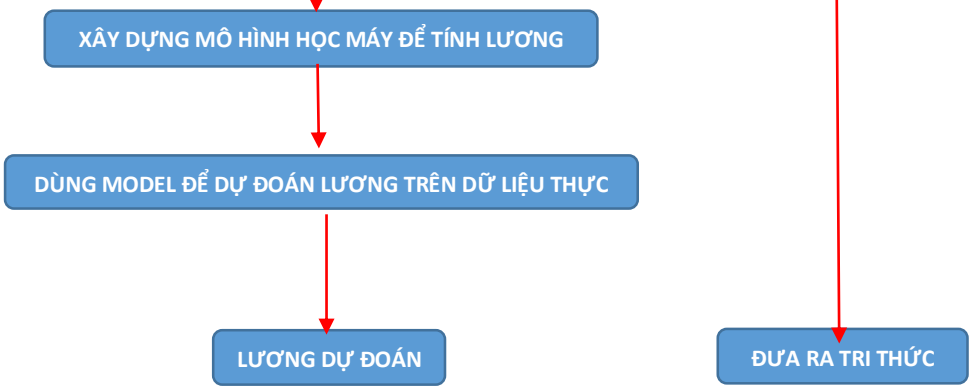
***

**Người trình bày:** Hoàng Đức Long

**Gmai:** *hoangduclongg@gmail.com*

**SĐT:** 0384856300

# 1.SƠ ĐỒ TỔNG QUAN

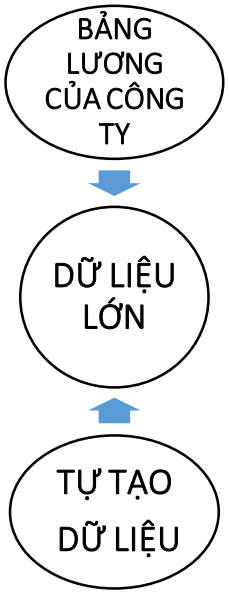
| STT | MÔ TẢ | LƯU ĐỒ |
|---|---|---|
| 1 | Phát biểu bài toán. |  |
| 2 | Thu thập dữ liệu từ nhiều nguồn khác nhau. |  |
| 3 | Phân tích dữ liệu |  |
| 4 | Tiền xử lí |  |
| 5 | Giải quyết bài toán. |  |

# 2.QUY TRÌNH THỰC HIỆN

## 2.1.Phát biểu bài toán

Tính tiền lương cho 1 nhân viên ở 1 công ty công nghệ

## 2.2.Thu thập dữ liệu

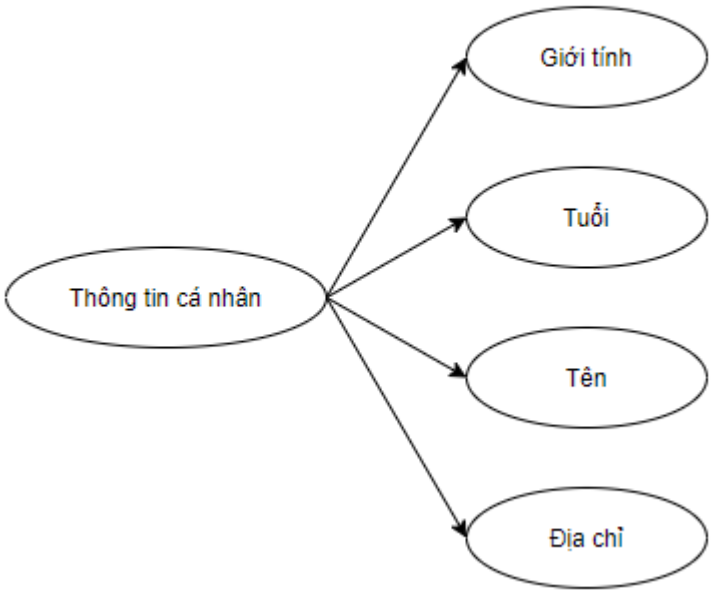



## 2.3.Phân tích dữ liệu

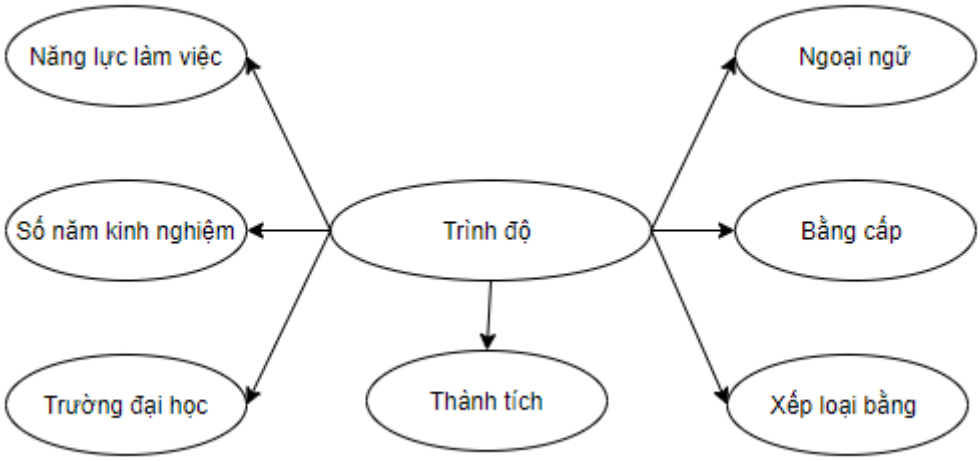
## 2.3.1.Cách tạo ra dữ liệu

### 2.3.1.1.Phân tích các trường dữ liệu:

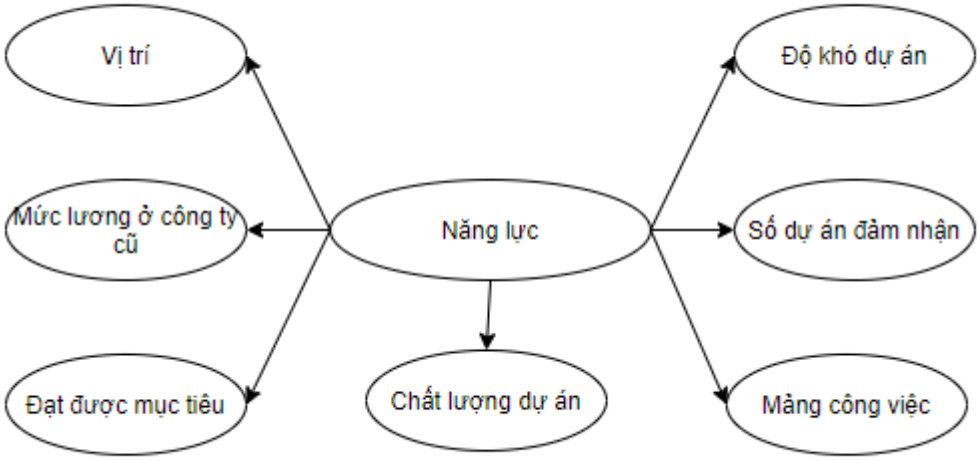
- **Thông tin cá nhân :**



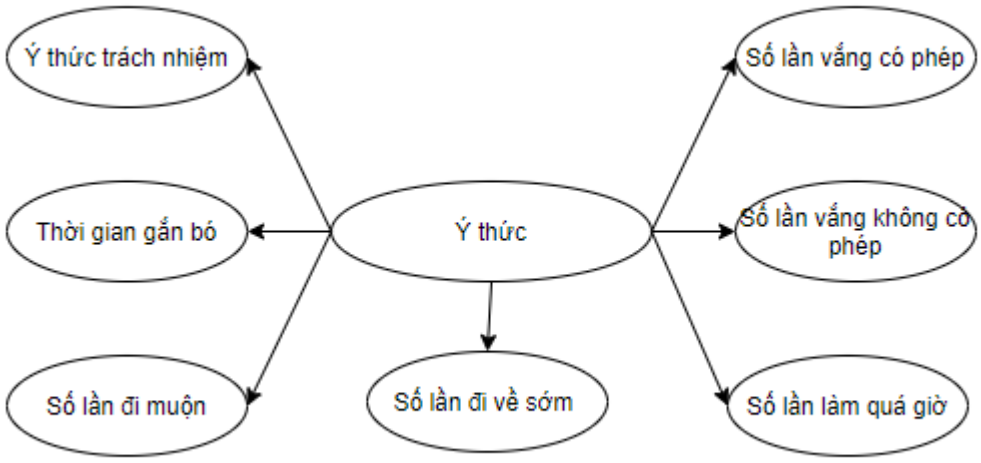
- **Trình độ :**



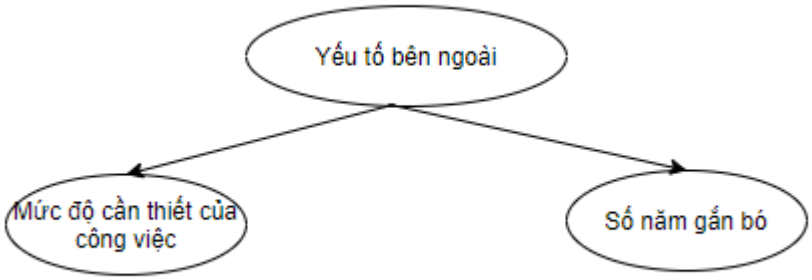

- **Năng lực :**




- **Ý thức :**

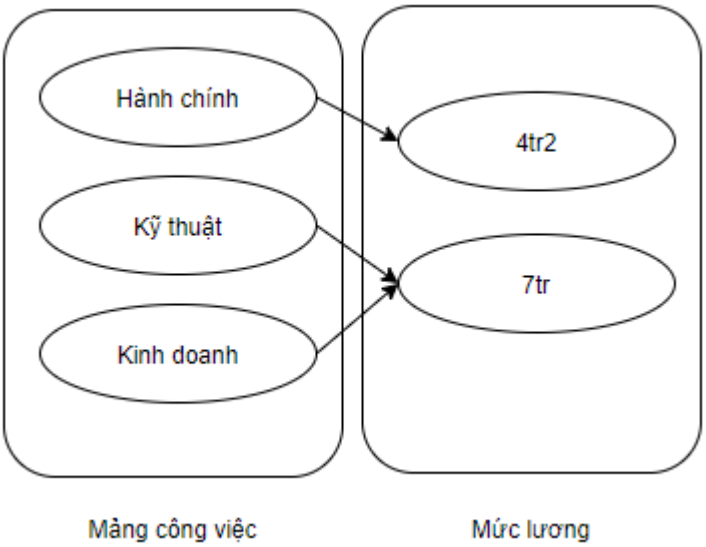


- **Các yếu tô bên ngoài :**

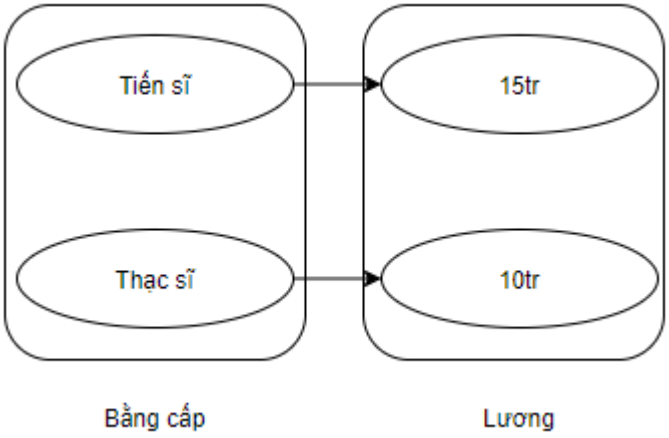



## 2.3.1.2.Giả sử các ràng buộc của dữ liệu:

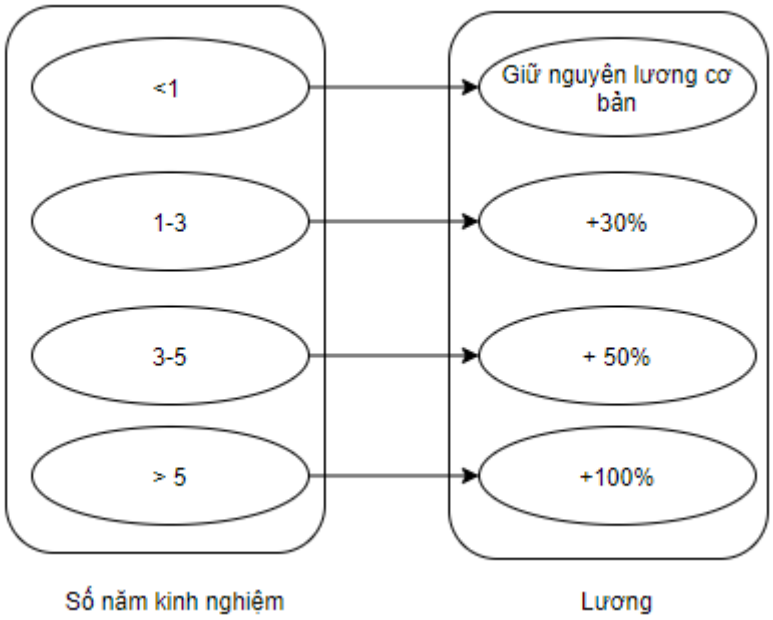
- ❖ Giả sử trong vai trò của 1 công ty, dùng 5 thuộc tính dữ liệu để định mức lương :
    1. Mảng công việc.
    2. Bằng cấp.
    3. Số năm kinh nghiệm.
    4. Vị trí.
    5. Năng lực công việc.
- ❖ Cách thức tính lương dựa vào các thuộc tính :
    1. Mảng công việc :



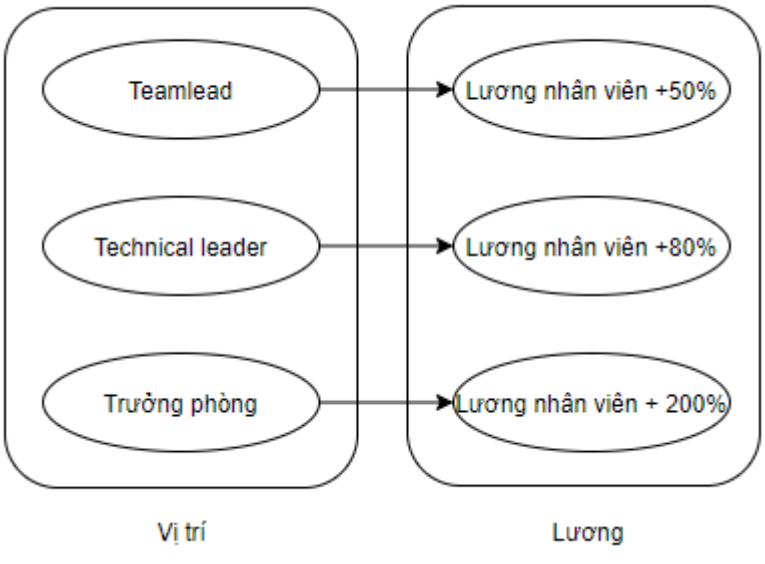
2. Bằng cấp:



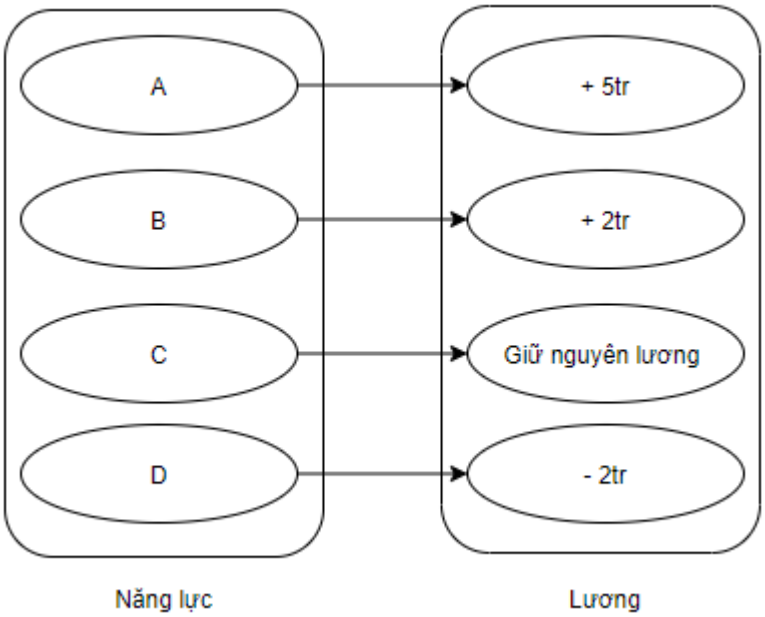

3. Số năm kinh nghiệm:




4. Vị trí:



5. Năng lực công việc:




### 2.3.2.Dữ liệu từ bảng lương của công ty.

### 2.3.2.1.Làm sạch dữ liệu.

- Loại bỏ các trường dữ liệu không cần thiết.
- Hiện tại vì chưa có dữ liệu đang dùng dữ liệu mô phỏng.

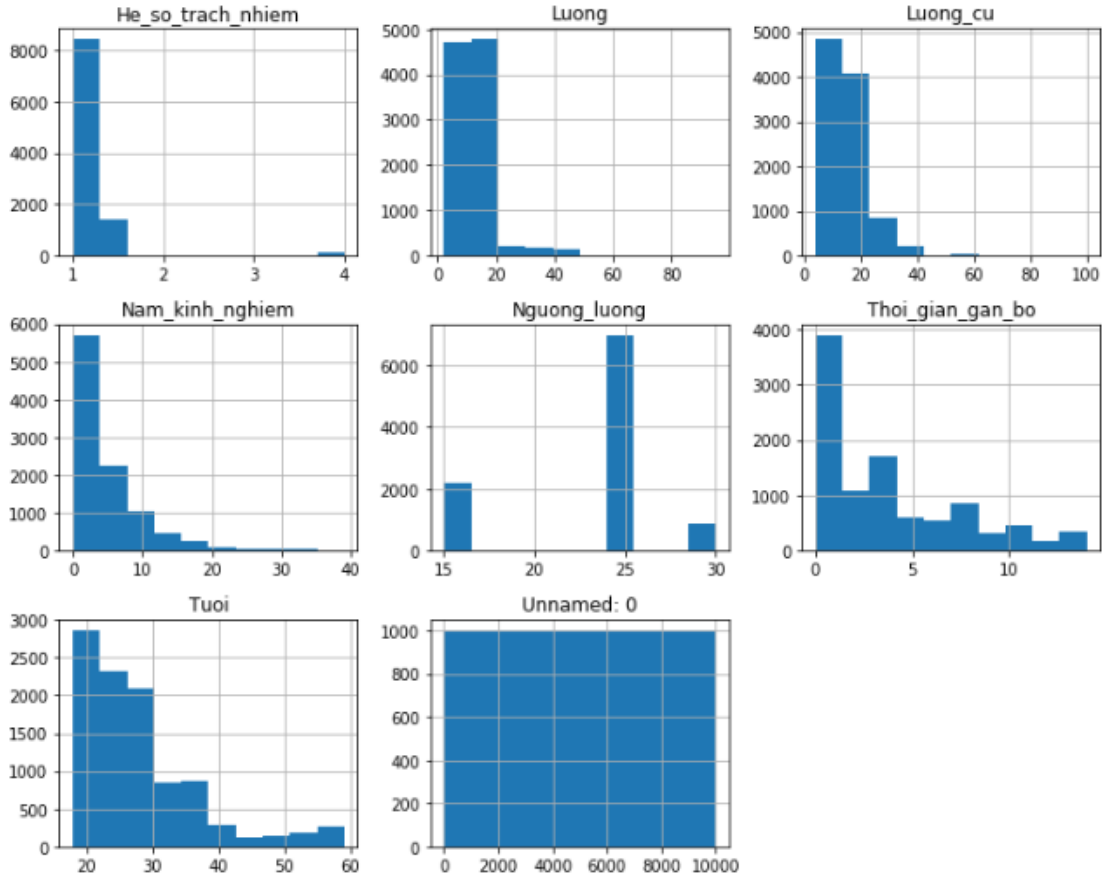
### 2.3.2.2.Thống kê và trực quan hóa dữ liệu.

- Vì chưa có dữ liệu thật nên bước này sẽ dùng dữ liệu mô phỏng.
- Thống kê các trường dữ liệu(số hàng mỗi trường,kiểu dữ liệu).

```
#rawdf.info()
salary_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 6 columns):
Bang_cap           10000 non-null object
Mang_cong_viec     10000 non-null object
Nam_kinh_nghiem    10000 non-null int64
Vi_tri             10000 non-null object
Nang_luc_work      10000 non-null object
Luong              10000 non-null float64
dtypes: float64(1), int64(1), object(4)
memory usage: 468.8+ KB
```

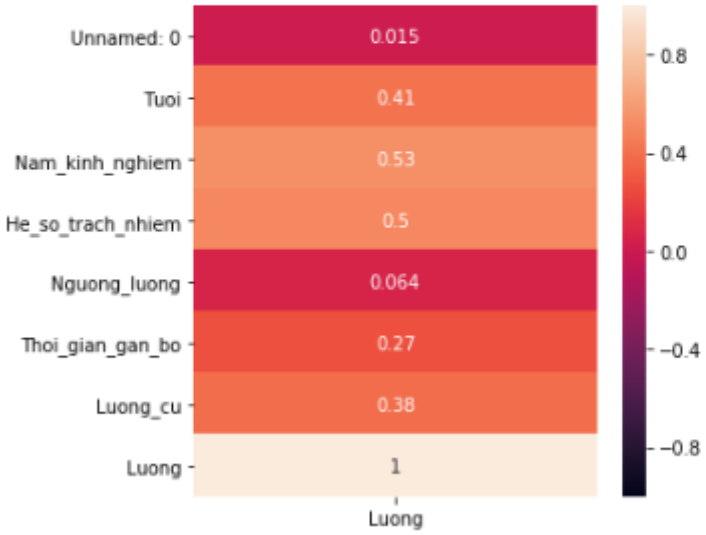
- Biểu đô thể hiện mức độ phân bố các trường dữ liệu.




- Biểu đồ tương quan trường **Lương** với các trường khác.

```
plt.figure(figsize=(5, 5))
#sns.heatmap(rawdf.corr()[['Luong']], annot=True, vmin=-1, vmax=1)
sns.heatmap(rawdf.corr()[['Luong']], annot=True, vmin=-1, vmax=1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x276f18e3e80>
```



- Biểu đồ tương quan giữa các trường với nhau .

```
plt.figure(figsize=(10, 8))
#sns.heatmap(rawdf.corr(), annot=True, square=True, vmin=-1, vmax=1)
sns.heatmap(rawdf.corr(), annot=True, square=True, vmin=-1, vmax=1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x276f1965a20>
```


## 2.4.Tiền xử lí

### 2.4.1.Lựa chọn các trường dữ liệu.

- ❖ Giả sử trong vai trò của 1 công ty, dùng 5 thuộc tính dữ liệu để định mức lương :
    1. Mảng công việc.
    2. Bằng cấp.
    3. Số năm kinh nghiệm.
    4. Vị trí.
    5. Năng lực công việc.

### 2.4.2.Chuẩn hóa dữ liệu

Chuyển đổi dữ liệu theo cách one-hot representation.

| Năng lực làm việc |
|---|
| A |
| B |
| C |
| D |

| A | B | C | D |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

## 2.5.Giải quyết bài toán

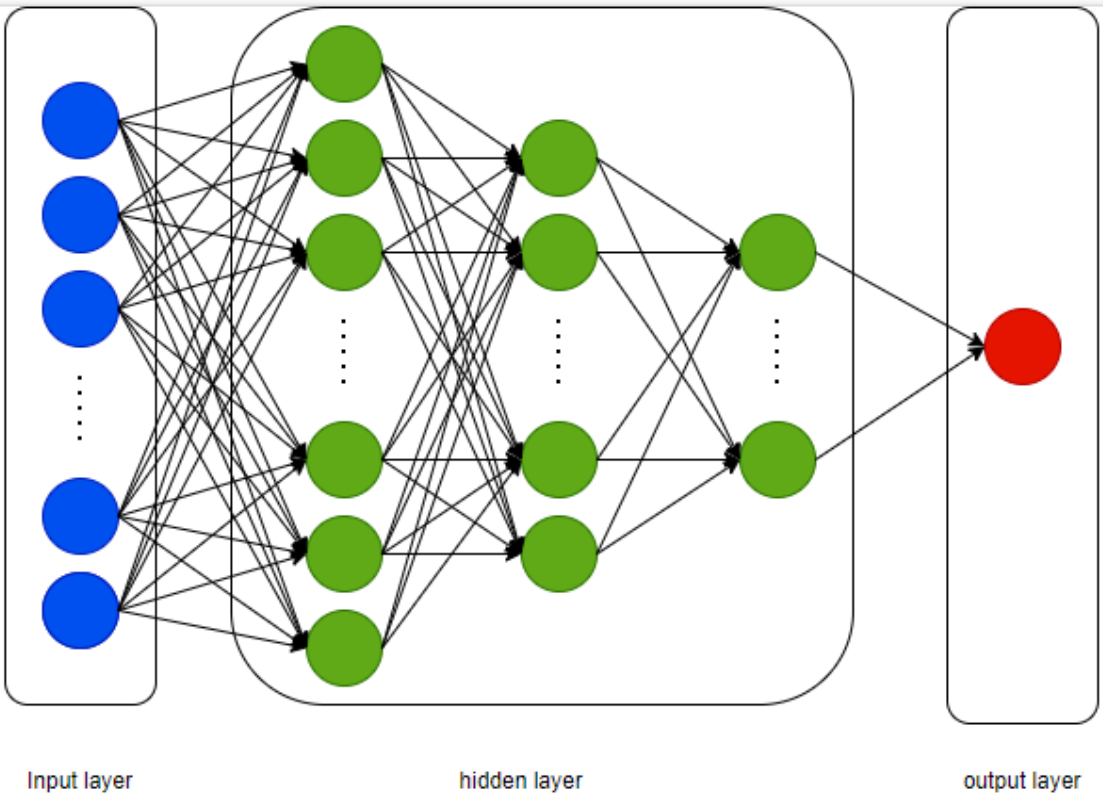
### 2.5.1.Rút ra tri thức từ phân tích bài toán.

- Từ kết quả thống kê,trực quan hóa bằng biểu đồ ... Ta có thể rút ra được tri thức như sau:
    1. Ai là người lương cao nhất công ty, thông tin người đó ....
    2. Số người lương cao hơn mức N.
    3. Từ các biểu đồ ta có thể biết được số người lương trên 15 triệu chiếm bao nhiêu %

    .....

### 2.5.2.Xây dựng mô hình học máy để tính lương.

- Với bài toán này thì ta sẽ dựng mô hình học máy hồi quy để dự đoán lương.
- Mô hình thử nghiệm là mô hình hồi quy ANN với kiến trúc :



1. Input layer : Dữ liệu được chuẩn hóa.
2. Hidden layer: Gồm 3 lớp :
    - Lớp 1 có 32 nodes.
    - Lớp 2 có 16 nodes.
    - Lớp 3 có 8 nodes.
3. Output layer : 1 nodes thể hiện lương dự đoán

### 2.5.3.Sử dụng model đã xây dựng để dự đoán.

❖ Đánh giá mô hình với tập test (tập test gồm 1000 dữ liệu) với độ chênh lệch giữa lương mô phỏng và dự đoán 1 triệu.

```
# Đánh giá độ chính xác với ngưỡng chênh < 1tr
accuracy = (abs(y_test - y_predict) <= 1.0).mean() * 100
print("Độ chính xác dự đoán lương = ", accuracy,"%")
```

```
Độ chính xác dự đoán lương =  96.3 %
```

### 2.5.4.Lương dự đoán.

❖ Thử nghiệm model với dữ liệu.

```
for i in range(10):
    print("predict = ", y_predict[i])
    print("label = ", y_test[i])
    print('\n')
```

```
predict =  [10.490693]
label =  [11.1]


predict =  [46.993397]
label =  [47.]


predict =  [8.446825]
label =  [9.1]


predict =  [16.19112]
label =  [15.5]


predict =  [18.863745]
label =  [19.]


predict =  [14.680143]
label =  [15.]


predict =  [8.618092]
label =  [9.1]


predict =  [11.959455]
label =  [12.]


predict =  [29.04948]
label =  [30.2]


predict =  [12.625004]
label =  [12.]
```