Alvin Anderson - UTS Machine Learning 2301978006

I. Discrete random variables adalah sebuah variable yang jumlah kemungkinan akan nilainya bisa kita prediksi. Contohnya adalah kasus melempar koin.

IF(X==heads), X=0
IF(X==tails), X = 1

Dalam kasus ini, nilai yang mungkin keluar sebagai output dari melempar koin hanya ada 2, yaitu heads / tails. Jika kita melempar koin, kemungkinan output dari lemparan kita hanyalah kedua nilai tersebut. Kemungkinan yang terbatas ini (2) disebut Discrete random variables. Inti dari Discrete random variables adalah mempunyai kemungkinan yang terbatas (Meskipun terlihat tidak terbatas). Contoh kasusnya ada dibawah seperti ini.

Y = Number of ants born tommorow

Kita, tentu pasti kesulitan menghitung jumlah semut yang akan lahir di hari esok. Jumlahnya mungkin tampak sulit dihitung seperti 4 miliar atau bahkan sampai quadra triliun. Namun, itu tetap bisa dihitung atau Number of ants born tummorow punya jumlah yang terbatas (finite number).

Selanjutnya adalah continuous random variables. Continuous random variables memiliki nilai yang tidak terbatas. Kita tidak bisa membatasi kemungkinan dalam continuous random variables. Sebagai contoh :

Z = Mass of a random animals lived in the world

Massa dari hewan acak yang kita pilih, bisa mempunyai nilai berapa saja. Di range 0 – 100kg mungkin. Massa seekor binatang bisa aja 0.008732kg. Lalu bisa juga massa binatang lainnya adalah 7.832122 atau 7.343212. Kita bisa lihat dari sini, bahwa kita tidak bisa membatasi kemungkinan massa seekor binatang. Massa dari binatang acak yang hidup di dunia ini bisa seberat atau seenteng apapun. Kondisi inilah yang disebut oleh continuous random variables yang memilki kemungkinan yang tidak terbatas (infinite values).

Analysis

```
In [4]: # 1b) Check NaN Values

        num_vars = dataset.columns[dataset.dtypes != 'object']
        dataset[num_vars].isnull().sum()
```

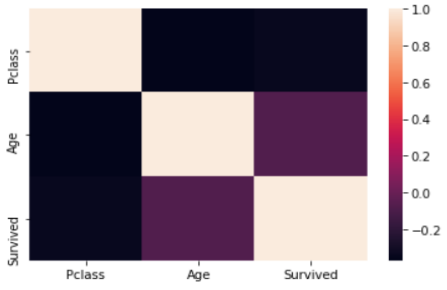
```
Out[4]: PassengerId      0
        Pclass           0
        Age            177
        SibSp            0
        Parch            0
        dtype: int64
```

Ada 177 nilai NaN didalam dataset ‘titanic.txt’.

```
In [8]: # 1c) Correlation Between 2 - 3 Variables

        corr_data = data[['Pclass','Age','Survived']]
        correlations = corr_data.corr()
        sns.heatmap(correlations,
                    xticklabels = correlations.columns,
                    yticklabels = correlations.columns)
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x2dc5f493e50>
```



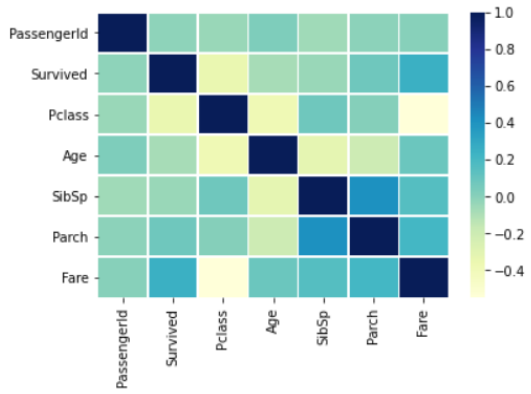
Di nomor 1c, saya mengambil 3 variabel, yaitu Survived, Age, dan PClass. Dari antara 3 variabel tersebut, Age dan Survived memiliki hubungan (walaupun lemah). Sedangkan pada variabel lainnya, tidak ditemukan hubungan apapun (warna hitam pekat)

Correlation

```
In [10]: # 1d) Correlation Between Independent Variables and Dependent Variables

         correlations = data.corr()
         sns.heatmap(correlations,
                     xticklabels = correlations.columns,
                     yticklabels = correlations.columns,
                     cmap="YlGnBu",
                     linewidths = 0.5)
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x2dc5fc33760>
```



Dari gambar diatas, kita bisa melihat bagaimana variabel satu bisa mempengaruhi variabel lainnya. Menurut gambar diatas, yang mempunyai hubungan yang paling kuat adalah SibSp dan Parch. Lalu dilanjutkan oleh Survived dan Fare, lalu Parch dan Fare dan seterusnya.

Linear Regression

```
In [13]: from sklearn.linear_model import LinearRegression
         regressor = LinearRegression()
         regressor.fit(x_train,y_train) #actually produces the linear eqn for the data
```

Sklearn adalah sebuah library dalam python yang mempunyai fungsi LinearRegression. LinearRegression mengandung data training (otomatis men-generate weight dan bias untuk training data)

Activation Function

```
In [14]: # 3a) Prediction of passenger survived

         y_pred = regressor.predict(x_test)
         y_pred = np.round(y_pred)
```

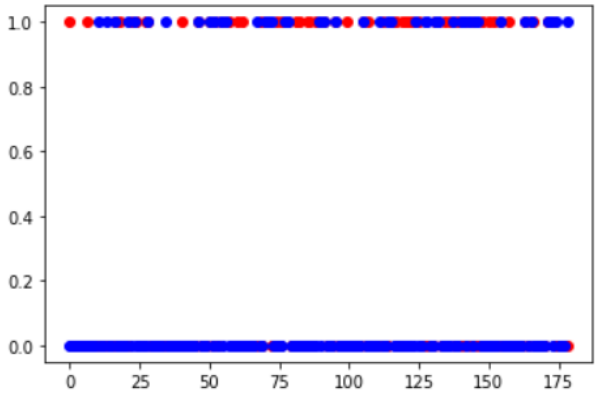
Fungsi dari np.round adalah untuk membulatkan hasil dari prediksi kita ke 0 atau 1. Jika prediksi kita lebih kecil dari 0.5, maka y_pred akan diklasifikasikan sebagai 0. Sedangkan jika prediksi kita lebih besar dari 0.5, makan y_pred akan diklasifikasikan sebagai 1. Hal ini untuk mendapatkan akurasi tepat untuk kita nantinya. Karena, pada kolom Survived, hanya ada nilai 0 dan 1 atau nilai True or False.

Y_Predicted vs Y_Actual

```
|: # 5) Perbandingan antara Hasil prediksi variabel survival dan hasil survival yang sebenarnya

plt.scatter(x = list(range(0, len(y_test))), y = y_test, color = 'red')
plt.scatter(x = list(range(0, len(y_pred))), y = y_pred, color = 'blue')
```

```
|: <matplotlib.collections.PathCollection at 0x2dc5fea2460>
```



Ini adalah gambaran perbandingan antara prediksi kita dan hasil yang sebenarnya. Bisa dilihat dari gambar ini bahwa model ini memiliki akurasi sekitar 70%.

Pembahasan

Sebenarnya, Linear Regression kurang cocok digunakan dalam kasus ini. Linear Regression digunakan untuk memprediksi variable yang bersifat continuous / mempunyai infinite values / ada kemungkinan yang tidak terbatas dalam memprediksi valuesnya. Sedangkan pada kasus ini yang terjadi adalah kita ingin memprediksi apakah seseorang selamat dalam kecelakaan titanic. Hanya ada 2 output kemungkinan dari kasus ini, yaitu selamat atau tidak atau jika dinilaikan dimatematika mempunyai nilai 0 atau 1 / True or False. Output nilai yang terbatas ini termasuk kedalam discrete variables. Sehingga, sebenarnya kasus ini lebih cocok diprediksi menggunakan klasifikasi.