**Bài toán Titanic: predict survised/not survised passenger on Titanic**

[https://www.kaggle.com/rafaelpleite/titanic-analysis-and-predict/notebook?fbclid=IwAR20w0-oDMWrg9vd5Vxg6znQF0pw9jFExopa1uBCX7Dpyz7nf7mhonDL5J0](Link:%20https://www.kaggle.com/rafaelpleite/titanic-analysis-and-predict/notebook?fbclid=IwAR20w0-oDMWrg9vd5Vxg6znQF0pw9jFExopa1uBCX7Dpyz7nf7mhonDL5J0)

Đây là bài toán phân lớp classification 0: not survised và 1: survised

Từ Initial Dataset

**Step 1: Exploratory Data Analysis**

* Visualize data (nhóm gộp từng cặp Pclass – Surviced; Fare – Surviced, Sex-Surviced… -> so sánh mối tương quan giữa tỷ lệ sống sót và các feature khác -> đưa ra dự đoán ( hành khách ở khoang hạng cao, trả nhiều tiền, giới tính nữ có tỷ lệ sống sót cao hơn…)

**Step 2: Pre-processed Dataset**

* Data curation, data cleaning (data nào thiếu thì correct lại,)
* Remove redundant features (feature có tỷ lệ missing cao (vd: số cabin) hoặc string type feature (vd: name, ticket) -> các features này không ảnh hưởng đến outcome predict -> remove)
* Normalize data to bring all variable to same range [0,1]

**Step 3: Data splitting:** 20% use as test set, 80% as train set

**Step 4: Build model**

* Build up a regression model using Keras from Tensorflow to handle missing data (Age, Fare)
* Build up a regression model using Keras from Tensorflow to predict Survival
* Predicted Survival

**Step 5:Evaluate Model Performance:**

* Accuracy metric: model có độ chính xác 81% và đây là tỉ lệ giữa tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử
* Confusion matrix:  thể hiện có bao nhiêu điểm dữ liệu thực sự thuộc vào một class, và được dự đoán là rơi vào một class

0: not survised và 1: survised

|  |  |  |
| --- | --- | --- |
| Total 819 | Predicted as:0 | Predicted as: 1 |
| True 0 | 508 | 41 |
| True 1 | 128 | 214 |
|  |  |  |

508: Model đoán đúng 508 người không sống sót.

41: 41 người không sống sót nhưng model đoán sai thành sống sót

128: 128 người sống sót nhưng model đoán sai thành không sống sót

214: Model đoán đúng 128 người sống sót