Airlines Classification

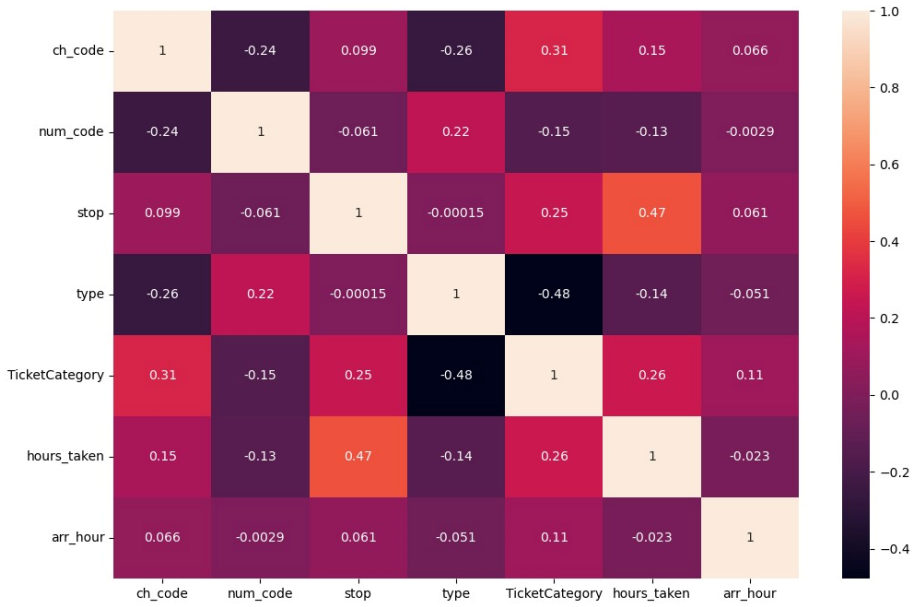
Team ID: 14_SC

Team members:

- Abdelrhman Mohamed Abdelsalam
- Omar Khaled Ahmed Abdullah
- Ibrahim Youssef Mustafa
- Rehab Hosam Ahmed Mokhtar
- Maivel Maher Isaac

- **Preprocessing:**
  - **TicketCategory column:** the categories have been encoded by using ordinal encoder which handle the unknown values by replacing them with specific value, in our case unknown values take value 4 (number of categories + 1 -zero based).
  - **Date column:** convert all the records to the standard format (MM/DD/YYYY), then separate them to 3 columns (day, year and month).
  - **Ch_code column:** the values have been encoded by using ordinal encoder and the unknown values take value 9 (number of ch_codes + 1 -zero based).
  - **time_taken column:** it has been separated into hours and minutes columns and have been represented in a numeric format instead of string format.
  - **dep_time & arr_time columns:** it has been converted from string format into date time format then the column has been separated into hour and minute columns.
  - **Stop column:** string split has been applied to get the number of stops. Then, the non and null valued rows have been converted into zero -numeric value- and 2+ valued rows into 2 numeric value.
  - **Type column:** by using ordinal encoder the values have been encoded and the unknown values take value 2 (number of types + 1 -zero based).
  - **Route column:** the column has been separated into source and destination columns by applying string split. Then, each column has been encoded by using ordinal encoder and unknown values have been handled by taking value 6 (number of source and destination countries + 1 -zero based).

- **Feature selection:** feature correlation has been applied and the airline column has been dropped as it has the same correlation value of ch_code column with the TicketCategory, feature Type has the highest correlation (0.48).



- **Null values:** null values have been handled by giving them value **-1** which gives highest prediction accuracy.
- **Scaling:** data scaling has been applied on the training set. Then, the scaler model has been saved.

Encoders

- Label encoder has been replaced with **ordinal encoder** as it is able to handle unknown values and replace them with specified value.
- By using joblib library the encoders models have been saved.

Used models

1. RidgeClassifier:

a. Accuracy of:
RidgeClassifier(alpha=0.1):0.7104824944839931

2. MLPClassifier:

a. Accuracy of: MLPClassifier():0.8862869988759835
b. Accuracy of: MLPClassifier(activation='relu', hidden_layer_sizes=(150, 100, 50), max_iter=50, solver='lbfgs'):0.7662670163606844
c. Accuracy of: MLPClassifier(activation='tanh', hidden_layer_sizes=(150, 100, 50), max_iter=50,solver='sgd'):0.8349777278214895
d. Accuracy of: MLPClassifier(activation='relu', hidden_layer_sizes=(150, 100, 50), max_iter=50,solver='adam'):0.9352233462387078
e. Accuracy of: MLPClassifier(activation='tanh', hidden_layer_sizes=(150, 100, 50), max_iter=50,solver='adam'):0.9402397901835894
f. Accuracy of: MLPClassifier(activation='tanh', hidden_layer_sizes=(150, 100, 50),max_iter=100,solver='adam'):0.944444444444444 (take > 10 mins)

3. DecisionTreeClassifie:

a. Accuracy of: DecisionTreeClassifier(random_state=42): 0.959056658756921
b. Accuracy of: DecisionTreeClassifier(max_depth=20, random_state=42): 0.9547895591357562
c. Accuracy of: DecisionTreeClassifier(max_depth=50, random_state=42): 0.959056658756921

4. KNeighborsClassifier:

a. Accuracy of: KNeighborsClassifier():0.8935098455518088

5. **LogisticRegression:**
    a. Accuracy of: LogisticRegression(multi_class='multinomial', max_iter=100) :0.7278423046500978
    b. Accuracy of: LogisticRegression(multi_class='multinomial', max_iter=50) :0.715186711627326
    c. Accuracy of: LogisticRegression(C=0.01, solver='liblinear'): 0.689480038299821
    d. Accuracy of: LogisticRegression(C=0.1, solver='lbfgs'):0.7187044669247742
6. **RandomForestClassifier:**
    a. RandomForestClassifier(n_jobs=-1, random_state=42): **0.9621581116523042** (Best)

Accuracy of: RandomForestClassifier(n_jobs=-1, random_state=42):0.9621581116523042

	precision	recall	f1-score	support
0.0	0.94	0.95	0.95	13048
1.0	0.95	0.96	0.95	4405
2.0	0.97	0.96	0.96	20847
3.0	0.99	0.98	0.98	9742
accuracy			0.96	48042
macro avg	0.96	0.96	0.96	48042
weighted avg	0.96	0.96	0.96	48042

All the used models have been saved using libjob library.

Total training time: 3.6 minutes.

Total test time: 23.1123 seconds.

Conclusion: RandomForestClassifier has the highest accuracy **0.96**. Ordinal encoding has been used as it is better than label encoding, it has the ability to handle

unknown values. Missing values have been given value **-1** as it gives highest accuracy.