Travel Insurance Analysis

Machine Learning Project

Coding and Big Data Course

Class IT-4 2020

Group members name:

Ervino Alifio Ramadhan – 001202000133

Markus Raja Sinabutar – 001202000038

Mulya Fajar Ningsih Alwi – 001202000101

Rafli Ersandy – 001202000111

Samuel Pandohan Terampil Gultom – 001202000095

Samuel Pandohan Terampil Gultom – 001202000095

Pointers

Presentation Highlights

1. **Motives**

   What are we trying to solve?
2. **Resources**

   What algorithm are we using?
3. **Demonstration**

   Demonstrating our model and its accuracy
4. **Practicality**

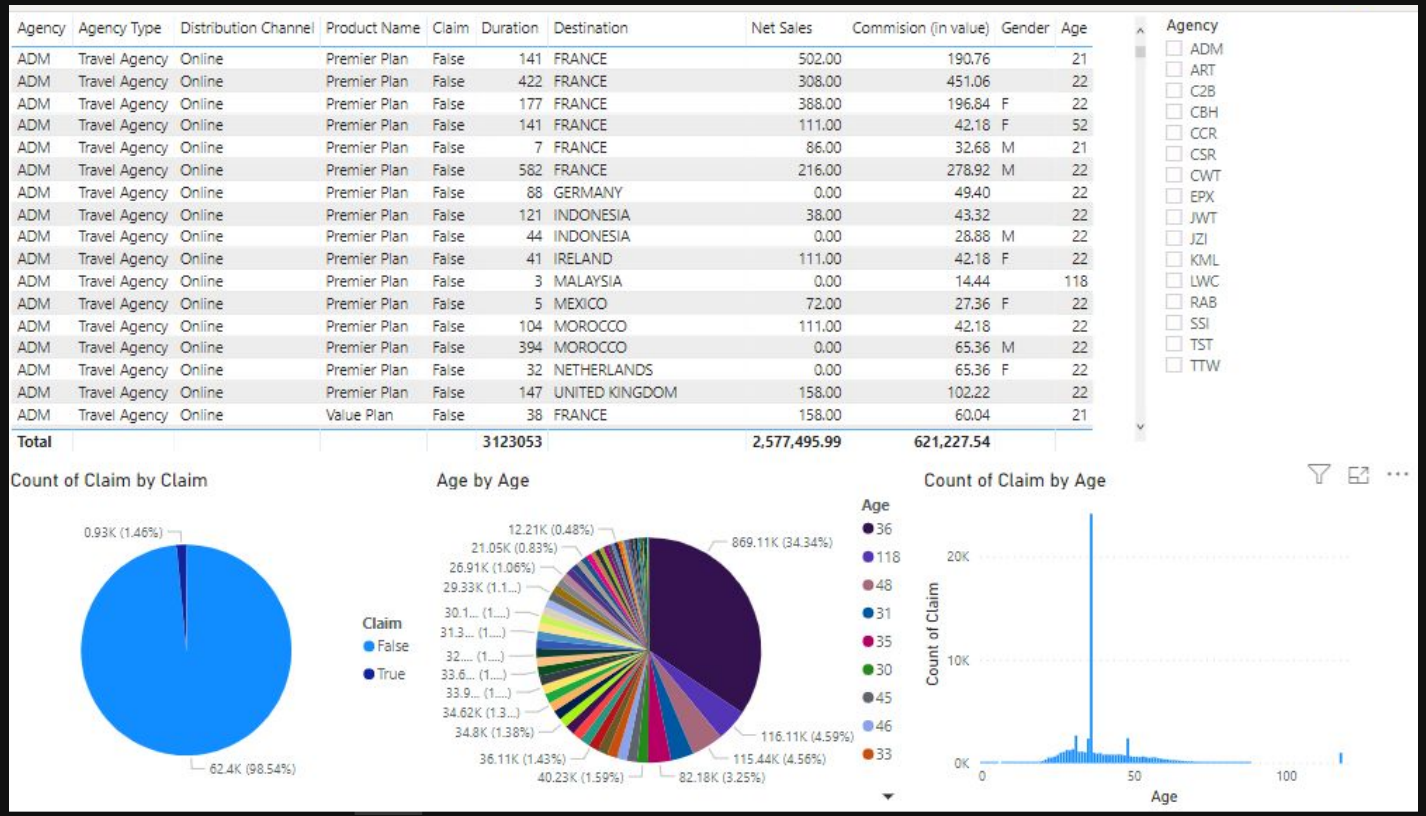
   How can it help people?

Our dataset

For our dataset we use a travel insurance data in Kaggle, below is the link to our dataset:

https://www.kaggle.com/mhdzahier/travel-insurance

Visualization

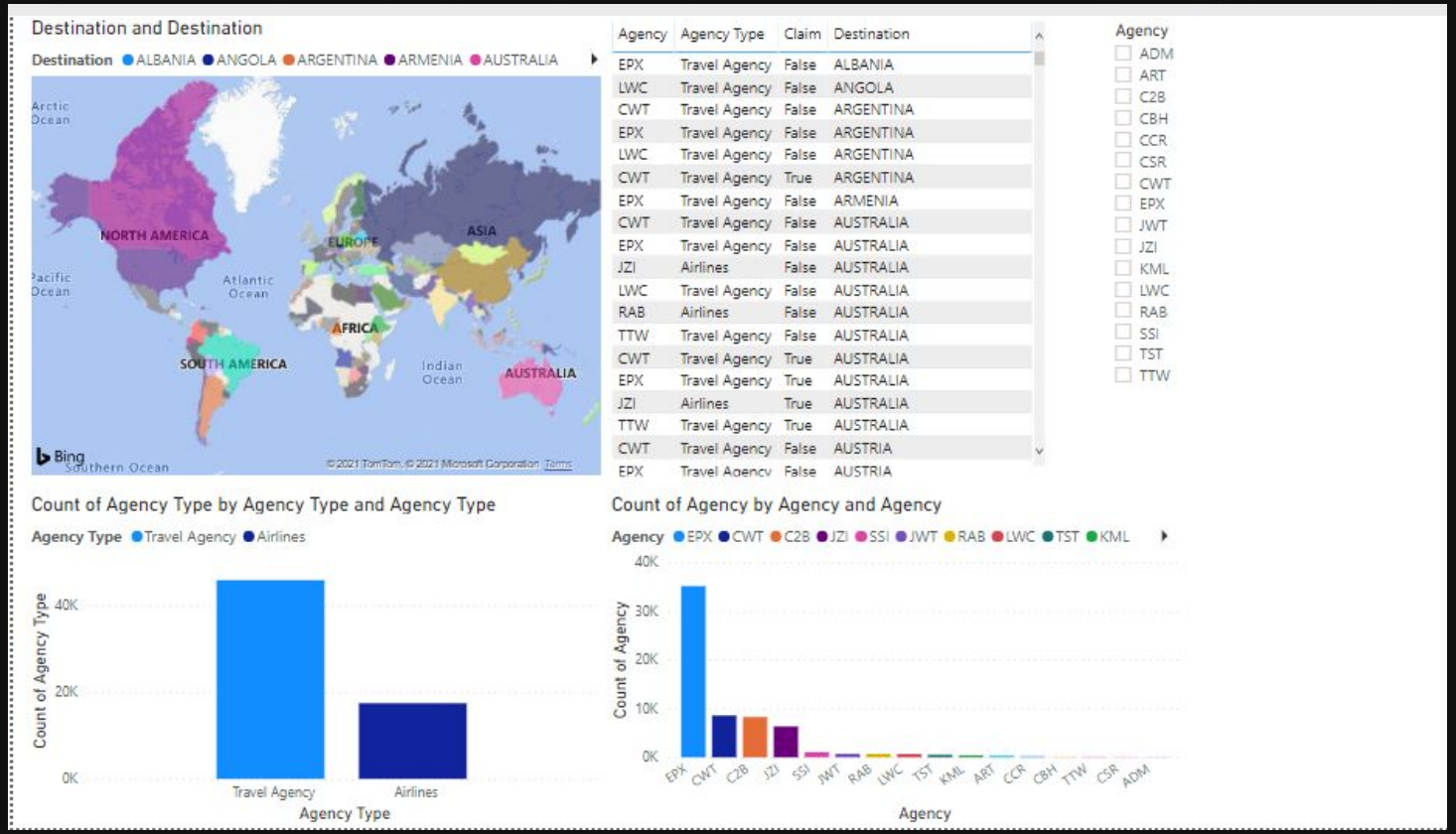


Agency	Agency Type	Distribution Channel	Product Name	Claim	Duration	Destination	Net Sales	Commision (in value)	Gender	Age
ADM	Travel Agency	Online	Premier Plan	False	141	FRANCE	502.00	190.76		21
ADM	Travel Agency	Online	Premier Plan	False	422	FRANCE	308.00	451.06		22
ADM	Travel Agency	Online	Premier Plan	False	177	FRANCE	388.00	196.84	F	22
ADM	Travel Agency	Online	Premier Plan	False	141	FRANCE	111.00	42.18	F	52
ADM	Travel Agency	Online	Premier Plan	False	7	FRANCE	86.00	32.68	M	21
ADM	Travel Agency	Online	Premier Plan	False	582	FRANCE	216.00	278.92	M	22
ADM	Travel Agency	Online	Premier Plan	False	88	GERMANY	0.00	49.40		22
ADM	Travel Agency	Online	Premier Plan	False	121	INDONESIA	38.00	43.32		22
ADM	Travel Agency	Online	Premier Plan	False	44	INDONESIA	0.00	28.88	M	22
ADM	Travel Agency	Online	Premier Plan	False	41	IRELAND	111.00	42.18	F	22
ADM	Travel Agency	Online	Premier Plan	False	3	MALAYSIA	0.00	14.44		118
ADM	Travel Agency	Online	Premier Plan	False	5	MEXICO	72.00	27.36	F	22
ADM	Travel Agency	Online	Premier Plan	False	104	MOROCCO	111.00	42.18		22
ADM	Travel Agency	Online	Premier Plan	False	394	MOROCCO	0.00	65.36	M	22
ADM	Travel Agency	Online	Premier Plan	False	32	NETHERLANDS	0.00	65.36	F	22
ADM	Travel Agency	Online	Premier Plan	False	147	UNITED KINGDOM	158.00	102.22		22
ADM	Travel Agency	Online	Value Plan	False	38	FRANCE	158.00	60.04		21
Total					**3123053**		**2,577,495.99**	**621,227.54**		

General Visualization by Claim

https://app.powerbi.com/groups/me/reports/1b979bd5-bede-484c-be1c-51209b71c347?ctid=24959766-5c6f-4228-b658-2eaabf9d7581

Visualization



Specified Visualization by Agency

https://app.powerbi.com/groups/me/reports/1b979bd5-bede-484c-be1c-51209b71c347?ctid=24959766-5c6f-4228-b658-2eaabf9d7581

What are we trying to solve?

From this dataset we want to know how much travel insurance is used for traveling for travelers and we also want to find out whether someone will use travel insurance or not based on several factors or inputs that given.

What algorithm are we using?

Because we want to solve the problem between yes or not. So, we are using decision tree for our algorithm.

```
[ ]  from sklearn import tree
     clf = tree.DecisionTreeClassifier()
```

What algorithm are we using?

```
[ ]  #Importing Dataset
     import numpy as np
     import pandas as pd

     #Setting Dataset to Variable
     df=pd.read_csv('travel insurance.csv')

     #Previewing Datasets
     df
```

	Agency	Agency Type	Distribution Channel	Product Name	Claim	Duration	Destination	Net Sales	Commision (in value)	Gender	Age
0	CBH	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	81
1	CBH	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	71
2	CWT	Travel Agency	Online	Rental Vehicle Excess Insurance	No	65	AUSTRALIA	-49.5	29.70	NaN	32
3	CWT	Travel Agency	Online	Rental Vehicle Excess Insurance	No	60	AUSTRALIA	-39.6	23.76	NaN	32
4	CWT	Travel Agency	Online	Rental Vehicle Excess Insurance	No	79	ITALY	-19.8	11.88	NaN	41
...	...	...	...	...	...	...	...	...	...	...	...
63321	JZI	Airlines	Online	Basic Plan	No	111	JAPAN	35.0	12.25	M	31
63322	JZI	Airlines	Online	Basic Plan	No	58	CHINA	40.0	14.00	F	40
63323	JZI	Airlines	Online	Basic Plan	No	2	MALAYSIA	18.0	6.30	M	57
63324	JZI	Airlines	Online	Basic Plan	No	3	VIET NAM	18.0	6.30	M	63
63325	JZI	Airlines	Online	Basic Plan	No	22	HONG KONG	26.0	9.10	F	35

63326 rows × 11 columns

What algorithm are we using?

```
[ ]  #Previewing NaN Values
     df.isnull().sum().any
```

```
     <bound method Series.any of Agency                    0
     Agency Type                0
     Distribution Channel       0
     Product Name               0
     Claim                      0
     Duration                   0
     Destination                0
     Net Sales                  0
     Commision (in value)       0
     Gender                 45107
     Age                        0
     dtype: int64>
```

What algorithm are we using?

```
[ ]  #Deleting NaN Values
     df = df.dropna()

     #Previewing Datasets after Deleting NaN Values
     df
```

	Agency	Agency Type	Distribution Channel	Product Name	Claim	Duration	Destination	Net Sales	Commision (in value)	Gender	Age
0	CBH	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	81
1	CBH	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	71
5	JZI	Airlines	Online	Value Plan	No	66	UNITED STATES	-121.0	42.35	F	44
11	JZI	Airlines	Online	Basic Plan	No	1	MALAYSIA	-18.0	6.30	M	47
12	KML	Travel Agency	Online	Premier Plan	No	53	NORWAY	-130.0	49.40	F	48
...	...	...	...	...	...	...	...	...	...	...	...
63321	JZI	Airlines	Online	Basic Plan	No	111	JAPAN	35.0	12.25	M	31
63322	JZI	Airlines	Online	Basic Plan	No	58	CHINA	40.0	14.00	F	40
63323	JZI	Airlines	Online	Basic Plan	No	2	MALAYSIA	18.0	6.30	M	57
63324	JZI	Airlines	Online	Basic Plan	No	3	VIET NAM	18.0	6.30	M	63
63325	JZI	Airlines	Online	Basic Plan	No	22	HONG KONG	26.0	9.10	F	35

18219 rows × 11 columns

What algorithm are we using?

```
[ ]  inputs = df.drop("Claim", axis = "columns")
     target1 = df.drop("Agency", axis = "columns")
```

```
[ ]  from sklearn.preprocessing import LabelEncoder
     le_agency = LabelEncoder()
     le_agencytype = LabelEncoder()
     le_distchan  = LabelEncoder()
     le_destination = LabelEncoder()
     le_gender  = LabelEncoder()
     le_claim   = LabelEncoder()
```

```
[ ]  inputs["agency_n"] = le_agency.fit_transform(inputs["Agency"])
     inputs["agencytype_n"] = le_agencytype.fit_transform(inputs["Agency Type"])
     inputs["distchan_n"] = le_distchan.fit_transform(inputs["Distribution Channel"])
     inputs["destination_n"] = le_destination.fit_transform(inputs["Destination"])
     inputs["gender_n"] = le_gender.fit_transform(inputs["Gender"])
     target1["claim_n"] = le_claim.fit_transform(target1["Claim"])
```

What algorithm are we using?

```
[ ] inputs
```

	Agency	Agency Type	Distribution Channel	Product Name	Duration	Destination	Net Sales	Commision (in value)	Gender	Age	agency_n	agencytype_n	distchan_n	destination_n	gender_n
0	CBH	Travel Agency	Offline	Comprehensive Plan	186	MALAYSIA	-29.0	9.57	F	81	3	1	0	43	0
1	CBH	Travel Agency	Offline	Comprehensive Plan	186	MALAYSIA	-29.0	9.57	F	71	3	1	0	43	0
5	JZI	Airlines	Online	Value Plan	66	UNITED STATES	-121.0	42.35	F	44	8	0	1	79	0
11	JZI	Airlines	Online	Basic Plan	1	MALAYSIA	-18.0	6.30	M	47	8	0	1	43	1
12	KML	Travel Agency	Online	Premier Plan	53	NORWAY	-130.0	49.40	F	48	9	1	1	56	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
63321	JZI	Airlines	Online	Basic Plan	111	JAPAN	35.0	12.25	M	31	8	0	1	36	1
63322	JZI	Airlines	Online	Basic Plan	58	CHINA	40.0	14.00	F	40	8	0	1	15	0
63323	JZI	Airlines	Online	Basic Plan	2	MALAYSIA	18.0	6.30	M	57	8	0	1	43	1
63324	JZI	Airlines	Online	Basic Plan	3	VIET NAM	18.0	6.30	M	63	8	0	1	81	1
63325	JZI	Airlines	Online	Basic Plan	22	HONG KONG	26.0	9.10	F	35	8	0	1	28	0

18219 rows × 15 columns

What algorithm are we using?

```
[ ] target1
```

	Agency Type	Distribution Channel	Product Name	Claim	Duration	Destination	Net Sales	Commision (in value)	Gender	Age	claim_n
0	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	81	0
1	Travel Agency	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	71	0
5	Airlines	Online	Value Plan	No	66	UNITED STATES	-121.0	42.35	F	44	0
11	Airlines	Online	Basic Plan	No	1	MALAYSIA	-18.0	6.30	M	47	0
12	Travel Agency	Online	Premier Plan	No	53	NORWAY	-130.0	49.40	F	48	0
...	...	...	...	...	...	...	...	...	...	...	...
63321	Airlines	Online	Basic Plan	No	111	JAPAN	35.0	12.25	M	31	0
63322	Airlines	Online	Basic Plan	No	58	CHINA	40.0	14.00	F	40	0
63323	Airlines	Online	Basic Plan	No	2	MALAYSIA	18.0	6.30	M	57	0
63324	Airlines	Online	Basic Plan	No	3	VIET NAM	18.0	6.30	M	63	0
63325	Airlines	Online	Basic Plan	No	22	HONG KONG	26.0	9.10	F	35	0

18219 rows × 11 columns

What algorithm are we using?

```
[ ] inputs_n = inputs.drop(["Agency", "Agency Type", "Distribution Channel", "Destination", "Gender", "Product Name", "Commision (in value)"], axis="columns")
    target_n = target1["claim_n"]
```

```
[ ] inputs_n
```

	Duration	Net Sales	Age	agency_n	agencytype_n	distchan_n	destination_n	gender_n
0	186	-29.0	81	3	1	0	43	0
1	186	-29.0	71	3	1	0	43	0
5	66	-121.0	44	8	0	1	79	0
11	1	-18.0	47	8	0	1	43	1
12	53	-130.0	48	9	1	1	56	0
...	...	...	...	...	...	...	...	...
63321	111	35.0	31	8	0	1	36	1
63322	58	40.0	40	8	0	1	15	0
63323	2	18.0	57	8	0	1	43	1
63324	3	18.0	63	8	0	1	81	1
63325	22	26.0	35	8	0	1	28	0

18219 rows × 8 columns

What algorithm are we using?

```
[ ]  target_n
```

```
0        0
1        0
5        0
11       0
12       0
        ..
63321    0
63322    0
63323    0
63324    0
63325    0
Name: claim_n, Length: 18219, dtype: int64
```

What algorithm are we using?

```
[ ] from sklearn import tree
    clf = tree.DecisionTreeClassifier()
```

```
[ ] clf.fit(inputs_n, target_n)
```

```
    DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                           max_depth=None, max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, presort='deprecated',
                           random_state=None, splitter='best')
```

```
[ ] clf.score(inputs_n, target_n)
```

```
    0.9956089796366431
```

```
[ ] #12 = duration real, 30 = net sales real, 35 = age real, 8 = agency_n (JZI),
    #0 = agencytype_n (Airlines), 1 = distchan_n (online), 81 = destination_n (vietnam), 1 = gender_n (male)
    clf.predict([[12,30,35,8,0,1,81,1]])
```

```
    array([1])
```

```
[ ] #186 = duration real, -29.0 = net sales real, 81 = age real, 3 = agency_n (CBH),
    #1 = agencytype_n (Travel Agency), 0 = distchan_n (offline), 43 = destination_n (MALAYSIA), 0 = gender_n (female)
    clf.predict([[186,-29.0,81,3,1,0,43,0]])
```

```
    array([0])
```

Demonstrating our model

The demo is available at the link below:

https://colab.research.google.com/drive/1K8MoVwy46lHajb4D65FXCykVVIFECsrr?authuser=2#scrollTo=pGq3g4RmFfnr

Model accuracy score

```
[ ]  clf.score(inputs_n, target_n)
```

```
0.9956089796366431
```

From the calculation of the accuracy score above, it can be seen that the accuracy of the model is 99% and it can be said that the model is quite accurate.

How can it help people?

From this dataset and our predictions, we help travelers to make a decision whether they need to use travel insurance for their trip or not based on some of the factors or inputs that the traveler has.