

```
from google.colab import files
uploaded = files.upload()
```

יש לבחור קבצים TravellInsur...ediction.csv
TravellInsurancePrediction.csv(text/csv) - 115409 bytes, last modified: 29.12.2025 - 100% done
Saving TravelInsurancePrediction.csv to TravelInsurancePrediction (1).csv

```
import pandas as pd
df = pd.read_csv('TravelInsurancePrediction.csv')
df.head()
```

	Unnamed: 0	Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad
0	0	31	Government Sector	Yes	400000	6	1	No	No
1	1	31	Private Sector/Self Employed	Yes	1250000	7	0	No	No
			Private						

Next steps: [Generate code with df](#) [New interactive sheet](#)

חלק 1 – בחירת DATA וטיעינה

בדאטה זו כל שורה מייצגת אדם פוטנציאלי שנבדק האם רכש ביטוח נסיעות. הנתונים כוללים מאפיינים דמוגרפיים והתנהגותיים כגון גיל, סוג תעסוקה, הכנסה שנתית, מספר בני משפחה, מצב בריאותי והרגלי טישה.

המשתנה **TravelInsurance** הוא העמודה (Target) המשתנה המנובא

- ערך 1 מציין שהאדם רכש ביטוח נסיעות
- ערך 0 מציין שהאדם לא רכש ביטוח נסיעות

וממשאות לחיזוי ההחלטה האם לרכוש ביטוח Features-שאר העמודות משמשות כ

שאלות תיאוריות – חלק 1

מהו למידה מונחית (Supervised Learning)?

המטרה (Target) וגם תווית ידועה מראש (Features) למידה מונחית היא סוג של למידה מכונה שבה המודל מואמן על דатаה שמקיל גם נתונים קלט. היא לומוד קשר בין הנתונים לבון התווית כדי לבצע חיזוי על נתונים חדשים.

מדוע נדרש>Data עם תוויות?

התוויות מאפשרות למודל לדעת מהי התשובה הנכונה עבור כל דוגמה בDATA. ללא תוויות, המודל לא יוכל למדוד שגיאה או לומוד כיצד לשפר את עצמו.

מדוע חשוב להבין את ההבדל בין Feature-7-Target?

הוא המשטנה שאותו מנוטים לחזות, Target הוא משטנה המספק מידע שעשו להשפיע על החיזוי (כגון גיל או הכנסה שנתית), ואילו Feature במקורה זה האם האדם רכש ביטוח נסיעות.

מדוע חשוב להבין את ההבדל בין Feature-7-Target?

הבנייה החקלאית מאפשרת לבחור עיבוד נתונים מותאים, לזרות משטנים לא רלוונטיים ולפרש את תוצאות המודל בקרה נכונה.

Double-click (or enter) to edit

חלק 2 – ניתוח נתונים קרני (EDA)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1987 entries, 0 to 1986
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Unnamed: 0        1987 non-null   int64  
 1   Age              1987 non-null   int64  
 2   Employment Type  1987 non-null   object  
 3   GraduateOrNot    1987 non-null   object  
 ...   ...             ...           ...    

```

```

4 AnnualIncome      1987 non-null    int64
5 FamilyMembers     1987 non-null    int64
6 ChronicDiseases   1987 non-null    int64
7 FrequentFlyer     1987 non-null    object
8 EverTravelledAbroad 1987 non-null    object
9 TravelInsurance    1987 non-null    int64
dtypes: int64(6), object(4)
memory usage: 155.4+ KB

```

▼ ניתוח מבנה הדadata

בדקתי את מבנה הדadata, סוג המשתנים והאם קיימים ערכים חסרים. ניתן לראות כי הדadata כולל שילוב של משתנים () באמצעות הפקהה נומריים ומשתנים קטגוריאליים, דבר שידרש טיפול שונה בכל סוג

```
df.describe(include='all')
```

	Unnamed: 0	Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	Eve
count	1987.000000	1987.000000		1987	1987	1.987000e+03	1987.000000	1987.000000	1987
unique	Nan	Nan		2		NaN	Nan	Nan	2
top	Nan	Nan	Private Sector/Self Employed		Yes	NaN	Nan	Nan	No
freq	Nan	Nan	1417		1692	NaN	NaN	NaN	1570
mean	993.000000	29.650226		NaN	9.327630e+05	4.752894	0.277806	NaN	
std	573.741812	2.913308		NaN	3.768557e+05	1.609650	0.448030	NaN	
min	0.000000	25.000000		NaN	3.000000e+05	2.000000	0.000000	NaN	
25%	496.500000	28.000000		NaN	6.000000e+05	4.000000	0.000000	NaN	
50%	993.000000	29.000000		NaN	9.000000e+05	5.000000	0.000000	NaN	
75%	1489.500000	32.000000		NaN	1.250000e+06	6.000000	1.000000	NaN	

▼ ניתוח סטטיסטי ראשון

אפשרת לקבל תמונה כללית על הנתונים, כולל ערכים ממויצעים, טווחים וערכים שכיחים. שלב זה מסיע בזיהוי חיריגות () באמצעות הפקהה. והבנת התפלגות הכתנים בדadata.

```
df['TravelInsurance'].value_counts()
```

TravelInsurance	count
0	1277
1	710

```
dtype: int64
```

```
df['TravelInsurance'].value_counts(normalize=True)
```

TravelInsurance	proportion
0	0.642677
1	0.357323

```
dtype: float64
```

בדיקת איזון המחלקות

בדיה זו חשוב כדי להבין האם קיימת בעית חוסר איזון בין מחלקות, שעלולה להשפיע. בדקתי את התפלגות המשתנה המנובא על ביצועי המודל והערכתנו.

EDA – שאלות תיאורטיות

הוא שלב קritisי לפני אימון מודל EDA מודיען?

אפשר להבין את מבנה הדאטה, לזהות בעיות כמו ערכים חסרים, חסר איזון בין משתנים או משתנים חריגים, וכך לבחור תהליך שימוש מתאים לפני אימון המודל.

EDA אילו בעיות עלולות להיווצר אם מודלים על?

או פרשנות לא נכון של הביצועים Overfitting, עשוי לגרום לאימון מודל על דאטה בעיתוי, מה שיוביל לתוצאות שגויות EDA דלג על

מהי בעיתות חסר איזון בין מחלקות ולמה היא חשובה?

חסר איזון מתרחש כאשר מחלוקת אחת מופיעה הרבה יותר מאשר אחרות. במצב כזה המודל עלול להעדיף את המחלוקת הדומיננטית ולספק תוצאות לא אמינות.

(חלק 3 – ניתוח ערכים חסרים)

```
df.isnull().sum()
```

	0
Unnamed: 0	0
Age	0
Employment Type	0
GraduateOrNot	0
AnnualIncome	0
FamilyMembers	0
ChronicDiseases	0
FrequentFlyer	0
EverTravelledAbroad	0
TravellInsurance	0

```
dtype: int64
```

בדיקות ערכים חסרים בדאטה

בשלב זה בדקתי האם קיימים ערכים חסרים בדאטה ובאיזה עמודות הם מופיעים. זיהוי ערכים חסרים הוא שלב חשוב לפני אימון מודל, מאחר ורוב מודלי למידת המכונה אינם יודעים להתמודד עם ערכים חסרים.

אסטרטגיית טיפול בערכים חסרים

במשתנים נומריים עדיף להשלים ערכים חסרים באמצעות החזירן, מאחר והוא פחת מושפע מערכים חריגים.

כדי לשמר את התפלגות הדאטה (Most Frequent) במשתנים קטגוריאליים עדיף להשלים באמצעות הערך השכיח.

בחירה זו מאפשרת לשמר על כמות הנתונים ולמנוע אי-בודד מיידע שעלול להיגרם מכחיקת שורות.

שאלות תיאורטיות – ערכים חסרים

מודע דאטה אמיתי מכל ערכים חסרים?

ערכים חסרים יכולים להיווצר עקב טיעיות האזנה, נתונים שלא נאספו, או חסר מענה של משתמשים. בדאטה מהעולם האמיתי זו תופעה נפוצה.

Imputation מה?

היא טכניקה להשלמת ערכים חסרים באמצעות ערכים מחושבים כגון ממוצע, חציון או ערך שכיח Imputation

מודע מחייב שורות היא לרוב רענון גרען?

מחיקת שורות גורמת לאי-בודד מיידע ועלולה להקטין משמעותית את גודל הדאטה או לפגוע בההתפלגות הנתונים.

מתי נוריד עמודה ומתי נשלים ערכים חסרים?

אם עמודה מכילה אחוז גבוה מאוד של ערכים חסרים או שאינה רלוונטית לבעה, ניתן להסיר אותה. במקרים אחרים עדיף להשלים את הערכים בהתאם למאפייניהן Imputation באמצעות שיטות

מה הרעיון בחלק הזה? (בשפה פשוטה) Features I-Target – הפרדת 🔗

אנחנו אומרים למודל:

✗ זה מה שאתה צריך לנחש (Target)

זהה המידע שモותר לך להשתמש בו (Features)

```
df = df.drop(columns=['Unnamed: 0'])
```

```
X = df.drop('TravelInsurance', axis=1)
y = df['TravelInsurance']
```

▼ Features I-Target חלק 4 – הפרדת

בשלב זה הפרדתי את הדטה לשני רכיבים מרכזיים:

- Features כל המשתנים המשמשים כקלט למודל – (X)
- Target המשתנה שאוטו המודל צריך לחזות – (y)

אילו שאר העמודות משמשות במידע מסביר לצורכי החיזוי. הפרדה זו היא שלב קריטי לפני TravellInsurance בפרויקט זה המשתנה המנובא הוא אימון מודל מידת מכונה.

▼ שאלות תיאוריות – חלק 4

7- מודיען Features כובוה להפריד בין?

במהלך האימון (Data Leakage) הפרדה זו מונעת מצב שבו המודל נחשף לערך שאוטו הוא אמור לחשות, וכן נמנעת זיגית מידע.

מה המשמעות של "למידה" במודול?

בצורה מיטבית Target-למודול מתקיים שבו המודל מתאים את הפרמטרים הפנימיים שלו כך שיוכל למפות בין ה-

אילו עמודות הוסרו ואילו הושארו? מודיען?

מאחר והן מכילות מידע רלוונטי-Features- מאחר והוא המשתנה המנובא. שאר העמודות נשארו כ-Target-הפרדה C עם תוויות ידועות מראש, ואילו ב- Supervised Learning.

Supervised Learning-7-Unsupervised Learning מה ההבדל בין?

אין תוויות והמטרה היא לאלהות מבנים (Clusters) על דאטה עם תוויות ידועות מראש, ואילו ב- Unsupervised Learning. ודף סיום בדאטה.

שלב 5.1 – חלוקה הדטה Train / Test חלק 5 – חלוקה

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42,
    stratify=y
)
```

7-Train / Test – חלוקה הדטה

המשמש להערכת ביצועי המודול על נתונים (Test) המשמש לאימון המודול, וסתם בדיקה (Train) בשלב זה חילקתי את הדטה לשני חלקים: סט אימון (Train) וחישומים (Test).

החלוקת בוצעה בצורה סטרטיפית לפי המשתנה המנובא, כדי לשמור על יחס דומה בין המחלקות בשני הסטים.

▼ שאלות תיאוריות – Train / Test Split

מודיען אסור להעירך מודול על דאטה שאיתו הוא אומן?

הערכתה על דאטה שאיתו המודול אומן תיתן תוצאות אופטימיות מדי ולא תשקף את יכולת ההכללה של המודול על נתונים חדשים.

Overfitting מהו?

הוא מצב שבו המודול מתקיים היטב על המידה לדאטה, כולל רעש וחריגות, וכך מבצע חיזוי גרוע על נתונים חדשים.

Data Leakage מה?

מתרכח כאשר מידע מסוים הבודקה דלפ' לתהיליך האימון, וגורם להערכת ביצועים לא אמיןיה.

מודיען חלוקה סטרטיפית חשובה בעקבות סיווג?

חלוקת סטרטיפית שומרת על יחס דומה בין המחלקות ומונעת מצב שבו אחת המחלקות מייצגת בצורה לא הוגנת בסט הבדיקה.

בנין Pipeline I-ColumnTransformer

```
numeric_features = X.select_dtypes(include=['int64', 'float64']).columns
categorical_features = X.select_dtypes(include=['object']).columns

print("Numeric features:", numeric_features)
print("Categorical features:", categorical_features)
```

```
Numeric features: Index(['Age', 'AnnualIncome', 'FamilyMembers', 'ChronicDiseases'], dtype='object')
Categorical features: Index(['Employment Type', 'GraduateOrNot', 'FrequentFlyer', 'EverTravelledAbroad'], dtype='object')
```

```
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
```

```
numeric_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

categorical_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])
```

```
from sklearn.compose import ColumnTransformer

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_pipeline, numeric_features),
        ('cat', categorical_pipeline, categorical_features)
    ]
)
```

```
X_train_processed = preprocessor.fit_transform(X_train)
X_test_processed = preprocessor.transform(X_test)
```

לעיבוד נתונים Pipeline בנית

לעיבוד הנתונים לפני אימון המודלים. המשתנים הנוומרים עברו השלמת ערכים חסרים באמצעות חיזיון ולאחר מכן Pipeline בשלב זה נבנה One-Hot Encoding.

בוצע רק על סט האימון כדי למנוע זליגת מידע fit-כאשר תהליך ההעיבוד בוצע באמצעות ColumnTransformer, כל שלבי הפעולה בוצעו באמצעות

– שאלות תיאוריות

מודלים מודולים דרושים קלט נומי בלבד?

מרבית מודלי למידת המכונה מבצעים חישובים מתמטיים, ולכןם יודעים לעבוד ישירות עם טקסט או קטגוריות.

חשיבות רף לחלק ממהמודלים fit מודול מבצעים?

מודלים המבוססים על מוחקים או גרדיאנטים רגשים לסקאלה הנתונים, בעוד שמודלים מבוססי עצים פחות מושפעים מכך.

רף על נתונים האימון fit מודול מבצעים?

על כל הדאטה עלול לגרום לזליגת מידע מטט הבדיקה לתהליכי האימון ולפגוע באיכות הערכת המודל fit ביצוע.

1 Logistic Regression 2 KNN 3 Decision Tree

```
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
```

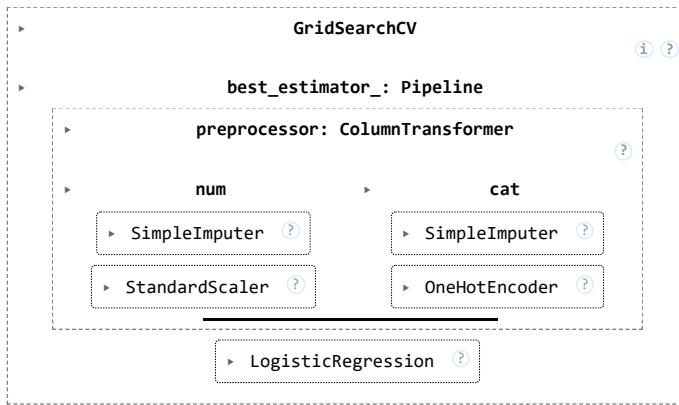
```
log_reg = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('model', LogisticRegression(max_iter=1000))
```

```

])
log_reg_params = {
    'model_C': [0.01, 0.1, 1, 10]
}

log_reg_grid = GridSearchCV(
    log_reg,
    log_reg_params,
    cv=5,
    scoring='accuracy'
)
log_reg_grid.fit(X_train, y_train)

```

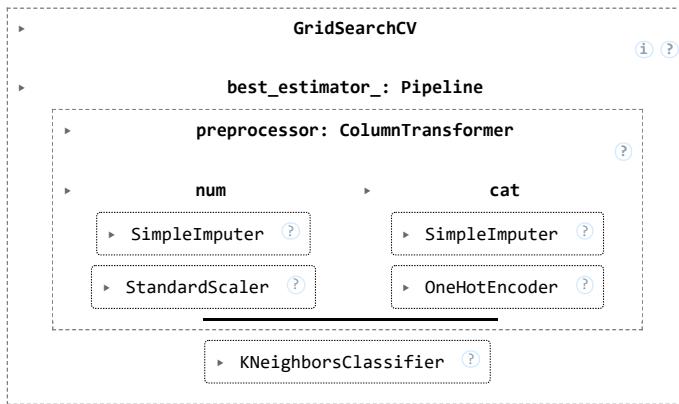


```

knn = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('model', KNeighborsClassifier())
])
knn_params = {
    'model_n_neighbors': [3, 5, 7, 9]
}

knn_grid = GridSearchCV(
    knn,
    knn_params,
    cv=5,
    scoring='accuracy'
)
knn_grid.fit(X_train, y_train)

```



```

tree = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('model', DecisionTreeClassifier(random_state=42))
])
tree_params = {
    'model_max_depth': [None, 5, 10, 20],
    'model_min_samples_split': [2, 5, 10]
}

tree_grid = GridSearchCV(
    tree,

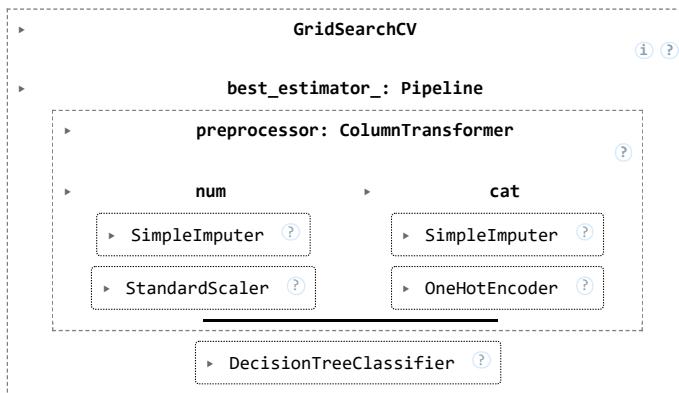
```

```

        tree_params,
        cv=5,
        scoring='accuracy'
    )

tree_grid.fit(X_train, y_train)

```



```

import pandas as pd

results = pd.DataFrame({
    'Model': ['Logistic Regression', 'KNN', 'Decision Tree'],
    'Best CV Accuracy': [
        log_reg_grid.best_score_,
        knn_grid.best_score_,
        tree_grid.best_score_
    ]
})

results

```

	Model	Best CV Accuracy	
0	Logistic Regression	0.782874	
1	KNN	0.802387	
2	Decision Tree	0.835740	

Next steps: [Generate code with results](#) [New interactive sheet](#)

aicmon מודלים והשוואתם

כל מודל נבחרו היפר-פרמטרים GridSearchCV: Logistic Regression, KNN I-Decision Tree. המודלים שנויים באמצעות Cross-Validation עם $cv=5$.

כדי לבחור את המודל בעל הביצועים הטובים ביותר, השוואת המודלים בוצעה על סמך דיקט ממוצע ב-

שאלות תיאוריות – איךון מודלים

"מה פירוש "aicmon מודל"

aicmon מודל הוא תהליך שבו המודל לומד להתאים את הפרמטרים שלו כך שיבצע חיזוי מדויק ככל האפשר על סמך הדאטה.

היפר Hyperparameter?

Hyperparameter שנקבע מראש והוא נלמד ישירות מהדאטה, כגון עומק עץ או מספר שכנים ב-KNN.

מדוע מודלים שונים נתונים תוצאות שונות על אותו דאטה?

כל מודל מבוסס על הנחות ומתמטיקה שונות, ולכן כל אחד>Dפוסים אחרים בדאטה ומציג ביצועים שונים.

חלק 8 – בחירת המודל הטוב ביותר

```
results
```

Model	Best CV Accuracy
0 Logistic Regression	0.782874
1 KNN	0.802387
2 Decision Tree	0.835740

Next steps: [Generate code with results](#) [New interactive sheet](#)

```
best_model_name = results.loc[results['Best CV Accuracy'].idxmax(), 'Model']
best_model_name
```

```
'Decision Tree'
```

```
best_model = max(
    [log_reg_grid, knn_grid, tree_grid],
    key=lambda x: x.best_score_
)
```

בחירה המודל הטוב ביותר

נבחר המודל שהציג את הדיקן הממוצע הגבוה ביותר. בחירה זו מבוססת על ביצועים Cross-Validation לאחר השוואת ביצועי המודלים באמצעות על סט אימון בלבד, ללא שימוש בסט הבדיקה, כדי למנוע הטיה בתהילך הבחירה.

פרק 9 – הערכת המודל על Test Data

```
y_pred = best_model.predict(X_test)
```

```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test, y_pred)
accuracy
```

```
0.8140703517587939
```

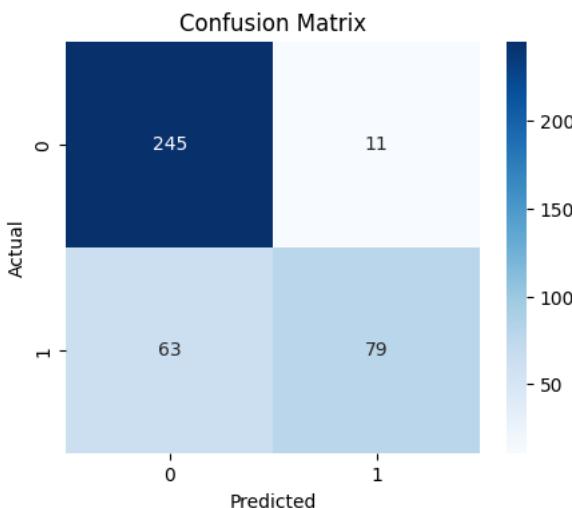
```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
cm
```

```
array([[245,  11],
       [ 63,  79]])
```

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```



```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.80	0.96	0.87	256
1	0.88	0.56	0.68	142
accuracy			0.81	398
macro avg	0.84	0.76	0.77	398
weighted avg	0.82	0.81	0.80	398

הערכת ביצועי המודל על סט הבדיקה

בשלב זה הערכתית את המודל הנבחר על סט הבדיקה, אשר לא שימוש בתהילן האימון. בוצעה הערכה באמצעות מדדי Accuracy, Confusion Matrix Precision, Recall ומדד F1.

שילוב ממדדים אלו מאפשר להבין לא רק את רמת הדיווק הכללית, אלא גם את סוג השגיאות שהמודל מבצע.

שאלות תיאוריות – הערכת מודל

בלבד אינו מספיק מדויק?

אינו מבדיל בין סוגים של שגיאות ועלול להטעות במיוחד במקרים בהם קיימת חוסר איזון בין מחלקות Accuracy.

מה הם False Positive ו-False Negative?

הוא מצב שבו המודל חזה שלילי אך הערך האמתי False Negative הוא מצב שבו המודל חזה חיובי אך הערך האמתי שלילי, ואילו True Positive.

חישוב יותר מ-Recall-Precision?

חישוב יותר מאשר החמץת מקרים חיובים מהוות בעיה חמורה יותר מאשר דחיות שגוי, לדוגמה במצבים רפואיים או זיהוי סיכון.

פרק 10 – חיזוי על נתונים חדשים

```
new_data = pd.DataFrame([
    {
        'Age': 35,
        'Employment Type': 'Private Sector/Self Employed',
        'GraduateOrNot': 'Yes',
        'AnnualIncome': 600000,
        'FamilyMembers': 3,
        'ChronicDiseases': 0,
        'FrequentFlyer': 'Yes',
        'EverTravelledAbroad': 'Yes'
    },
    {
        'Age': 50,
        'Employment Type': 'Government Sector',
        'GraduateOrNot': 'No',
        'AnnualIncome': 400000,
        'FamilyMembers': 4,
        'ChronicDiseases': 1,
        'FrequentFlyer': 'No',
    }
])
```

```

        'EverTravelledAbroad': 'No'
    },
    {
        'Age': 28,
        'Employment Type': 'Private Sector/Self Employed',
        'GraduateOrNot': 'Yes',
        'AnnualIncome': 800000,
        'FamilyMembers': 1,
        'ChronicDiseases': 0,
        'FrequentFlyer': 'Yes',
        'EverTravelledAbroad': 'Yes'
    }
])

```

```
new_predictions = best_model.predict(new_data)
new_predictions
```

```
array([1, 0, 0])
```

```
new_probabilities = best_model.predict_proba(new_data)
new_probabilities
```

```
array([[0.42857143, 0.57142857],
       [0.8          , 0.2          ],
       [0.87121212, 0.12878788]])
```

חיזוי על נתונים חדשים

בשלב זה יצרתי דוגמאות חדשות המדומות לקווות פוטנציאליים, וביצעת חיזוי באמצעות המודל הנבחר. המודל זהה הן את המחלקה (רכישה / אי-רכישה) והן את ההסתברות לכל תחזית.

שימוש בהסתברויות מאפשר לקבל החלטות מושכלת יותר, ולא רק חיזויBINARIO.

שאלות תיאורטיות – חיזוי על נתונים חדשים

מהו Preprocessor? מודיע חייבים להשתמש באוטו?

כדי להבטיח שהנתונים החדשניים יעברו את אותו תהליך עיבוד כמו נתוני האימון, וכך לשמר על עקבותיו ודיוק בחיזוי.

מחדש נתונים חדשים fit מודיע אסור לבצע?

מחדר ישנה את פרמטרי העיבוד והמודל, ויפגע באמינות החיזוי וביכולת הכללה fit ביצוע.

מה מייצגת הסתברות חיזוי?

הסתברות חיזוי מייצגת את רמת הביטחון של המודל בהשטייכות הדוגמה למחלקה מסוימת.

כיצד משתמשים בביטחון החיזוי בקבלת החלטות?

ניתן לקבוע סף החלטה, או להשתמש בהסתברויות לצורך לקווות, ניהול סיכונים וקבלת החלטות עסקיות.

יאלהה, סיום חלק 11 – רפלקציה מסכמת

רפלקציה מסכמת

מה היה החלק המאתגר ביותר בפרויקט?

מלא שטפפל גם בערכים חסרים, גם בקידוד משתנים קטגוריאליים וגם במניעת זליגת מידע בין סט Pipeline החלק המאתגר ביותר היה בהנחתה. האימון לסט הבדיקה.

אילו טעויות נפוצות קיימות בפרויקט?

הערכת המודל על נתונים שבהם Train/Test-על כל הדטה לפני החלוקה ל preprocessing ביצוע, EDA-טעויות נפוצות כוללות דילוג על שלב ה-אימון.

מודיע איקות הדטה חשובה יותר מבחרית מודל

דטה איקוטי מאפשר לכל מודל ללמידה דפוסים אכטיטיים, בעוד שדטה בעיתוי יכול לתרום גוראות גם אם נשתמש במודל מתקדם.