ບົດຝຶກຫັດ2: Data Preprocessing

ລະຫັດນັກສືກສາ:205N008319

ຊືື່ ແລະ ນາມສະກຸນ: ທ້າວເຊັງວ່າງ ບຼົ່ງໄມ

ຫ້ອງ: 3CW1

1. ຈົ່ງນໍາໃຊ້ຄໍາສັ່ງPython ເພື່ອສະແດງຂໍ້ມຸນຈາກຊຸດຂໍ້ມູນData.csvໃສ່ໃນຕາຕະລາງໃຫ້ສົມບູນ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Country | Age | Salary | Purscased |
|  |  |  |  |  |
| 0 | France | 44.0 | 72000.0 | No |
|  |  |  |  |  |
| 1 | Spain | 27.0 | 48000.0 | Yes |
|  |  |  |  |  |
| 2 | Germany | 30.0 | 54000.0 | No |
|  |  |  |  |  |
| 3 | Spain | 38.0 | 61000.0 | No |
|  |  |  |  |  |
| 4 | Germany | 40.0 | NaN | Yes |
|  |  |  |  |  |
| 5 | France | 35.0 | 58000.0 | Yes |
|  |  |  |  |  |
| 6 | Spain | NaN | 52000.0 | No |
|  |  |  |  |  |
| 7 | France | 48.0 | 79000.0 | Yes |
|  |  |  |  |  |
| 8 | Germany | 50.0 | 83000.0 | No |
|  |  |  |  |  |
| 9 | France | 37.0 | 67000.0 | Yes |
|  |  |  |  |  |

2. ຈົ່ງນຳໃຊ້ຄຳສັ່ງ info()ເພື່ອສະແດງຂໍ້ມຸນໃນຕາຕະລາງໃຫ້ສົມບູນ

|  |  |  |  |
| --- | --- | --- | --- |
|  | Column (Features) | Not-Null Count | Dtype |
|  |  |  |  |
| 0 | Country | 10 not-nul | float64 |
|  |  |  |  |
| 1 | Age | 9 not-nul | float64 |
|  |  |  |  |
| 2 | Salary | 9 not-null | float64 |
|  |  |  |  |
| 3 | Purchased | 10 not-nul | object |
|  |  |  |  |

1. ຈົ່ງກວດສອບຂໍ້ມູນສູນຫາຍ(Missing Values) ດ້ວຍຄຳສັ່ງ isna(), isnull().values.any(),

isnull(),isnull().sum(), isnull().sum().sum(), ແລ້ວລາຍງານວ່າມີຂໍ້ມູນໃດແດ່ສູນຫາຍ

1. ຈົ່ງແກ້ໄຂ້ບັນຫາຂໍ້ມູນສູນຫາຍດ້ວຍການເຮັດPreprocessing ເຊຶ່ງສາມາດນຳໃຊ້ຫຼັກການຕ່າງໆໃນເວັບໄຊຈົວຢ່າງນີໍ້: https://stackabuse.com/python-how-to-handle-missing-dataframe-values-in-pandas/
2. ຈົ່ງໃຊ້ຄຳສັ່ງ Python ເພື່ອແບ່ງຊຸດຂໍ້ມູນData.csvອອກເປັນ2 ຊຸດຄື: X (Features)

ແລະ y (Label)

ຄຳສັ່ງຕົວຢ່າງ:

df= pd.read\_csv('Data.csv')

1. = df.iloc[:, :-1].values y = df.iloc[:, -1].values

X

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Country | |  | Age | | Salary |
|  |  |  |  |  |  |  |  |
| 0 |  |  | France |  | 44.0 | | 72000.0 |
|  |  |  |  |  |  |  |  |
| 1 |  |  | Spain |  | 27.0 | | 48000.0 |
|  |  |  |  |  |  |  |  |
| 2 |  | Germany | |  | 30.0 | | 54000.0 |
|  |  |  |  |  |  |  |  |
| 3 |  |  | Spain |  | 38.0 | | 61000.0 |
|  |  |  |  |  |  |  |  |
| 4 |  | Germany | |  | 40.0 | | NaN |
|  |  |  |  |  |  |  |  |
| 5 |  |  | France |  | 35.0 | | 58000.0 |
|  |  |  |  |  |  |  |  |
| 6 |  |  | Spain |  | NaN | | 52000.0 |
|  |  |  |  |  |  |  |  |
| 7 |  |  | France |  | 48.0 | | 79000.0 |
|  |  |  |  |  |  |  |  |
| 8 |  | Germany | |  | 50.0 | | 83000.0 |
|  |  |  |  |  |  |  |  |
| 9 |  |  | France |  | 37.0 | | 67000.0 |
|  |  |  |  |  |  |  |  |
| y (Label) | |  |  |  |  |  |  |
|  |  |  |  | | |  |  |
|  |  |  | Purscased | | |  |  |
|  | |  |  |  | |  |  |
| 0 | |  |  | No | |  |  |
|  | |  |  |  | |  |  |
| 1 | |  |  | Yes | |  |  |
|  | |  |  |  | |  |  |
| 2 | |  |  | No | |  |  |
|  | |  |  |  | |  |  |
| 3 | |  |  | No | |  |  |
|  | |  |  |  | |  |  |
| 4 | |  |  | Yes | |  |  |
|  | |  |  |  | |  |  |
| 5 | |  |  | Yes | |  |  |
|  | |  |  |  | |  |  |
| 6 | |  |  | No | |  |  |

7 Yes

8 No

9 Yes

1. ຈົ່ງໃຊ້ຄຳສັ່ງPythonເພື່ອແປງຄ່າຂອງXແລະ y ຄຳສັ່ງຕົວຢ່າງ:

from sklearn.compose import ColumnTransformer from sklearn.preprocessing import OneHotEncoder

ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [0])], remainder='passthrough')

X = np.array(ct.fit\_transform(X))

ຄ່າຂອງX ທີື່ແປງແລ້ວຮອບ1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Country | |  |  |  | Age | |  | Salary |
| 0 | 1.0 0.0 0.0 | | |  | 44.0 | | | | 72000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 1 | 0.0 0.0 1.0 | | |  | 27.0 | | | | 48000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 2 | 0.0 1.0 0.0 | | |  | 30.0 | | | | 54000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 3 | 0.0 0.0 1.0 | | |  | 38.0 | | | | 61000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 4 | 0.0 1.0 0.0 | | |  | 40.0 | | | | 63777.77777777778 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 5 | 1.0 0.0 0.0 | | |  | 35.0 | | | | 58000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 6 | 0.0 0.0 1.0 | | |  | 38.77777777777778 | | | | 52000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 7 | 1.0 0.0 0.0 | | |  | 48.0 | | | | 79000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 8 | 0.0 1.0 0.0 | | |  | 50.0 | | | | 83000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
| 9 | 1.0 0.0 0.0 | | |  | 37.0 | | | | 67000.0 | |
|  |  |  |  |  |  |  |  |  |  |  |
|  | | |  |  |  |  |  |  |  |  |
|  | |  | |  |  |  |  |  |  |  |
| from sklearn.preprocessing import StandardScaler | | | | | | | |  |  |  |
| sc = StandardScaler() | | |  |  |  |  |  |  |  |  |
| X\_train[:, 3:] = sc.fit\_transform(X\_train[:, 3:]) | | | | | | |  | |  |  |
| X\_test[:, 3:] = sc.transform(X\_test[:, 3:]) | | | | | |  | | |  |  |
| ຄ່າຂອງX ທີ່ແປງແລ້ວຮອບ2 | | | |  |  |  |  |  |  |  |
|  |  | Country | |  |  |  | Age | |  | Salary |
|  |  | | |  | | | | |  |  |
| 0 | 1.0 0.0 0.0 | | | -0.19159184384578545 | | | | |  | -1.0781259408412425 |
|  |  | | |  | | | | |  |  |
| 1 | 0.0 0.0 1.0 | | | -0.014117293757057777 | | | | |  | -0.07013167641635372 |
|  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 2 | 0.0 1.0 0.0 | 0.566708506533324 | 0.633562432710455 |
|  |  |  |  |
| 3 | 0.0 0.0 1.0 | -0.30453019390224867 | -0.30786617274297867 |
|  |  |  |  |
| 4 | 0.0 1.0 0.0 | -1.9018011447007988 | -1.420463615551582 |
|  |  |  |  |
| 5 | 1.0 0.0 0.0 | 1.1475343068237058 | 1.232653363453549 |
|  |  |  |  |
| 6 | 0.0 0.0 1.0 | 1.4379472069688968 | 1.5749910381638885 |
|  |  |  |  |
| 7 | 1.0 0.0 0.0 | -0.7401495441200351 | -0.5646194287757332 |
|  |  |  |  |
| 8 | 0.0 1.0 0.0 | **-1.4661817944830124** | **-0.9069571034860727** |
|  |  |  |  |
| 9 | 1.0 0.0 0.0 | **-0.44973664397484414** | **0.2056403393225306** |
|  |  |  |  |

ຄຳສັ່ງຕົວຢ່າງ:

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

y = le.fit\_transform(y)

ຄ່າຂອງyແປງແລ້ວ

|  |  |
| --- | --- |
|  | Purscased |
|  |  |
| 0 | 0 |
|  |  |
| 1 | 1 |
|  |  |
| 2 | 0 |
|  |  |
| 3 | 0 |
|  |  |
| 4 | 1 |
|  |  |
| 5 | 1 |
|  |  |
| 6 | 0 |
|  |  |
| 7 | 1 |
|  |  |
| 8 | 0 |
|  |  |
| 9 | 1 |
|  |  |