Data Science Use Cases in Healthcare

Data Science Use Cases adalah tugas dunia nyata yang konkret untuk diselesaikan menggunakan data yang tersedia. Data Science Use Cases dapat berupa masalah yang harus dipecahkan, hipotesis yang harus diperiksa, atau pertanyaan yang harus dijawab. Dasarnya, melakukan Data Science berarti menggunakan Use Case di dunia nyata.
Data Science Use Cases di sekuruh industri mencerminkan semakin pentingnya ilmu data, meningkatnya fokus pada MLOps, dan kebutuhan akan otomatis dalam berbagai industri. (Kadam, 2022). Setiap industru, termasuk healthcare menyadari pentingnya menggunakan data scince untuk merampingkan operasi, meningkatkan ROI, membuat keputusan bisnis yang tepat, dan mengoptimalkan proses.

Dalam hal Data Science Use Cases in Healrthcare, penggunaan dan interpretasi yang benar dari data yang tersedia tidak hanya bermanfaat bagi pemasar di sektor, tapi juga dapat membantu diagnosis penyakit serius secara tepat waktu dan bahkan menyelamatkan nyawa orang. Secara khusus, beberapa bidang Healthcare yang sangat khusus seperti genetika, kedokteran reproduksi, onkologi, bioteknolgi, radiografi, diagnostik prediktif, dan farmasi telah pindah ke tingkat yang sama sekali baru berkat membuka potensi penuh data mereka.

Data Science Use Cases in Healthcare : Prediksi Kanker Payudara.

Menurut World Cancer Research Fund International, kanker payudara adalah jenis kanker yang paling umum di kalangan wanita dan yang paling umum kedua secara keseluruhan. Kemajuan metode penambangan data dalam kombinasi dengan algoritma pembelajaran mesin telah memungkinkan untuk memprediksi risiko kanker payudara, mendeteksi potensi anomali pada tahap awal, memperkirakan dinamikanya, dan karenanya mengembangkan rencana optimal untuk melawan penyakit.

Masalah utama dengan menggunakan himpunan data tersebut dalam pengklasifikasi adalah bahwa mereka dapat sangat tidak seimbang. Akibatnya, algoritma cenderung mengklasifikasikan kasus-kasus baru sebagai kasus non-patologis. untuk menilai keakuratan model tersebut secara lebih efisien, masuk akal untuk menerapkan skor F1 sebagai metrik evaluasi, karena memperkirakan positif palsu (kesalahan tipe I) dan negatif palsu (kesalahan tipe II) daripada kasus ketika algoritma mengklasifikasikan entri dengan benar.

Mari kita lihat kumpulan data dunia nyata tentang Kanker Payudara dari salah satu negara bagian AS. Fitur-fitur tersebut dijelaskan secara rinci dalam dokumentasi, tetapi singkatnya, mereka adalah rata-rata, kesalahan standar, dan nilai terbesar dari atribut dari setiap gambar digital yang menampilkan inti sel dari massa payudara seorang wanita. Setiap entri dari himpunan data sesuai dengan seorang wanita dengan tumor ganas atau jinak (yaitu, semua wanita yang bersangkutan memiliki beberapa jenis tumor). Atribut termasuk jari-jari sel, tekstur, kehalusan, kekompakan, cekung, simetri, dll.

```
import pandas as pd

cancer_data = pd.read_csv('data.csv')
pd.options.display.max_columns = len(cancer_data)
print(f'Number of entries: {cancer_data.shape[0]:,}\n'
      f'Number of features: {cancer_data.shape[1]:,}\n\n'
      f'Number of missing values: {cancer_data.isnull().sum().sum()}\n\n'
      f'{cancer_data.head(2)}')
Number of entries: 569
Number of features: 33

Number of missing values: 569

         id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0  842302         M        17.99         10.38          122.8     1001.0
1  842517         M        20.57         17.77          132.9     1326.0

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017

   symmetry_mean  fractal_dimension_mean  radius_se  texture_se  perimeter_se  \
0         0.2419                 0.07871     1.0950      0.9053         8.589
1         0.1812                 0.05667     0.5435      0.7339         3.398

   area_se  smoothness_se  compactness_se  concavity_se  concave points_se  \
0   153.40       0.006399         0.04904       0.05373            0.01587
1    74.08       0.005225         0.01308       0.01860            0.01340

   symmetry_se  fractal_dimension_se  radius_worst  texture_worst  \
0      0.03003              0.006193         25.38          17.33
1      0.01389              0.003532         24.99          23.41

   perimeter_worst  area_worst  smoothness_worst  compactness_worst  \
0            184.6      2019.0            0.1622             0.6656
1            158.8      1956.0            0.1238             0.1866

   concavity_worst  concave points_worst  symmetry_worst  \
0           0.7119                0.2654          0.4601
1           0.2416                0.1860          0.2750
```

```
  fractal_dimension_worst  Unnamed: 32
0                 0.11890         NaN
1                 0.08902         NaN
```

Lepaskan kolom terakhir yang hanya berisi nilai yang hilang:

```
cancer_data = cancer_data.drop('Unnamed: 32', axis=1)
```

Berapa banyak wanita, dalam % yang dipastikan mengidap kanker (tumor payudara ganas)?

```
round(cancer_data['diagnosis'].value_counts()*100/len(cancer_data)).convert_dtypes()
```

```
B    63
M    37
Name: diagnosis, dtype: Int64
```

37% dari semua responden menderita kanker payudara, sehingga dataset sebenarnya agak seimbang.
Karena nilai variabel diagnosis bersift kategoris, kita harus mengkodekannya ke dalam bentuk numerik untuk penggunaan lebih lanjut dalam pembelajaran mesin. Sebelum melakukannya mari kita pisahkan data menjadi fitur prediktor dan diagnosis variabel target:

```
X = cancer_data.iloc[:, 2:32].values
y = cancer_data.iloc[:, 1].values

# Encoding categorical data
from sklearn.preprocessing import LabelEncoder

labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
```

Mari buat rangkaian kereta dan pengujian, lalu skalakan fiturnya:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Untuk memodelkan data kita, mari kita tetapkan algoritma pemodeln berikut dengan parameter defaultnya : k-nearest neighbor (KKN) dan regresi logistik:

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression

# KNN
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
knn_predictions = knn.predict(X_test)

# Logistic regression
lr = LogisticRegression()
lr.fit(X_train, y_train)
lr_predictions = lr.predict(X_test)
```

Karena kumpulan data agak seimbang, maka dapat menggunakan metrik evaluasi skor akurasi untuk mengindentifikasi model yang paling akurat:

```
from sklearn.metrics import accuracy_score

print(f'Accuracy scores:\n'
    f'KNN model:\t\t   {accuracy_score(y_test, knn_predictions):.3f}\n'
    f'Logistic regression model: {accuracy_score(y_test, lr_predictions):.3f}')
```

```
Accuracy scores:
KNN model:        0.956
Logistic regression model: 0.974
```

Berdasarkan data, model regresi logistik berkinerja terbaik saat memprediksi tumor payudara ganas atau jinak untuk wanita dengan patologi payudara yang terdeteksi dalam bentuk apapun. Secara potensial kedepan, mengingat algoritme regresi logistik tidak memiliki parameter penting untuk disesuaikan, kita dapat bermain dengan pendekatan berbeda untuk melatih atau menguji pemisahan dan berbagai teknik pemodelan prediktif. (datacamp, 2022).

References:

Datacamp. (2022, April). *Data Science Use Cases Guide*. Retrieved September 17, 2022, from https://www.datacamp.com/blog/data-science-use-cases-guide

Kadam, M. (2022, Mei 30). *Bacancy*. Retrieved September 17, 2022, from Data Science Use in Retail & Healthcare Industries: https://www.bacancytechnology.com/blog/data-science-use-cases/amp/