Source:
https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Reference:

https://medium.com/analytics-vidhya/diabetes-130-us-hospitals-for-years-1999-2008-e18d69beea4d

Count of instances: 100000
Count of attributes: 55

Clustering:

Random Forest:

```
Accuracy: 0.6207625036847794
Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.70      0.66     10952
           1       0.60      0.53      0.56      9402

    accuracy                           0.62     20354
   macro avg       0.62      0.61      0.61     20354
weighted avg       0.62      0.62      0.62     20354
```

- Accuracy: The overall accuracy of the classification model is 0.6207625036847794, which means the model correctly predicted the target class for approximately 62.08% of the samples.
- Classification Report: The report provides various metrics such as precision, recall, and F1-score for each class (0 and 1) in the dataset.
    - Precision: It measures the accuracy of positive predictions, i.e., the percentage of correctly predicted positive samples out of all positive

predictions. For class 0, the precision is 0.63, and for class 1, the precision is 0.60.
    - Recall: It measures the percentage of true positive samples that were correctly identified by the model out of all actual positive samples. For class 0, the recall is 0.70, and for class 1, the recall is 0.53.
    - F1-score: It is the harmonic mean of precision and recall, providing a balance between the two. For class 0, the F1-score is 0.66, and for class 1, the F1-score is 0.56.

    - Support: It represents the number of samples in each class. For class 0, the support is 10952, and for class 1, the support is 9402.
- Macro Avg: The macro average calculates the average of the metrics across all classes, giving equal weight to each class. In this case, the macro average precision, recall, and F1-score are all around 0.62.
- Weighted Avg: The weighted average calculates the average of the metrics across all classes, weighted by the number of samples in each class. In this case, the weighted average precision, recall, and F1-score are all around 0.62.
- In the report you provided, the accuracy of the model is approximately 0.62, which means it correctly predicted the target class for around 62.08% of the samples. While accuracy is a commonly used metric, it may not provide a complete picture of a model's performance, especially if the classes are imbalanced.
- To better assess the model, it is important to consider other evaluation metrics as well, such as precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify positive and negative samples and its balance between precision (accuracy of positive predictions) and recall (sensitivity to positive samples).
- In addition to the metrics, it is essential to understand the specific requirements and constraints of your problem. For example, if the task involves identifying a rare event, such as a disease, and recall (identifying all positive cases) is crucial, a model with a high recall may be preferred, even if the overall accuracy is lower.

KNN:

Accuracy: 0.5581212538076054

:Classification Report
precision recall f1-score support

10952	0.61	0.64	0.58	0
9402	0.49	0.46	0.52	1
accuracy			0.56	20354
macro avg	0.55	0.55	0.55	20354
weighted avg	0.56	0.56	0.56	20354

In this report:

- Accuracy: The overall accuracy of the KNN algorithm in predicting the "readmitted" value is 0.5581212538076054, which means the model correctly predicted the target class for approximately 55.81% of the samples.
- Classification Report: The report provides various metrics such as precision, recall, and F1-score for each class (0 and 1) in the dataset.
    - Precision: It measures the accuracy of positive predictions, i.e., the percentage of correctly predicted positive samples out of all positive predictions. For class 0, the precision is 0.58, and for class 1, the precision is 0.52.
    - Recall: It measures the percentage of true positive samples that were correctly identified by the model out of all actual positive samples. For class 0, the recall is 0.64, and for class 1, the recall is 0.46.
    - F1-score: It is the harmonic mean of precision and recall, providing a balance between the two. For class 0, the F1-score is 0.61, and for class 1, the F1-score is 0.49.
    - Support: It represents the number of samples in each class. For class 0, the support is 10952, and for class 1, the support is 9402.
- Macro Avg: The macro average calculates the average of the metrics across all classes, giving equal weight to each class. In this case, the macro average precision, recall, and F1-score are all around 0.55.
- Weighted Avg: The weighted average calculates the average of the metrics across all classes, weighted by the number of samples in each class. In this case, the weighted average precision, recall, and F1-score are all around 0.56.

This report provides a summary of the KNN algorithm's performance in terms of accuracy and various metrics for each class. It helps in understanding the model's effectiveness in predicting the "readmitted" value and the balance between precision and recall for different classes.

خوشه بندی:

dbscan روش

The silhouette score is a metric used to evaluate the quality of clustering results. It measures how well each sample in a cluster is assigned to that cluster and how distinct the clusters are from each other.

The silhouette score ranges from -1 to 1, where:

- A score close to 1 indicates that the samples are well-clustered, with each sample being close to its own cluster and far from other clusters.
- A score around 0 suggests overlapping clusters or ambiguous assignments of samples.
- A score close to -1 indicates that the samples are incorrectly clustered, with samples being assigned to the wrong clusters rather than their own.

In general, a higher silhouette score indicates better clustering results, where the clusters are well-defined and well-separated. However, it's important to note that the silhouette score should be interpreted in the context of the specific dataset and problem at hand.

The silhouette_score function in scikit-learn calculates the average silhouette score for all samples in a dataset, given the feature matrix and corresponding cluster labels. It provides a convenient way to assess the quality of clustering algorithms and to compare different clustering solutions.

When I use all the columns as features, clustering performs as:

```
-1      100831
 0          35
 8          30
 11         29
 13         25
        ...
 7           4
 107         4
 77          3
 108         3
 91          3
Name: cluster, Length: 114, dtype: int64
```

the cluster with the label "-1" has the highest count, with 100,831 data points assigned to it. This indicates that a large majority of the data points were grouped into this cluster. The other clusters have significantly smaller counts, with the second-largest cluster having only 35 data points.

important to note that the choice of features for clustering can significantly impact the resulting clusters. By including only a subset of columns as features, you may be emphasizing certain aspects of the data while disregarding others. This can lead to different cluster assignments and potentially reveal different structures or groupings within the data.
When I choose only a selection of features:

```
# Select features for clustering
features_db3 = diabet[['time_in_hospital',
'num_lab_procedures', 'num_procedures',
                   'number_emergency',  'insulin',
'change', 'diabetesMed']]
```

```
Number of clusters decreased to 214:
0      9191
21     6653
7      6373
4      6001
19     4695
       ...
198       4
162       3
184       3
187       3
207       3
Name: cluster, Length: 214, dtype: int64
```

And silhouette_score also got a higher score which means the clustering is more precise.