Impact of Clustering on Regression Models

Group 4: Camy Ngo And Rena Ahn

DSP3: Final Report

GitHub Link to Project Repository: https://github.com/mscs5305-group4/clustering-patients

Table of Contents

1. Problem Statement

We received data on 114 patients with severe asthma, a chronic respiratory condition. When observing the patient data, we noticed there may be correlations between higher PEFR values and other provided demographic information (e.g. age, weight, height, smoking status, occupation, living environment). Our project aims to clearly identify and determine these correlations. To do so, we will calculate summary statistics of each patient data and combine them with demographic information to find trends among the data. Through these trends, our project will develop both a regression predicting patients' PEFR statistics and a clustering model. Ultimately, we will evaluate the effect of clustering on the performance of regression modeling through analysis.

2. Data Sources

The data for this project was received from the Division of Allergy and Respiratory Medicine at Soonchunhyang University Bucheon Hospital, South Korea.

PEFR_asthma_114_medinfo_07.15 Dataset: 14 columns, 99 valid rows

- Demographic Features: age, sex, smoking status, height, weight, BMI, occupation.
- Additional patient profile information (UID, BCODE).

99 Individual Patient Data Files: 14 columns, varying rows

- Features: pefr_am, pefr_pm, pefr_other.
- There is a considerable amount of missing data that must be handled.

3. Data Preprocessing and Cleaning

Handling Missing Data

- Empty cells and '-' were processed as missing data.
- Patient Files
    - Imputation Method Analysis
        - Forward fill, backward fill, linear interpolation, median (of column) imputation, and mean (of column) imputation methods were analyzed for performance.
        - Cumulative error was calculated to reveal forward fill performs well on shorter gaps and linear interpolation performs above average overall, making it a safe choice for longer imputations.
    - Data Integrity: Columns with more than 30% missing data are not imputed to prevent introducing bias.
    - Short Gaps (≤ 14 days missing) imputed with forward fill.
    - Long Gaps (> 14 days missing) imputed with linear interpolation.
- Demographic File
    - Address: 1 missing value
        - Missing value was later dropped in the process of preprocessing.
    - Occupation: 6 missing values
        - Missing value was later dropped or considered 'Medium' exposure.

Handling Erroneous Data

- Demographic File: no reasonably erroneous value found.
- Patient Files: values over 1000 filtered as missing.

occupation, address Columns

- The occupation and address columns were translated from Korean to English
- The occupation column was transformed to 'exposure' which organizes the level of exposure the patient may have to substances and particles harmful for asthma patients as 'Low', 'Medium', and 'High' [2].
    - For example, the 'Construction Worker' occupation will be transformed to 'High' exposure.
- The address column was transformed to 'region' because we determined specific address information will not be needed as patients are located all over Korea; the region or providence one lives in will be enough.

4. Statistic Summary Calculation

Processing of PEFR Values to Calculate Statistics

- Only patients with pefr_am values were used to calculate statistics and generate a processed dataset. This is because pefr values tend to vary according to the hour of the day. Therefore, using only the pefr_am column reduces this variance and different to keep a consistent dataset.
- Therefore, the dataset was reduced to 82 patients.
- Calculated Statistics: count, min, max, median, mean, sum, std, skew.

Feature Encoding

- Encoded categorical variables (sex, smoke, exposure, region) as we will use clustering models such as K-Means which work with numerical data.

Merging for Collective Patient Data

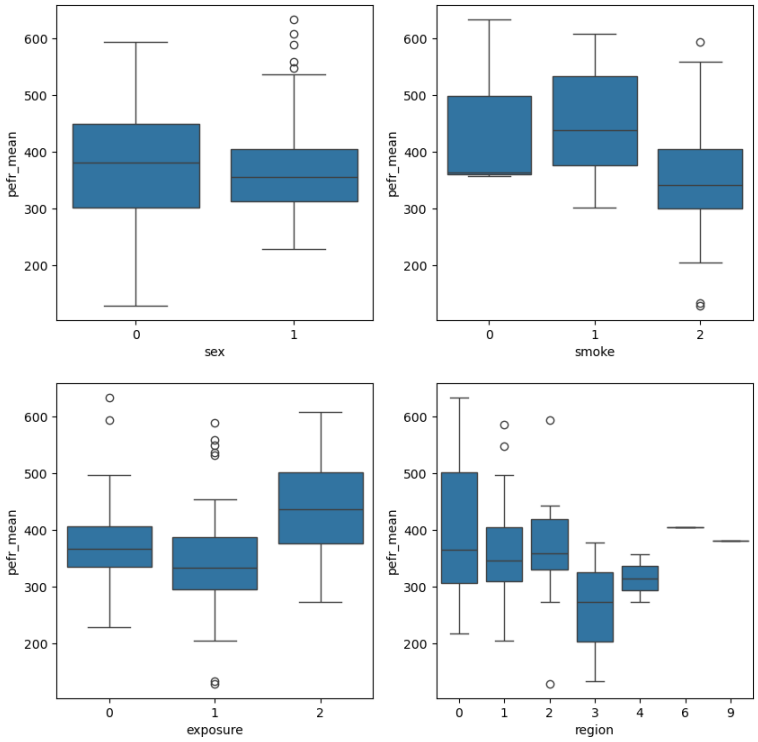
- Statistic data (calculated from patient PEFR values) and demographic data file merged.
- The final collective file has 82 patients and 19 columns.
    - Column features: ID, age, sex, smoke, smoke_amount, height, weight, BMI, BSA, exposure, region, pefr_count, pefr_min, pefr_max, pefr_median, pefr_mean, pefr_sum, pefr_std, pefr_skew

5. Exploratory Data Analysis (EDA)

EDA was performed in two different ways. First, it was performed before modeling on the whole dataset to identify overall trends with the purpose of providing a direction during modeling. Later, having developed clusters using clustering models, EDA was performed on each individual cluster with the purpose of comparing the clusters.

5.A. Population: Whole Dataset

Figure 1: boxplot of categorical features vs. pefr_mean



0: male,

1: female

0: smoker;

1: ex-smoker;

2: non-smoker

0: low exposure occupation;

1: medium exposure occupation;

2: high exposure occupation

The pefr_mean feature is influenced most by categorical features and has distinct medians and IQRs depending on the category.Figure 2 shows these relationships:

- The pefr_mean of females is slightly less and have smaller IQR than that of males.
- The pefr_mean of non-smokers are less than that of smokers and ex-smokers.
- The pefr_mean of high-exposure occupations are greater than that of low and medium exposure occupations.
- The pefr_mean of patients living in region 3 is the smallest of all regions and has a distinct IQR from the other regions.

5.B. Cluster Comparison

Figure 4: boxplot of categorical features vs. pefr_mean grouped by cluster

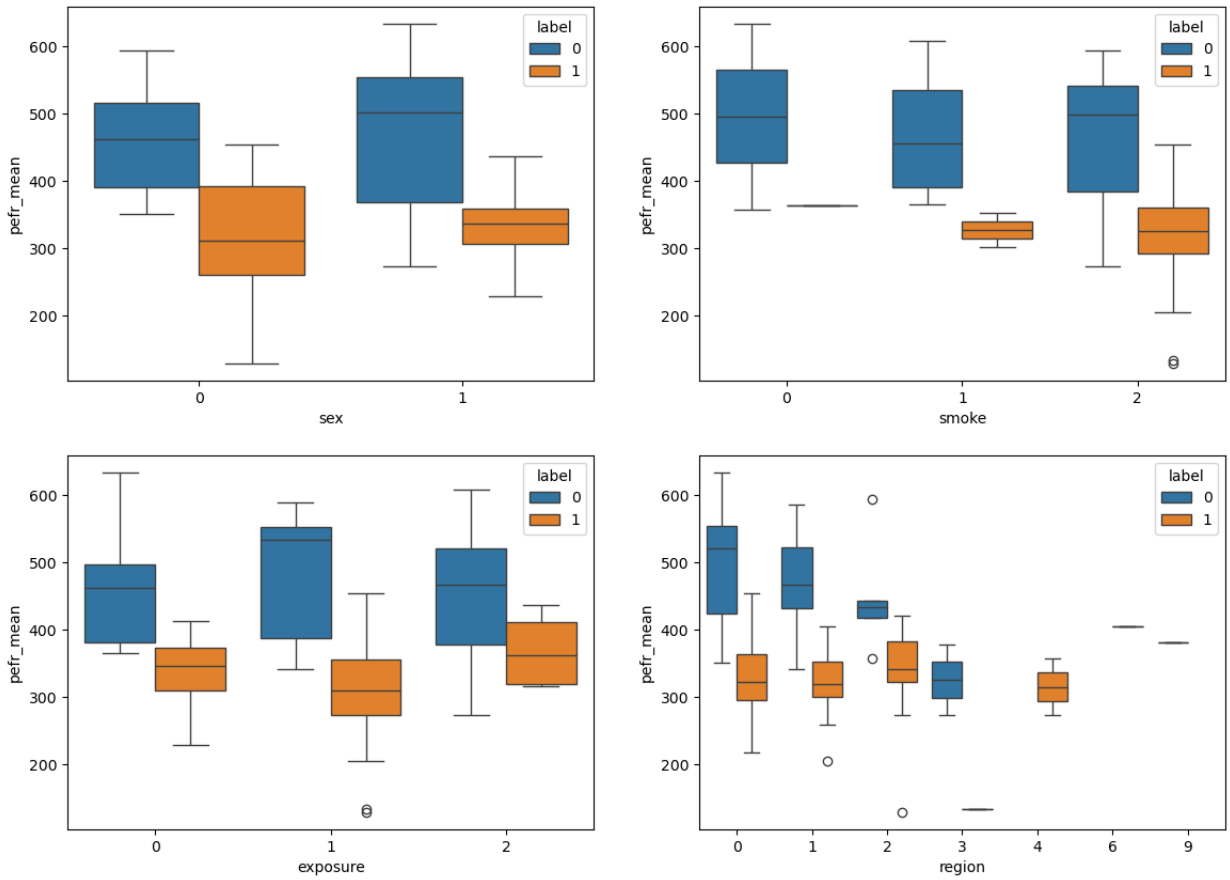


Figure 4 shows the pefr_mean values of each cluster is distinct from each other:

- The pefr_mean values of cluster 0 are greater than that of cluster 1.
- The pefr_mean values of cluster 0 have IQRs that are almost always distinct from the IQRs of cluster 1. That is, the IQRs do not overlap.

6. Modeling Results

Two types of models were developed: regression and clustering. First, each type of model underwent training, hyperparameter tuning, etc. as appropriate. Then, the regression model was trained on each cluster produced by the clustering model.

6.A. Regression Modeling

Three regression models—**Linear Regression**, **Lasso Regression**, and **Ridge Regression**—were trained and evaluated. **Linear Regression** was selected as the final model. This choice was motivated by the following reasons:

1. Linear Regression simplifies interpretation, making it easier to visualize and understand relationships between variables.
2. Regularization (Lasso and Ridge) showed no significant improvement, hence Linear Regression as the optimal choice.

Target Variable (y): pefr_mean

Attribute Variables (X): age, sex, smoke, smoke_amount, height, weight, BMI, BSA, exposure, region, pefr_count, pefr_min, pefr_median, pefr_max, pefr_sum, pefr_std, pefr_skew

The following is the evaluated metrics of the linear regression model:

```
{'R-squared (Train)': 0.9982289888958417,
 'R-squared (Test)': 0.9814949100574341,
 'RMSE': np.float64(10.928671570494606),
 'Intercept': np.float64(36.833374869711804),
 'Coefficients': array([-2.64485794e-02, -1.02153413e+00,  4.93196174e+00,  3.70256890e-01,
        -1.54989736e+00, -3.52560426e+00, -6.17829747e-01,  2.63707197e+02,
         1.27775711e+00, -7.61201736e-02, -3.10469015e-03, -9.69726082e-03,
         9.85914399e-01,  2.11103151e-02,  1.29711720e-06, -4.38090588e-02,
         7.21867123e-01])}
```

Feature Coefficients

```
Regression Equation:
y = -8.36 + (-0.00) * age + (-1.22) * sex + (-0.11) * smoke + (0.06) * smoke_amount + (0.27)
 * height + (-0.28) * weight + (0.84) * BMI + (-14.10) * BSA + (1.04) * exposure + (-0.26) *
 region + (-0.01) * pefr_count + (-0.01) * pefr_min + (0.97) * pefr_median + (0.02) *
 pefr_max + (0.00) * pefr_sum + (-0.27) * pefr_std + (1.31) * pefr_skew
```

The coefficients reflect the weight of each feature's impact on the target variable.

Cross-Validation Results

The data was split into 5 folds for validation, ensuring robustness of the results. The results from the 5-fold validation are as follows:

Fold	MSE	R2
1	119.435862	0.981495
2	54.533780	0.995011
3	65.974709	0.993317
4	33.017830	0.997669
5	44.632820	0.994902

Interpretation of Results

BMI and BSA Analysis

1. **BMI**: Represents body mass relative to height. The positive coefficient suggests individuals with higher BMI might exhibit different smoking behaviors or experience greater health impacts from smoking (e.g., reduced lung capacity, inflammation).
2. **BSA**: Balances both height and weight in physical surface terms. The negative coefficient suggests that greater BSA correlates inversely with **perf_mean**, likely reflecting unique physiological impacts associated with body composition.
3. **Height's Indirect Role**: Height affects BMI and BSA but does not directly influence smoking behavior. Instead, the combined metrics capture physical characteristics associated with smoking-related outcomes.
   For example, individuals of the same weight but varying heights will have different BMI and BSA values, potentially altering their smoking-related health impacts.

Model Evaluation Metrics

- Mean Squared Error (MSE):
  The average MSE (~63.92) confirms a reliable model fit, with lower values indicating better performance.
- R^2 (Coefficient of Determination):
  The consistently high R^2 values across folds (0.981–0.998) underscore the model's ability to explain the variance in the target variable effectively.

6.B. Clustering Modeling

The StandardScalar, PCA, and KMeans modules of the scikit-learn package were used to develop a K-means clustering model [1].

- StandardScalar was used to scale the dataset.
- PCA was used to reduce the dimensionality of the dataset and improve model performance.
- KMeans was used to train a K-Means clustering model.

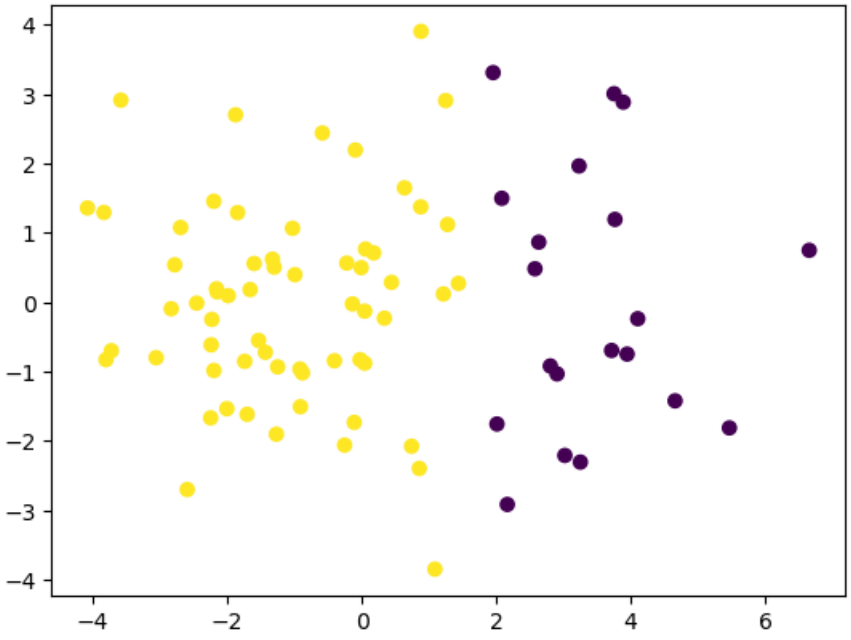
The silhouette_score, calinkski_harabasz_score, and davies_bouldin_score functions of the scikit-learn package were used to evaluate the clustering model [1].

Hyperparameter tuning was performed to determine the ideal number of components (n_components) for PCA and the ideal number of clusters (n_clusters) for the KMeans model. The analysis showed the best model performance with 12 components (n_components=12) and 2 clusters (n_clusters=2).

The clusters are visualized in Figure 5. They yielded the following results:

	silhouette	calinski harabasz	davies bouldin
Trial 1	0.23988039917111553	23.22468076729419	1.7905329576511109
Trial 2	0.27058383999081853	21.815529473034257	1.6863762059726048
Trial 3	0.2496619357096872	23.197813336003385	1.7648798474496472

Figure 5: KMeans model cluster visualizations



6.C. Analysis: Regression Modeling on Clusters

The trained LinearRegression model was then applied to Cluster 0 and Cluster 1.

Target Feature (y): pefr_mean

Attribute Features (X): age, sex, smoke, smoke_amount, height, weight, BMI, BSA, exposure, region, pefr_count, pefr_min, pefr_max, pefr_sum, pefr_std, pefr_skew

- pefr_median was removed due to the feature being almost perfectly correlated with the pefr_mean.

Regression equation on the whole dataset:

y = 384.72 + (-0.36) * age + (4.57) * sex + (-10.69) * smoke + (-0.28) * smoke_amount
+ (4.94) * height + (18.72) * weight + (-7.65) * BMI + (-1194.44) * BSA
+ (-1.38) * exposure + (-1.67) * region + (-0.10) * pefr_count
+ (0.22) * pefr_min + (0.45) * pefr_max + (0.00) * pefr_sum + (-1.06) * pefr_std
+ (-8.57) * pefr_skew

Regression Equation on Cluster 0:

y = 1933.58 + (-0.09) * age + (10.65) * sex + (-2.77) * smoke
+ (-0.24) * smoke_amount + (-1.49) * height + (31.93) * weight + (-36.57) * BMI
+ (-1454.93) * BSA + (-6.39) * exposure + (-4.25) * region + (-0.27) * pefr_count
+ (0.05) * pefr_min + (0.24) * pefr_max + (0.00) * pefr_sum + (-0.46) * pefr_std
+ (-1.48) * pefr_skew

Regression Equation on Cluster 1:

y = 2604.90 + (-0.06) * age + (6.32) * sex + (-2.32) * smoke + (-0.45) * smoke_amount
+ (-15.25) * height + (21.82) * weight + (-53.39) * BMI + (-51.61) * BSA
+ (-1.29) * exposure + (-3.78) * region + (-0.09) * pefr_count
+ (0.30) * pefr_min + (0.31) * pefr_max + (0.00) * pefr_sum + (-0.60) * pefr_std
+ (-22.79) * pefr_skew

The difference in regression equations proves clustering has an impact on regression models. Thus, it is reasonable to proceed in investigating the impact of clustering on the performance of regression models.

In addition to the Mean Squared Error and R-Squared calculated during Regression Modeling, Relative Error was also calculated to better compare results of the regression models.

The regression model's results on the whole dataset are the following:

Fold	Mean Squared Error	R-Squared	Relative Error
1	2365.391188	0.633512	1.657084
2	1999.509390	0.817059	1.598833
3	4175.302437	0.577034	0.994056
4	2589.772276	0.817179	2.841265
5	1430.315555	0.836624	1.168266
Avg.	2512.0581692	0.7362816	1.6519008

The regression model's results on Cluster 0 are the following:

Fold	Mean Squared Error	R-Squared	Relative Error
1	540.213189	0.868421	0.215165
2	272.484987	0.957658	0.161694
3	1542.512996	0.772229	0.383772
4	197.096961	0.977177	0.189287
5	500.215097	0.940425	0.200547
Avg.	1017.5077433	0.903182	0.1610404

The regression model's results on Cluster 1 are the following:

Fold	Mean Squared Error	R-Squared	Relative Error
1	1521.574435	0.560799	1.053520
2	2009.578023	0.714075	1.184441
3	328.741844	0.899019	0.446514
4	1150.116369	0.599758	0.852596
5	1150.863121	0.765872	0.821225
Avg.	3080.436896	0.7079046	0.8716592

7. Analysis of Results

Through the results of regression modeling on each cluster in Section 6.C, the impact of clustering on the performance of regression models can be analyzed.

The following is the difference between the results of the model on Cluster 0 and the results of the model on the whole dataset:

Fold	Mean Squared Error	R-Squared	Relative Error
1	-1825.177999	0.234909	-1.441919
2	-1727.024403	0.140599	-1.437139
3	-2632.789441	0.195195	-0.610284
4	-2392.675366	0.159998	-2.651978
5	-930.100458	0.103801	-0.967719
Avg.	-1901.5535334	0.166904	-1.4217878

The following is the difference between the results of the model on Cluster 1 and the results on the model on the whole dataset:

Fold	Mean Squared Error	R-Squared	Relative Error
1	-843.816753	-0.072713	-0.603564
2	10.068633	-0.102984	-0.424392
3	-3846.560593	0.321985	-0.547542
4	-1439.655907	-0.217421	-1.988669
5	-279.452434	-0.070752	-0.347041
Avg.	-1279.8834108	-0.0283306	-0.7822416

From the differences, it is shown that clustering improves the performance of the regression model overall. Particularly, a decrease in error metrics can be observed.

This impact is especially significant in Cluster 0: mean squared error reduced by an average of 1901.55, R-squared increased by an average of 0.167, and relative error reduced by an average of 1.42. In addition, all folds of Cluster 0 saw an improvement in the regression model.

There is less impact on Cluster 1: mean squared error reduced by an average of 1279.88, R-squared decreased by an average of 0.03, and relative error reduced by an average of 0.78. Overall, there was an improvement in the model, but old 2 had an increase in mean squared error. And although the average decrease in R-squared is small at -0.03, only fold 3 saw an increase in R-squared. Thus, clustering reduced the error of, but not R-squared of, the regression model for Cluster 1.

Citations

[1] *2.3. clustering*. scikit. (n.d.). https://scikit-learn.org/1.5/modules/clustering.html

[2] Better Health Channel. (2021, April 6). *Asthma and your workplace*. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/asthma-and-your-workplace