Técnicas de aprendizaje no supervisado aplicadas al análisis de absentismo laboral en empresa de Brasil

Presentado por:
Juan Felipe Osorio López | Aura Maria Molina Amaya
Yackeline Cristina Quintero López | Juan José Toro Villegas

Planteamiento del problema

El planteamiento del problema se centra en la necesidad de realizar un análisis exhaustivo de los datos de ausentismo laboral utilizando técnicas de aprendizaje no supervisado; en este sentido, se plantea la siguiente pregunta de investigación: ¿Cómo pueden las técnicas de aprendizaje no supervisado, como el clustering y la reducción de dimensionalidad, utilizarse para analizar el conjunto de datos de Absentismo en el trabajo y proporcionar información relevante para la gestión y prevención del ausentismo laboral en una empresa de mensajería en Brasil?

Metodología o procedimiento

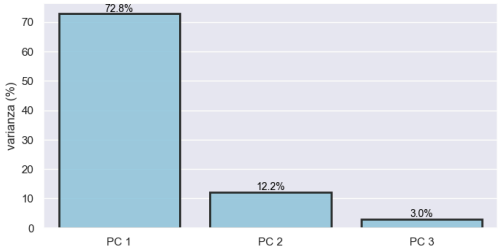
Inicialmente se realizó una búsqueda de información en centros de datos para elegir una base de datos, para esto se empleó el *UC Irvine Machine Learning Repository*, donde se encontró una base de datos que se ajusta a las necesidades propuestas para el desarrollo del estudio de caso. Una vez elegida la base de datos, se carga en python y se examinan los datos y se verifica su integridad y estructuctura; se continúa con el análisis exploratorio identificando la distribución y variabilidad; se procede a realizar la limpieza de los datos, eliminando filas duplicadas y variables irrelevantes que no aportan al análisis. Se explora la correlación de los datos y se identifican dos variables con mayor correlación peso y masa corporal, y para identificar patrones de asociación lineal entre los atributos eliminamos la variable masa corporal.

Aprendizaje NO Supervisado

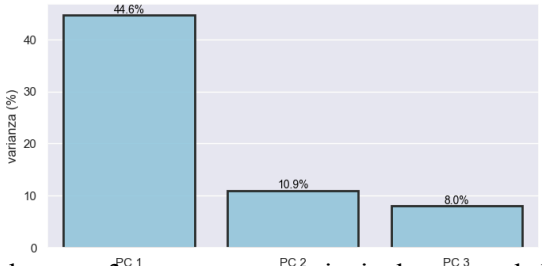
Se realiza la aplicación de los Algoritmos de Clustering (k-means escalado: definiendo el k óptimo, definición de las constantes para k-means, y evaluando las métricas de desempeño y K-means reducido: selección del k óptimo, definición de las constantes para k-means y evaluación de las métricas de desempeño)

Resultados

Reducción de la dimensionalidad: Se está reduciendo efectivamente la dimensionalidad de los datos a solo 3 componentes principales, lo que facilita la interpretación y el análisis posterior.



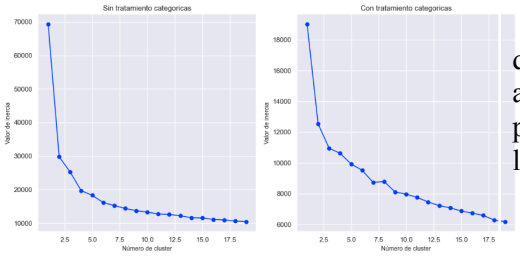
Al tomar 3 componentes principales, con el dataset sin atípicos reducido, estamos explicando 88% de la varianza.



Al tomar 3 componentes principales, con el dataset sin atípicos con tratamiento de la variable categórica reducido, estamos explicando 63.5% de la varianza.

Algoritmos de Clustering

K-MEANS (ESCALADO) : **K-óptimo (escalado)**



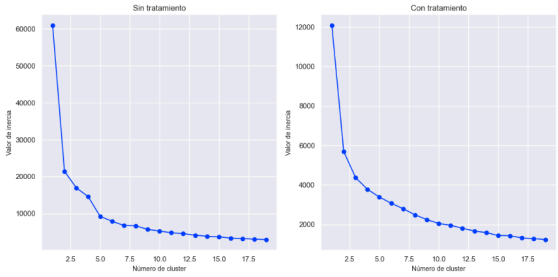
No es clara la curva para identificar el codo, puede ser que este dataset no es el óptimo para realizar este tipo de agrupaciones o algoritmos de clustering. También la posibilidad de tener variables que podrían influir en el posicionamiento de los puntos. La naturaleza de las variables que nos estén afectando.

Métricas de desempeño

Modelo	Inertia	Silhouette Score	Calinski-Harabasz Score
Modelo 1 K-Means escalado (4 cluster)	19681.4	0.282076	585.238
Modelo 2 K-Means escalado y tratado (7 cluster)	8659.53	0.151742	138.077
Modelo 3 K-Means escalado y tratado (4 cluster)	10150.8	0.217294	202.521

El modelo 1, a pesar de tener una inercia muy alta en comparación a los otros modelos, presenta unas métricas de silueta y Calinski mayor. Por lo tanto tomaremos como referencia el primer modelo (con las variables numéricas escaladas y sin tratamiento de las variables categóricas) y compararlo con los demás algoritmos.

K-MEANS (REDUCIDO) :k-óptimo (reducido)



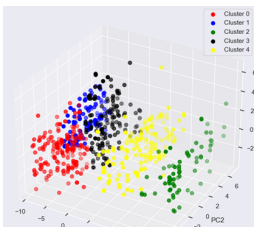
No es clara la curva para identificar el codo, puede ser que este dataset no es el óptimo para realizar este tipo de agrupaciones o algoritmos de clustering. También la posibilidad de tener variables que podrían influir en el posicionamiento de los puntos. La naturaleza de las variables que nos estén afectando.

Métricas de desempeño

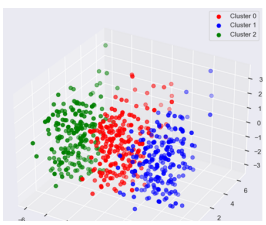
Modelo	Inertia	Silhouette Score	Calinski-Harabasz Score
Primer modelo K-Means Reducido	9233.61	0.387498	973.553
Segundo modelo con tratamiento categoricas K-Means Reducido	4368.44	0.331464	615.058

El primer modelo con PCA reducido presenta una inercia mucho mejor en comparación con el modelo de K-Means. Aumentando también en los resultados de silueta y Calinski. Y, en comparación con el modelo reducido con tratamiento de la variable categórica, a pesar de tener una inercia mayor, en las otras dos métricas es superior (aunque en silueta no es mucha la diferencia).

Calinski- 973

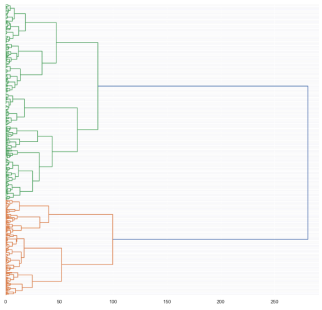


Calinski 615



El primer modelo tiene un Calinski-Harabasz Score más alto (973.553) en comparación con el segundo modelo (615.058). Un mayor Calinski-Harabasz Score indica una mejor separación entre los clusters y una mejor cohesión dentro de los clusters.

Hierarchical Clustering



El dendrograma muestra la estructura de agrupación jerárquica de los datos y sus ramas representan las fusiones de clusters y la altura en el eje e indica la distancia o similaridad entre los clusters que se fusionan, podemos decir que:

A medida que aumentamos el número de clusters, el Silhouette Score disminuye gradualmente, lo que indica una menor cohesión y separación entre los clusters.

El Calinski-Harabasz Score también disminuye al aumentar el número de clusters, lo que sugiere una menor separación y estructura en los clusters.

La mejor división parece ser entre 2 y 3 clusters, ya que tienen los Silhouette Score y Calinski-Harabasz Score más altos en comparación con 4, 5 y 6 clusters.

	Clusters	Silhouette Score	Calinski Harabasz Score
0	2	0.562577	1280.118254
1	3	0.475162	934.557077
2	4	0.333618	867.165999
3	5	0.319032	821.812863
4	6	0.309038	773.849303

DBSCAN

DBSCAN con dataset escalado

```
# Evaluacion del modelo
print(" DBSCAN ")
print('silhouette_score: ', silhouette_score(df1_out, modelo_db.labels_))
print('calinski_harabasz_score: ', calinski_harabasz_score(df1_out, modelo_db.labels_))
✓ 0.0s
DBSCAN
silhouette_score:  0.346327040250145
calinski_harabasz_score:  490.6637249326333
```

El modelo DBSCAN ajustado (Tuning) supera a los otros dos modelos en términos de Silhouette Score y Calinski-Harabasz Score, lo que indica una mejor capacidad para identificar clusters significativos y cohesivos en los datos.

El modelo DBSCAN escalado también muestra un rendimiento competitivo, especialmente en términos de Calinski-Harabasz Score, pero el DBSCAN ajustado logra una mejor separación y estructura de clusters en este caso.

Ajustar los hiperparámetros de los modelos de clustering puede tener un impacto significativo en su rendimiento, y es crucial para obtener resultados más precisos y útiles en el análisis de datos.

DBSCAN con dataset reducido PCA

```
# Evaluacion del modelo
print(" DBSCAN ")
print('silhouette_score: ', silhouette_score(X_pca, modelo_db.labels_))
print('calinski_harabasz_score: ', calinski_harabasz_score(X_pca, modelo_db.labels_))
✓ 0.0s
DBSCAN
silhouette_score:  0.09929757445402665
calinski_harabasz_score:  153.97375037535522
```

DBSCAN con dataset reducido PCA tuneado

```
# Evaluacion del modelo
print(" Tunning DBSCAN ")
print('silhouette_score: ', silhouette_score(X_pca, modelo_db_best.labels_))
print('calinski_harabasz_score: ', calinski_harabasz_score(X_pca, modelo_db_best.labels_))
✓ 0.0s
DBSCAN
silhouette_score:  0.4269978390321452
calinski_harabasz_score:  464.2893634763519
```

Gaussian Mixture Dataset escalado

```
# Evaluacion del modelo GMM
labels_ = model_gmm.predict(df1_out)

print("Gaussian Mixture Model")
print('silhouette_score: ', silhouette_score(df1_out, labels_))
print('calinski_harabasz_score: ', calinski_harabasz_score(df1_out, labels_))
✓ 0.0s
Gaussian Mixture Model
silhouette_score:  -0.04134806010183432
calinski_harabasz_score:  49.62411844725509
```

Silhouette Score: El modelo GMM con PCA tiene un Silhouette Score más alto (0.280) que el modelo GMM sin PCA (0.233). Esto indica que la distancia media entre las muestras de un cluster y las de los clusters vecinos es mayor en el modelo con PCA, lo que sugiere una mejor separación y cohesión de clusters.

Gaussian Mixture Dataset reducido PCA

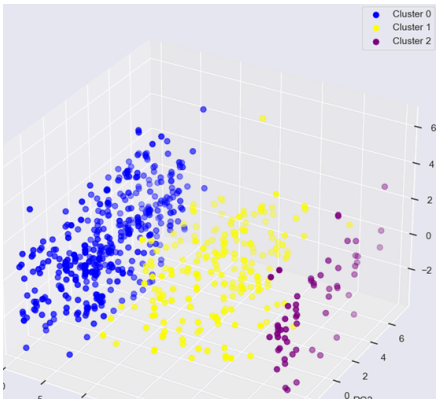
```
# modelo GMM PCA

model_gmm_pca = GaussianMixture(n_components=3, random_state=123, covariance_type='full').fit(X_pca)
✓ 0.0s


# Evaluacion del modelo GMM PCA
labels_ = model_gmm_pca.predict(X_pca)

print("Gaussian Mixture Model PCA")
print('silhouette_score: ', silhouette_score(X_pca, labels_))
print('calinski_harabasz_score: ', calinski_harabasz_score(X_pca, labels_))
✓ 0.0s
Gaussian Mixture Model PCA
silhouette_score:  0.4761069333519351
calinski_harabasz_score:  928.6698174784516
```

Calinski-Harabasz Score: Similarmente, el modelo GMM con PCA obtiene un Calinski-Harabasz Score más alto (145.411) en comparación con el modelo GMM sin PCA (118.960). Este puntaje más alto indica una estructura de clusters más densa y separada en el modelo con PCA.

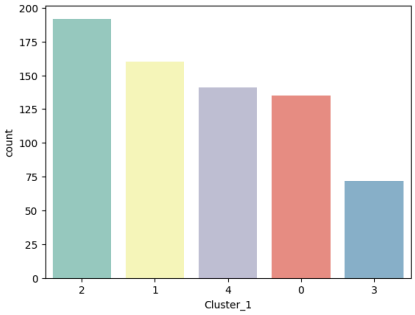


Reducción de la Superposición de Clusters: Ambos modelos muestran cierta superposición o confusión entre clusters, como se refleja en los Silhouette Scores relativamente bajos. Sin embargo, el modelo GMM con PCA logra reducir esta superposición en comparación con el modelo sin PCA, como lo indican los puntajes más altos.

Mejora en la separación de Clusters con PCA: La aplicación de PCA ha mejorado significativamente la capacidad de separación y la estructura de clusters en el modelo GMM. Esto se evidencia en el aumento en los puntajes de Silhouette y Calinski-Harabasz Scores, lo que sugiere una mejor definición y cohesión de los clusters.

Análisis

Partiendo del modelo seleccionado con el algoritmo de K-Means con el dataset reducido con PCA, la composición de los clusters, quedó distribuida así:



El cluster 2, se lleva la mayor cantidad de registros seleccionados por el algoritmo, por encima de los 180 datos, mientras que el cluster 1 lo sigue con una cantidad superior a los 150 registros. Los clusters 4 y 0 si mantienen aproximadamente la misma proporción de los datos entre 125 y 130. Finalmente, el cluster 3 si presenta la menor cantidad con un valor de registros inferior a los 70 datos.

Interpretación de los clusters formados:

Clúster	Descripción
Cluster 0	• La razón de ausencia corresponde a fisioterapia y consulta dental con 34% y 42.9 % respectivamente. • Mes de ausencia significativa entre los tres primeros meses (enero, febrero y marzo) y registrados los viernes en su mayor proporción. • En la estación de otoño se registra con un 54%. • Presentan una edad de 38 años. • El gasto en transporte en promedio está alrededor de 200 UM (unidades monetarias). • El ausentismo se caracteriza alrededor de las 2 horas en la mayor parte de los registros.
Cluster 1	• La razón de ausencia corresponde a enfermedades del sistema osteomuscular y del tejido conjuntivo con un 33.7% • Las ausencias en mayor proporción entre los meses de marzo y junio. Con día más significativo los lunes. • La mayoría tienen alrededor de 28 años, pero también presenta gran proporción en 37 y 38 años. • El gasto en transporte en promedio está alrededor de 222 UM (unidades monetarias). • El ausentismo se caracteriza alrededor de las 8 horas en la mayor parte de los registros.

Cluster 2	● La razón de ausencia corresponde a consulta médica representado en un 44.3% y consulta dental de un 26%. ● Ausencias concentradas en los últimos 5 meses desde agosto a diciembre. Siendo los martes con la mayor proporción. ● Se presenta en la estación de primavera con un 60%. ● La mayor cantidad tienen una edad de 28 años, pero su promedio oscila entre los 36 y 40 años. ● El gasto en transporte en promedio es alrededor de 227 UM (unidades monetarias). ● El ausentismo se caracteriza de 1 a 8 horas, con un promedio de 4.9 horas.
Cluster 3	● No presenta razón de ausencia en 59.7 %, seguido de enfermedades infecciosas y parasitarias con un 22.2 %. ● Ausencias registradas en el mes de octubre, con el miércoles de día de mayor proporción. ● Estan en esta agrupación se concentran los que tienen fallos disciplinarios. ● El rango de edad oscila entre los 36 y 50 años. ● El gasto en transporte en promedio es alrededor de 241 UM (unidades monetarias). ● No registran horas de ausentismo.
Cluster 4	● La razón de ausencia corresponde a consulta médica y lesiones, envenenamientos y algunas otras consecuencias de causas externas con un 40% y 23% respectivamente. ● Las ausencias en mayor proporción entre los meses de marzo y julio. Distribuidos todos los días de la semana. ● Presentan una edad de 28 años. ● El gasto en transporte en promedio es alrededor de 233 UM (unidades monetarias). ● El ausentismo se caracteriza de 2 a 8 horas, con un promedio de 7 horas.

Conclusiones

1. Desafíos en la aplicación de algoritmos de clustering: El análisis de los algoritmos de clustering, como K-Means y DBSCAN, revela algunos desafíos en la identificación de la estructura subyacente en los datos de ausentismo laboral. La falta de claridad en la identificación del codo en las curvas de codo y la posible influencia de variables no consideradas sugieren la necesidad de un enfoque más exhaustivo para seleccionar el número óptimo de clusters y mejorar la interpretación de los resultados.

2. La reducción de dimensionalidad mediante PCA permite explicar una proporción significativa de la varianza en los datos de ausentismo laboral, con un nivel aceptable de retención de información. Al tomar sólo tres componentes principales, se explica hasta el 88% de la varianza en el dataset sin atípicos. Esto sugiere que PCA es una técnica eficaz para simplificar la estructura de los datos y facilitar la interpretación de los resultados obtenidos a través de técnicas de clustering.

3. Los clusters identificados revelan patrones distintivos de ausentismo laboral dentro de la población estudiada. Los diferentes clusters muestran variaciones en cuanto a las razones de ausencia, el momento y la duración de las ausencias, así como características demográficas y laborales. Estos hallazgos proporcionan información valiosa para comprender las causas subyacentes del ausentismo laboral y pueden servir como base para el diseño de estrategias de gestión y prevención más efectivas en la empresa de mensajería en Brasil.

Recomendaciones

- Se sugiere realizar análisis adicionales para profundizar en la comprensión de las causas subyacentes del ausentismo laboral identificado en los clusters. Esto puede incluir la realización de encuestas o entrevistas con los empleados para obtener información cualitativa sobre sus motivos de ausencia, así como la exploración de factores externos, como el clima laboral o las políticas de la empresa, que puedan influir en el ausentismo.

- Implementar un monitoreo continuo del ausentismo laboral: Se recomienda establecer un sistema de monitoreo regular del ausentismo laboral utilizando herramientas analíticas y métricas relevantes. Esto permitirá a la empresa detectar tendencias emergentes, identificar patrones de ausentismo específicos y tomar medidas preventivas de manera proactiva para abordar los problemas antes de que se conviertan en problemas mayores.

- Basándose en los patrones identificados en los clusters y en el análisis adicional realizado, se recomienda desarrollar estrategias de intervención personalizadas dirigidas a grupos específicos de empleados. Estas estrategias pueden incluir programas de bienestar, capacitación en gestión del tiempo o políticas flexibles de trabajo que aborden las necesidades individuales y promuevan un ambiente laboral saludable y productivo.

Referencias

[1] M. A. Llano-Restrepo, “Redacción y publicación de artículos científicos,” *Ing. y Compet.*, vol. 8, no. 2, pp. 112–127, 2006.

[2] Datos del conjunto de datos "Absenteeism at Work", disponible en: https://archive.ics.uci.edu/dataset/445/absenteeism+at+work