**Propuesta para clasificar sitios de localización de proteínas aplicando y comparando modelos con Random Forest y K-means**

**Huamani Loredo, Alvaro**

20151966@aloe.ulima.edu.pe

Universidad de Lima

# INTRODUCCIÓN

El presente documento analizará un dataset de sitios de localización de proteínas (Nakai, 1996), para esto se empezará aplicando un modelo con Random Forest, luego se aplicarán métodos para intentar mejorar la precisión del modelo, luego se implementará un modelo de K-means con el fin de lograr desarrollar un modelo con una mayor precisión ya sea con entrenamiento supervisado o no supervisado; también se buscará medir la precisión del modelo no supervisado, ya que el dataset cuenta con una columna que es considerado la variable dependiente.

# BASES TEÓRICAS

## Random forest

Los bosques aleatorios son una combinación de predictores de árboles de manera que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque (Breiman, 2001).

Una ventaja de este modelo es que ya no es necesario usar un método como cross validations para medir la precisión del modelo, en este informe solo se usará el accurracy de la matriz de confusión.

## K-means

K-means es uno de los métodos de agrupamiento más populares. La idea básica es: dada una agrupación inicial pero no óptima, reubique cada punto en su nuevo centro más cercano, actualice los centros de agrupación calculando la media de los puntos miembros y repita el proceso de reubicación y actualización hasta que se satisfagan los criterios de convergencia (Mannor et al., 2011).

## Grid Search

Una forma de mejorar la precisión de un modelo sería jugar con los hiperparámetros manualmente, hasta que encuentre una gran combinación de valores de hiperparámetros. Este sería un trabajo muy tedioso y es posible que no tenga tiempo para explorar muchas combinaciones. Implementando Grid Seach todo lo que se necesita hacer es decidir con qué hiperparámetros se desea que experimente y qué valores probar, con esto Grid Search evaluará todas las combinaciones posibles de valores de hiperparámetros, utilizando la validación cruzada (Geron, 2019).

El autor (Geron, 2019) también menciona que cuando no se tiene idea del valor que debería tener un hiperparámetro, un enfoque simple es probar potencias consecutivas de 10, este es el criterio que se usara en la experimentación.

## Isolation forest

Isolation forest aísla las anomalías más cerca de la raíz del árbol en comparación con los puntos normales. Esta característica única permite a Isolation forest construir modelos parciales (a diferencia de los modelos completos en la creación de perfiles) y emplear solo una pequeña proporción de datos de entrenamiento para construir modelos efectivos (Liu, Ting, & Zhou, 2008).

# METODOLOGÍA Y EXPERIMENTACIÓN

Para la experimentación se hizo uso de la librería scikit learn de Python.

## Análisis de datos

El dataset de sitios de localización de proteínas cuenta con 1484 registros, teniendo una columna como variable dependiente y las otras se pueden considerar como variables independientes. Este dataset no cuenta con datos nulos.

Analizando la columna que clasifica estos datos, se puede observar en la figura 1 una distribución de datos que se agrupan más a ciertos valores, como es el caso del MIT, NUC y CYT; siendo las otras clases más pequeñas.

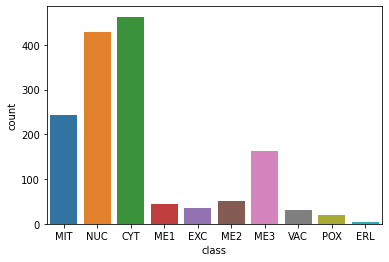


Fig. 1. Conteo de filas del dataset en base a su columna de clasificación

Analizando la primera columna con el nombre de “sequence\_name”, se puede observar en la figura 2 que de 1484 datos, 1462 datos son únicos y debido a que la primera columna no es una variable cuantitativa, ni se puede agrupar por la gran cantidad de tipos que hay, se decidió no tomar en cuenta esta columna en el modelo.

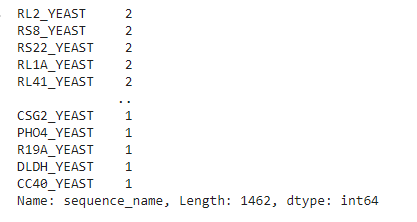


Fig. 2. Conteo de valores únicos en la columna de “sequence\_name”,

## Aplicación del modelo de Random Forest

Para una primera aplicación del modelo de Random Forest se usó como único parámetro el “n\_estimators” igual a 200. Como se observa en la figura 3, el accuray obtenido es 0.63, se puede concluir que este modelo no tiene una muy buena precisión.

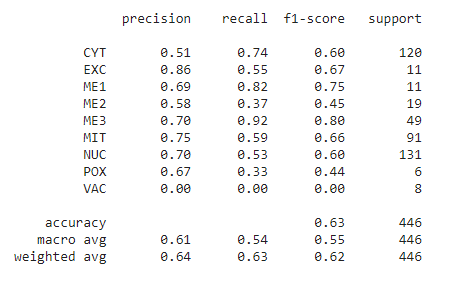


Fig. 3. Reporte del primer modelo de Random Forest en base a la matriz de confución

## Afinando hyperparámetros del modelo de Random Forest con Grid Search

Intentando mejorar la precisión del modelo, se aplicará el método de Grid Search, con la configuración mostrada en la figura 4.

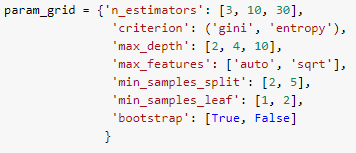


Fig. 4. Configuración de los hyperparámetros para el método de Grid Search

Como se puede observar en la figura 5, este método no logró mejorar la precisión de este modelo.

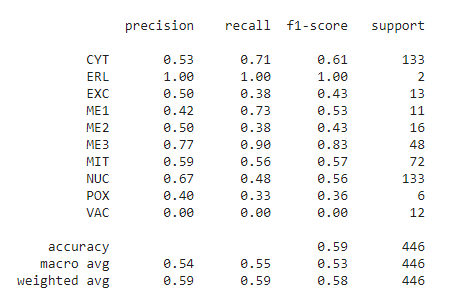


Fig. 5. Reporte del modelo de Random Forest usando Grid Search

Como se puede observar en la figura 5,

## Aplicación de K-means

Dado que no se logró una buena precisión con el modelo con Random Forest, se buscará clasificar los datos con el método de K-means.

Luego de aplicar el modelo se observa en la figura 6 que la clasificación resultante es más homogénea, a comparación de la clasificación con datos reales presentes en el dataset.

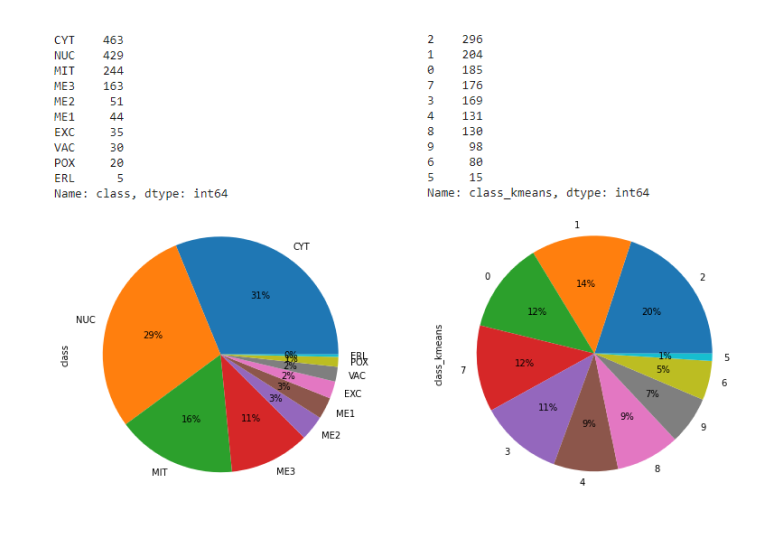


Fig. 6. Conteo de datos en base a la clasificación del dataset vs el conteo en base a la clasificación resultante del modelo de K-means

Intentando relacionar las clases resultantes del modelo con K-means, se observa en la figura 7 que en el estado actual, es muy difícil usar estas clases del dataset para medir la precisión del modelo con K-means

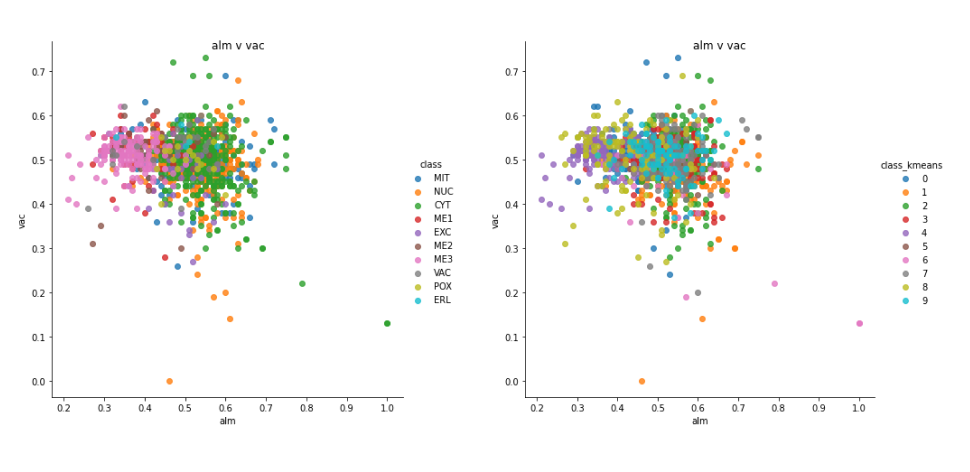


Fig. 7. Gráfica de puntos de las variables independientes alm y vac en base a las clases del dataset vs las del modelo K-means

## Reducción de clases

Debido a los resultados obtenidos anteriormente, y ya que 3 clases representan alrededor el 80% de los datos, se probará usar las clases más concurrentes: 'CYT','NUC','MIT'.

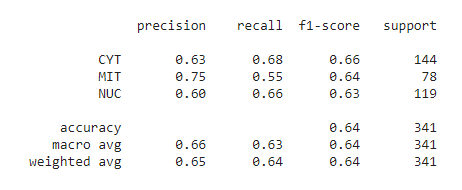


Fig. 8. Reporte del modelo de Random Forest usando 3 clases

Aplicando el método de Random Forest a esta data reducida se puede observar en la figura 8 que la precisión del modelo no mejora, pues se sigue obteniendo un accuracy de 0.64.

**Prueba con Isolation forest**

Aplicando el motodo de Isolation Forest teniendo en cuenta un 10% como posible data de contaminación, como se observa en la figura 9, se obtiene un resultado similar.

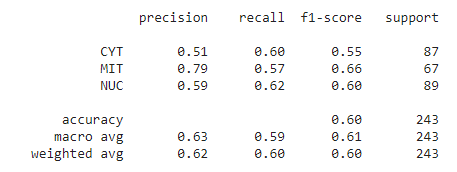


Fig. 9. Reporte del modelo de Random Forest usando 3 clases

## Comparación de la clasificación de la data con 3 clases con el modelo de K-means

Como se puede observar en la figura 10, ya se puede relacionar en cierta parte las clases del modelo con k-means y las clases reales reducidas a 3, sin embargo, no es certero, aun sigue siendo poco recomendable medir la precisión del modelo de k-means en esta data.

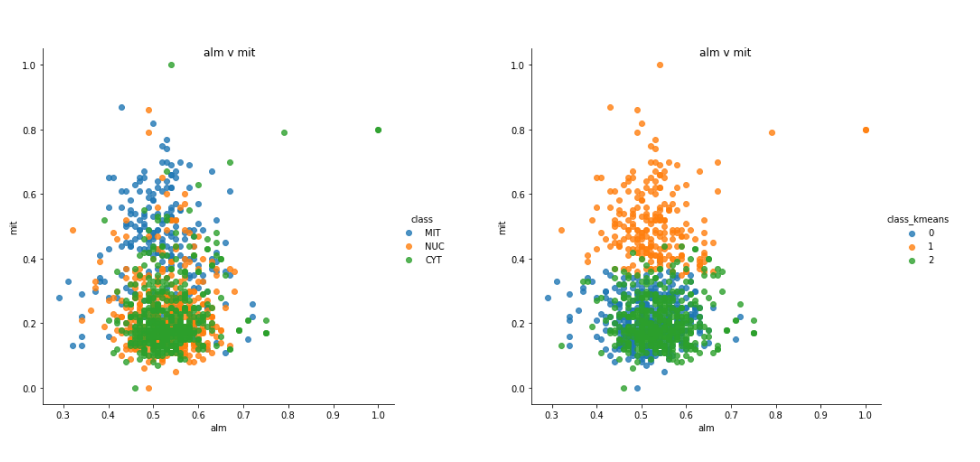


Fig. 10. Gráfica de puntos de las variables independientes alm y mit en base a las clases del dataset vs las del modelo K-means

# CONCLUSIONES

Luego del análisis aplicando los modelos de Random Forest y K-means se llega a la conclusión de que no se logra obtener una buena precisión para el modelo de Random Forest, y si se emplea métodos de aprendizajes no supervisados como k-means, las clases presentes en la data, no se logran relacionar con las predicciones del modelo, por lo que no es posible medir la precisión del modelo de k-means, también se intuye que hay un porcentaje elevado de data contaminada, pues aplicando isolation forest la precisión no mejoró, se recomienda probar con un data set más extenso y con más variables que aporten a la clasificación.

**Link del repositorio:**

[**https://colab.research.google.com/drive/1cQjFLoIcCM8Bjvg\_9YhlJjziChZXtrb5?usp=sharing**](https://colab.research.google.com/drive/1cQjFLoIcCM8Bjvg_9YhlJjziChZXtrb5?usp=sharing)

[**https://github.com/alvaroenrique/ML\_course-project3**](https://github.com/alvaroenrique/ML_course-project3)

**Referencias**

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Geron, A. (2019). *Hands–On Machine Learning with Scikit–Learn and TensorFlow 2nd edition*. O’Reilly Media.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., … Zhang, X. (2011). K-Means Clustering. In *Encyclopedia of Machine Learning* (pp. 563–564). https://doi.org/10.1007/978-0-387-30164-8\_425

Nakai, K. (1996). UMI Machine Learning Repository: Yeast Data Set. Retrieved December 11, 2020, from https://archive.ics.uci.edu/ml/datasets/Yeast