CVC DATA

UNIVERSITYHACK 2024

DATATHON

Predicción de Antígenos

En una empresa industrial de biotecnología donde se producen antígenos para el desarrollo de vacunas quieren mejorar el proceso de fábricación.

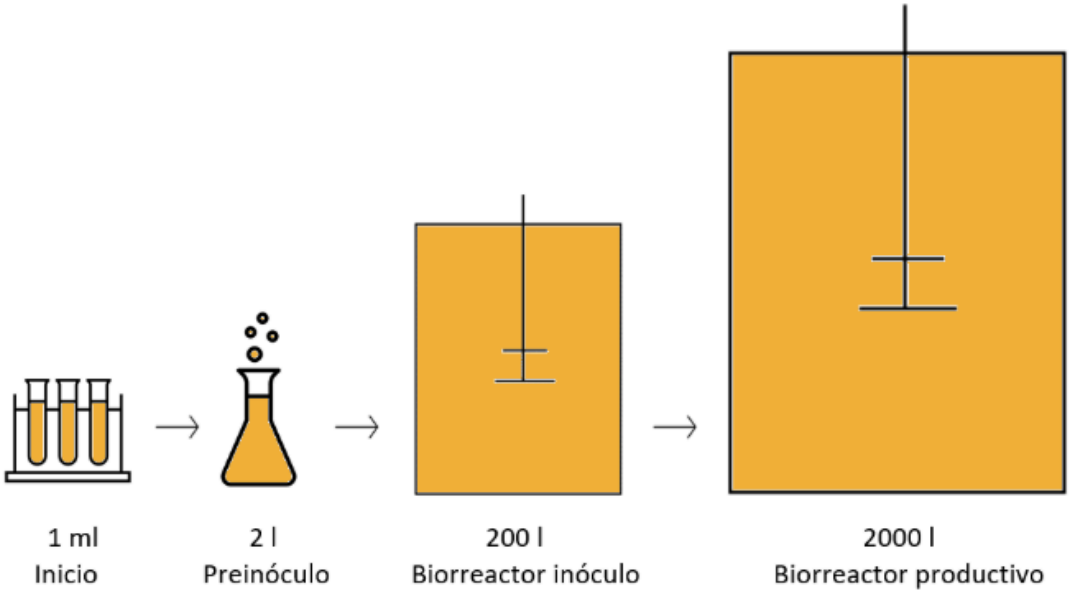
CARLOS OLIVER
VICTOR NOGUERA
CARLOS PORTILLA

ÍNDICE

Trabajo Desarrollado

Hemos desarrollado un modelo de **Machine Learning supervisado** para la predicción de antígenos en vacunas, utilizando técnicas de regresión. A pesar de contar con un número elevado de **features**, disponíamos de un **número limitado de muestras** para entrenar el modelo, lo que hizo que la métrica **RMSE** fuera la más adecuada para evaluar el desempeño del modelo. El modelo fue implementado en **Python** utilizando dos enfoques principales: **Random Forest** y **XGBRegressor**.

Pero antes de ello, tuvimos que investigar sobre el proceso de la fabricación de vacunas y entender mejor los datos.



Análisis Exploratorio

Durante el análisis exploratorio, observamos que los **outliers** presentes en los datos afectaban significativamente el rendimiento del modelo, especialmente debido a que el error en el **RMSE** es cuadrático. Además de los **outliers**, identificamos que algunas variables, como las **temperaturas** y **humedades**, influían de manera considerable en la capacidad predictiva del modelo.

A continuación, realizamos un estudio sobre el impacto de estas variables en el modelo:

Impacto de Temperaturas y Humedades en la Fabricación de Vacunas

Las **temperaturas** y **humedades** son factores críticos en la fabricación de vacunas debido a su impacto directo en la estabilidad y eficacia de los componentes activos. Las vacunas suelen contener proteínas, antígenos o virus inactivos, que son sensibles a las condiciones ambientales. A continuación, se detallan los aspectos más relevantes:

1. **Estabilidad de los Componentes Activos**:

Las temperaturas y humedades inadecuadas pueden desnaturalizar los componentes activos de la vacuna, lo que afecta su capacidad para generar una respuesta inmune eficaz. Esto resalta la importancia de mantener condiciones ambientales controladas para preservar la estabilidad de las vacunas.

2. **Almacenamiento y Transporte**:

Las vacunas deben mantenerse dentro de un rango de temperatura controlada, generalmente entre **2°C y 8°C**, durante su almacenamiento y transporte. Cualquier desviación de estas condiciones puede alterar la eficacia del producto. Además, un nivel inapropiado de humedad puede comprometer la integridad del envase y del producto.

3. **Procesos de Fabricación**:

Durante la producción de vacunas, es esencial mantener condiciones ambientales controladas para garantizar que las reacciones químicas necesarias se lleven a cabo correctamente. Temperaturas y humedades inadecuadas pueden afectar procesos clave, como la **liofilización**, alterando la calidad del producto final.

4. **Control de la Contaminación**:

Un ambiente con **alta humedad** favorece el crecimiento de microorganismos no deseados, lo que podría contaminar las vacunas. Además, temperaturas extremas pueden dificultar el control de calidad y la seguridad del producto.

Este formato está diseñado para ser claro y bien estructurado, facilitando la comprensión de cómo las **temperaturas** y **humedades** afectan tanto al modelo como al proceso de fabricación de vacunas.

Manipulación de Variables y Argumentación

Para la manipulación de datos, utilizamos la librería **Pandas** y realizamos pruebas con diferentes estrategias, como el uso de **medianas**, **medias**, **proporciones**, **índices** y **periodos de días**. Tras experimentar con estos métodos, optamos por utilizar **medianas** para reducir los outliers, ya que las **medias** no fueron efectivas y dejaban persistir valores atípicos que seguían afectando el modelo.

Justificación de la Selección del Modelo

El modelo **XGBoost** fue seleccionado debido a su rendimiento superior, obteniendo un **RMSE de 205**, en comparación con **Random Forest**, que no penalizaba tan eficazmente los datos atípicos. **XGBoost** resultó ser el modelo más adecuado, ya que mejoró la capacidad de generalización frente a los outliers y proporcionó mejores resultados en términos de precisión.