Conclusiones, Dificultades y Caminos Abiertos

Conclusiones

Debido a la pequeña muestra de datos, los tiempos de entrenamiento y ajuste de los modelos han sido muy manejables. Si hubiéramos trabajado con una cantidad real de datos, los tiempos de procesamiento habrían sido significativamente más largos, y habría sido necesario realizar un estudio más profundo para reducir estos tiempos optimizando la búsqueda de hiperparámetros y el entrenamiento de los modelos.

En cuanto al estudio de los modelos no supervisados, se concluye que los outliers, que normalmente se desecharían para limpiar la muestra, proporcionan información valiosa sobre la presencia de cáncer. El problema es que estos outliers se superponen para cada variable, lo que reduce su capacidad informativa y predicción durante la clusterización de los datos.

Hemos desarrollado la aplicación 'CancerDetector.py', diseñada para probar los modelos entrenados para la detección del cáncer y la identificación del tipo de cáncer.

Funcionalidad de la Aplicación

La aplicación permite poner a prueba diversos modelos de aprendizaje supervisado para predecir la presencia de cáncer basándose en varias biomarcas. Los usuarios pueden seleccionar entre diferentes modelos de aprendizaje supervisado, organizados de menor a mayor efectividad:

1. Regresión Lineal
2. Regresión Logística
3. Random Forest
4. KNN (K-Nearest Neighbors)
5. AdaBoost
6. Gradient Boosting
7. Voting Classifier

Predicción del Tipo de Cáncer

En la pestaña 'Predicción del Tipo de Cáncer', la aplicación permite predecir el tipo de cáncer basado en los parámetros introducidos. El modelo encargado de esta predicción es un conglomerado de varios modelos de aprendizaje supervisado, combinados mediante la técnica de Voting Classifier.

Los modelos utilizados en el 'Voting Classifier 2' son:

1. Random Forest
2. Gradient Boosting
3. LightGBM
4. XGBoost

Introducción de Datos

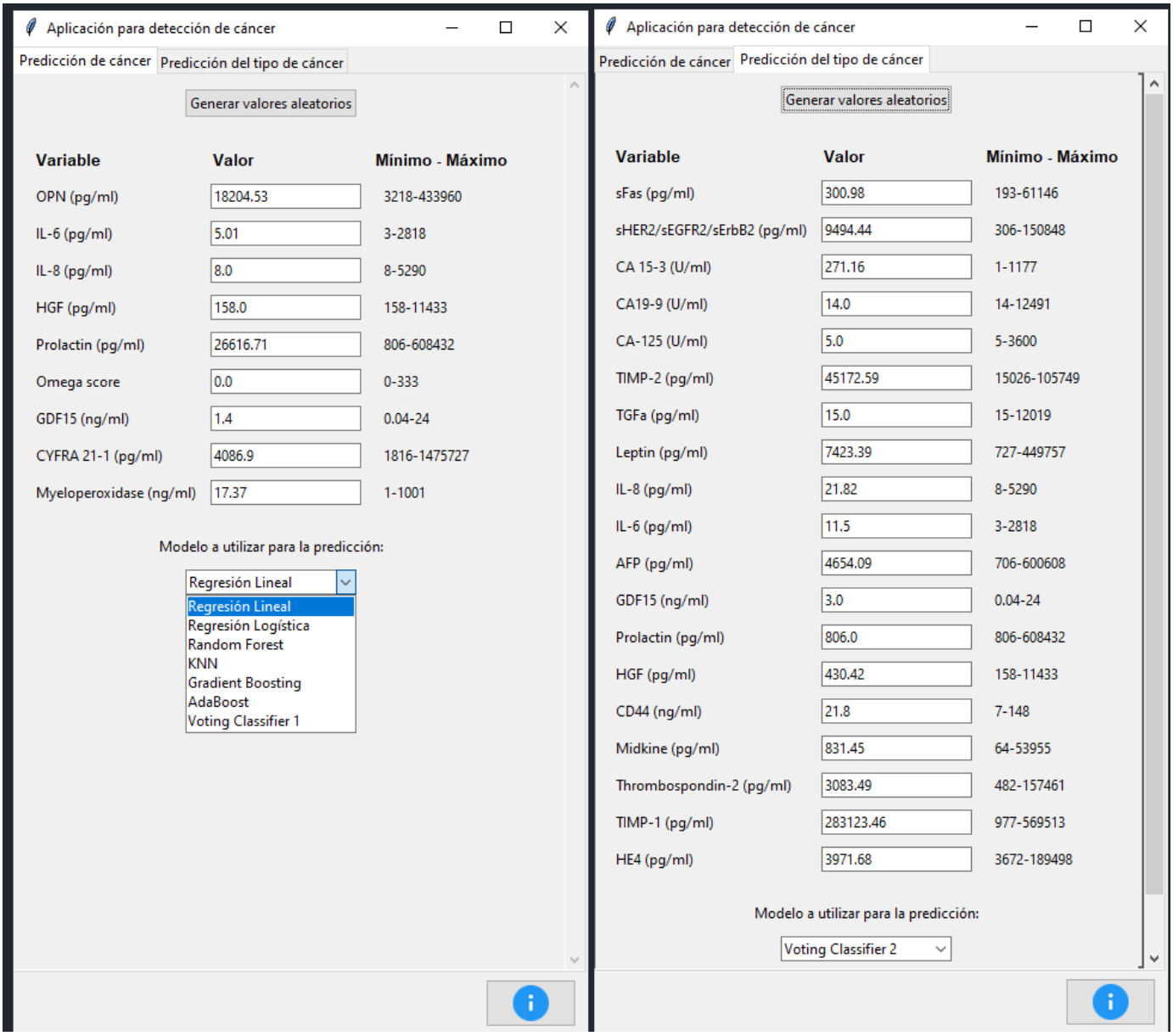
Los usuarios tienen la opción de ingresar manualmente los valores de las variables. Alternativamente, pueden generar valores aleatorios para las biomarcas utilizando un modelo de generación de datos sintéticos, CTGAN, que está entrenado para crear muestras de datos siguiendo una distribución similar a la de los datos originales del estudio.

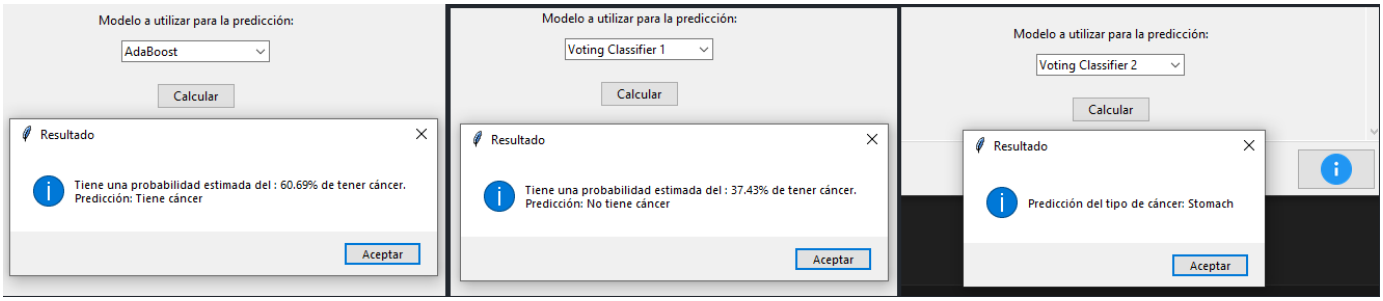
Resultado de la Predicción

Una vez ingresados los datos y seleccionado el modelo, al presionar el botón 'Calcular', se mostrará en una ventana emergente la probabilidad de tener cáncer, que variará según el modelo elegido, junto con la predicción correspondiente.

Conclusión Final de la Aplicación

'CancerDetector.py' es una herramienta versátil que permite evaluar distintos modelos de aprendizaje supervisado para la detección y clasificación del cáncer, ofreciendo opciones tanto para la entrada manual de datos como para la generación de datos sintéticos, y facilitando así un análisis comprensivo y personalizado.





Dificultades

A continuación, presentamos las dificultades que hemos encontrado durante el desarrollo del proyecto:

Falta de Datos Reales

La escasez de datos auténticos y diversos es un obstáculo significativo para validar los modelos en escenarios reales de producción. Sin una cantidad suficiente de datos representativos, es difícil evaluar cómo se comportarán los modelos en la práctica. Esta limitación puede afectar la precisión y la aplicabilidad de los modelos, impidiendo una evaluación exhaustiva de su rendimiento y capacidad para generalizar a nuevos datos.

Dependencia del Entorno de Desarrollo

Durante el entrenamiento y guardado de los modelos, se ha observado que deben ser cargados bajo el mismo entorno de máquina virtual (VM) en el que fueron entrenados. Esta dependencia del entorno original crea una fragilidad en el sistema, dificultando la portabilidad y escalabilidad de los modelos. La necesidad de replicar exactamente el entorno de desarrollo puede limitar la flexibilidad y la capacidad de implementar los modelos en diferentes plataformas o contextos operativos.

Predicciones Incorrectas Iniciales

En un primer momento, las predicciones de la aplicación eran todas positivas, detectando cáncer en todos los casos. Este problema surgió porque no se tuvo en cuenta que los registros introducidos por el usuario debían preprocesarse para asemejarse al formato de los datos de entrenamiento. Sin un preprocesamiento adecuado, las discrepancias entre los datos de entrada y los datos de salida, es decir, utilizados para entrenar los modelos, pueden conducir a resultados incorrectos. Esta experiencia subraya la importancia de un pipeline de preprocesamiento robusto y consistente.

Limitaciones de Recursos Computacionales

El entrenamiento de modelos complejos, como los basados en técnicas de machine learning avanzadas, requiere una cantidad considerable de recursos computacionales. Esto puede incluir tiempo de procesamiento, memoria y capacidad de almacenamiento, lo cual puede ser un desafío especialmente sin acceso a infraestructura avanzada, como clústeres de computación o servicios en la nube. Estas limitaciones pueden ralentizar el desarrollo y la implementación de los modelos, incrementando los costos y afectando la eficiencia del proceso.

Complejidad en la Interpretación de Resultados

Los modelos de machine learning, especialmente aquellos considerados "caja negra" como las redes neuronales profundas, son difíciles de interpretar y justificar. Esta falta de transparencia puede generar desafíos en la validación de los modelos y en la confianza de los resultados obtenidos. La interpretabilidad es crucial en aplicaciones médicas, donde las decisiones basadas en modelos deben ser comprensibles y justificables para los profesionales de la salud.

Gestión de la Calidad de los Datos

La calidad de los datos de entrada es crucial para el desarrollo de modelos fiables. Datos incompletos, ruidosos o sesgados pueden llevar a resultados incorrectos y modelos poco fiables. La limpieza y preparación adecuada de los datos es un proceso complejo y laborioso que requiere una atención meticulosa. Este proceso incluye la imputación de valores nulos, la normalización y estandarización de los datos, y la eliminación de outliers.

Evolución de los Datos y el Modelo

Los datos pueden cambiar con el tiempo debido a diversos factores, como cambios en la población estudiada, la aparición de nuevas tecnologías de diagnóstico o tratamientos, y cambios en las prácticas de recolección de datos. Estos cambios pueden hacer que los modelos entrenados pierdan precisión y relevancia. Es necesario implementar estrategias de mantenimiento y actualización continua de los modelos para asegurar que sigan siendo precisos y útiles a medida que los datos evolucionan.

Consideraciones Éticas y de Privacidad

El manejo de datos sensibles, especialmente en aplicaciones médicas, requiere una atención cuidadosa a las normativas de privacidad y ética, como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea y la Ley de Portabilidad y Responsabilidad de Seguros de Salud (HIPAA) en Estados Unidos. Estas normativas pueden complicar el acceso y uso de ciertos conjuntos de datos, limitando la capacidad para realizar análisis completos y entrenamientos de modelos sin comprometer la privacidad de los pacientes.

Integración con Sistemas Existentes

Integrar los modelos entrenados en los sistemas y flujos de trabajo existentes puede presentar dificultades técnicas significativas. Estos desafíos pueden incluir problemas de compatibilidad, diferencias en los formatos de datos, y limitaciones en el rendimiento. La integración exitosa requiere una cuidadosa planificación y posiblemente la adaptación de los sistemas existentes para asegurar que los modelos puedan ser implementados de manera efectiva y eficiente.

Evaluación y Validación del Modelo

Asegurar que el modelo generalice bien a nuevos datos y no esté sobreajustado a los datos de entrenamiento es un desafío constante. Se requieren técnicas robustas de validación, como la validación cruzada y el uso de conjuntos de datos de prueba independientes, para garantizar la fiabilidad y precisión de los modelos. Además, es crucial realizar un monitoreo continuo del rendimiento del modelo una vez implementado para identificar y corregir rápidamente cualquier disminución en su precisión o efectividad.

Caminos Abiertos

Calidad y Cantidad de Datos

Partimos de la premisa de que los datos del estudio son de muy buena calidad, preparados meticulosamente y seleccionando explícitamente las variables relevantes. No obstante, el problema principal es que la muestra de datos es pequeña. Para entrenar modelos más robustos y fiables, sería conveniente tener acceso a más datos o poder recolectar nuevos datos adicionales.

Generación de Datos Sintéticos

En caso de no tener acceso a nuevos datos reales, existe la opción de generar datos sintéticos utilizando el modelo CTGAN. Sin embargo, esta alternativa puede sesgar las predicciones del modelo, ya que, aunque CTGAN puede generar datos con distribuciones similares a las originales, no puede replicar completamente la compleja aleatoriedad ni las relaciones subyacentes entre los biomarcadores. Esto puede limitar la precisión y generalización de los modelos.

Procesamiento de Datos

Para el procesamiento de datos, se realizó un estudio sobre el relleno de valores nulos utilizando distintas técnicas. Dado que hay pocos registros faltantes, se optó por técnicas básicas de imputación. Sin embargo, es posible que se puedan obtener mejores resultados explorando enfoques más avanzados y personalizados para el manejo de datos faltantes.

Replicación de Modelos de Referencia

Hubiera sido beneficioso replicar los modelos CancerA1DE y CancerA2DE utilizados en el estudio original. Esto habría permitido una mejor comprensión del planteamiento, funcionamiento y rendimiento de dichos modelos, además de facilitar una comparación directa con nuestros modelos. Desafortunadamente, debido a dificultades técnicas y falta de tiempo, no fue posible realizar esta replicación.

Tendencia al Sobreajuste

Los modelos muestran una tendencia al sobreajuste debido a la limitada muestra de datos del estudio. Esta tendencia puede ser mitigada al

tener acceso a más datos de entrenamiento o utilizando técnicas de regularización más efectivas.

Ampliación y Refinamiento de Modelos

Se puede continuar el estudio ampliando la oferta de modelos y afinando los hiperparámetros de los modelos estudiados. Además, se pueden añadir más modelos competentes al Voting Classifier para hacerlo más robusto y mejorar su rendimiento.

Evaluación de Modelos No Supervisados

Tras el estudio, se concluyó que los modelos no supervisados no presentaron una mejora significativa en la predicción del cáncer. No obstante, es posible que se haya pasado por alto su potencial. Un posible enfoque es profundizar en el estudio de estos modelos y evaluar su influencia en la predicción del cáncer más detalladamente.

Variabilidad en las Predicciones

En la aplicación final, se observó que la probabilidad de predicción variaba considerablemente según los registros de entrada, a pesar de que la predicción final era la misma. Esto puede deberse al sobreajuste de los modelos, que afecta la confianza en las probabilidades asignadas. Es importante abordar esta variabilidad para asegurar que las probabilidades de predicción sean consistentes y fiables.

Conclusión final Caminos Abiertos

Para avanzar en el desarrollo de modelos de predicción de cáncer más robustos y fiables, es crucial acceder a más datos reales o generar datos sintéticos de manera cuidadosa. También es esencial mejorar el procesamiento de datos, replicar y comparar modelos de referencia, y mitigar la tendencia al sobreajuste. Además, explorar y refinar modelos adicionales, incluidos los no supervisados, puede proporcionar insights valiosos y mejorar el rendimiento general de los modelos.

Referencias:

- **Cuantificación de ADN libre en plasma sanguíneo de voluntarios sanos en una población bogotana - Salazar et all, Revista Nova**
- **Multi-cancer early detection tests: pioneering a revolution in cancer screening - Huiqin Jiang & Wei Guo, SpringerLink**