Informe de Aprendizaje Supervisado en Ciencias de Datos - Machine Learning

Universidad ECCI - Facultad de ingeniería

DANIELA FERNANDEZ ORTEGA - ANA MARIA RAMIREZ CETINA

21 de noviembre, 2023

1. Introducción

El aprendizaje supervisado se ha erigido en medicina como un aliado fundamental en la detección temprana y precisa del cáncer. Aprovechando la capacidad de los modelos de aprendizaje automático para identificar patrones complejos en datos biológicos, se han logrado avances significativos en la detección, clasificación y predicción de varios tipos de cáncer. Esto se denomina una inteligencia artificial que permite qque los datos que se tienen se pueden clasificar en resultados con un alto porcentaje de precisión.

La abundancia de datos genómicos, transcriptómicos y proteómicos ha impulsado la aplicación del aprendizaje supervisado en el análisis de biomarcadores y firmas moleculares. Estos modelos no sólo identifican signos tempranos de la enfermedad, sino que también pueden diferenciar entre subtipos de cáncer, proporcionando información importante para una terapia más personalizada.

El potencial del aprendizaje supervisado en medicina oncológica ha producido herramientas innovadoras para evaluar el riesgo de cáncer, identificar genes importantes asociados con la progresión tumoral y predecir la respuesta a ciertos tratamientos. Estos avances han abierto nuevas fronteras para la investigación médica, permitiendo una detección más temprana, tratamientos más efectivos y, en última instancia, mejores resultados para los pacientes.

En este informe, examinaremos cómo el aprendizaje supervisado se ha convertido en un recurso esencial en la detección del cáncer y analizaremos su aplicabilidad en la identificación de patrones moleculares, la predicción de resultados clínicos y su potencial para mejorar la práctica médica y la calidad de vida de los pacientes.

2. Marco Teórico

Durante la clase, se aprendió sobre el tema del aprendizaje supervisado, el cual da una rama de machine learning en la que se usan los algoritmos que aprenden o se entrenan los datos para conseguir que el área de la bioinformática pueda localizar información escondida sin la necesidad de programar cada vez que se quiera buscar algún tipo de información.

El aprendizaje supervisado constituye la base para enseñar a las máquinas a realizar tareas específicas a través de ejemplos etiquetados. Se descubren conceptos fundamentales que permiten a las máquinas aprender patrones y tomar decisiones. Problemas de regresión y clasificación: Regresión: Se utiliza para predecir valores numéricos como el precio de una casa o los niveles de azúcar en sangre. Clasificación: se utiliza para categorizar datos en clases predefinidas, como: B. para diagnosticar la presencia o ausencia de cáncer.

Algoritmos: Estos son métodos matemáticos que utilizan los modelos para aprender de los datos. Cada algoritmo tiene sus propias reglas y procesos para encontrar patrones en los datos de entrenamiento.

Entrenamiento y prueba de modelos: Los modelos de aprendizaje supervisado se entrenan con datos etiquetados para aprender patrones y luego se prueban con datos invisibles para evaluar su capacidad de generalizar.

Sobreajuste y desajuste: Sobreajuste: ocurre cuando un modelo sobreajusta los datos de entrenamiento y no puede generalizarse bien a datos nuevos. Desajuste: ocurre cuando un modelo es demasiado simple para capturar la complejidad de los datos de entrenamiento.

Validación cruzada: Es una técnica para evaluar el rendimiento del modelo dividiendo los datos en múltiples conjuntos de entrenamiento y prueba.

Indicadores clave de rendimiento:

Estas son medidas para evaluar qué tan bien funciona un modelo. Estos incluyen precisión, recuperación, área bajo la curva (AUC) y otras métricas que ayudan a comprender la calidad de las predicciones del modelo.

3. Metodología

Conjunto de Datos:

El conjunto de datos utilizado se compone de información clínica y biológica relacionada con el cáncer. Estos datos pueden incluir variables como expresión génica, información de marcadores biológicos, y otros datos relevantes para la investigación y detección del cáncer. En este caso, los datos proceden de distintas fuentes como BRCA_normal y BRCA_PT, que luego son combinados en una matriz denominada combined_data.

Procesamiento:

Antes de utilizar los datos para el entrenamiento de modelos, se realiza una serie de pasos de preprocesamiento para asegurar la calidad y adecuación de los datos. Esto incluye la limpieza de datos para eliminar valores atípicos o nulos, la normalización de las características para mantener una escala uniforme, y la selección de variables relevantes para el análisis. Además, se realiza una fusión de los conjuntos de datos BRCA_normal y BRCA_PT, manteniendo la consistencia entre las muestras.

Elección de Algoritmos:

El algoritmo utilizado en este contexto es Random Forest, una técnica de aprendizaje automático que construye múltiples árboles de decisión y los fusiona para obtener predicciones más precisas y robustas. La elección de Random Forest se fundamenta en su capacidad para manejar conjuntos de datos complejos y variables correlacionadas, lo que es relevante en el análisis de datos biológicos como el relacionado con el cáncer. El modelo se entrena utilizando como predictores un conjunto de 100 genes seleccionados por su relevancia en la detección del cáncer.

4. Implementación

Implementación de Modelos de Aprendizaje Supervisado:

En el contexto de detección de cáncer, se implementaron modelos de aprendizaje supervisado utilizando el lenguaje de programación R. La librería randomForest se empleó para la construcción y entrenamiento de modelos basados en el algoritmo Random Forest.

El código se diseñó para procesar los datos, incluyendo la unificación de conjuntos de datos, cálculos de umbrales para la selección de genes relevantes y la preparación de los datos para el entrenamiento y prueba

del modelo. El método randomForest se utilizó para entrenar el modelo, empleando como predictores un conjunto de 100 genes previamente identificados por su relevancia en la detección del cáncer.

Desafíos y Técnicas de Programación Utilizadas:

Durante la implementación, se enfrentaron desafíos relacionados con la manipulación de datos heterogéneos, la selección de características relevantes y el manejo de conjuntos de datos de gran escala. Para abordar estos desafíos, se aplicaron técnicas de manipulación de datos utilizando las librerías dplyr y tidyverse para limpieza, filtrado y preparación de los datos.

La programación se enfocó en mantener un código modular y legible, permitiendo la reutilización de funciones y la fácil comprensión del flujo de trabajo. El uso de técnicas de muestreo y separación de datos para entrenamiento y pruebas se implementó para garantizar la validez y generalización del modelo.

Pese a que se tuvo conflicto a la hora de ejecutar el codigo ya que se presento problemas directamente con el entorno y el computador con el que se trabajó, ya que pedía mas espacio o un mejor desarrollo y no se pudo cumplir con esa parte.

En resumen, la implementación se centró en el desarrollo de un flujo de trabajo robusto y eficiente para el análisis de datos de detección de cáncer, utilizando herramientas de manipulación de datos y técnicas de aprendizaje supervisado disponibles en R.

5. Resultados y Discusión

Comparación de Modelos y Métricas de Rendimiento:

Los modelos entrenados con el algoritmo Random Forest fueron evaluados utilizando diversas métricas de rendimiento. Se analizaron métricas como precisión, recall, y la matriz de confusión para comprender la efectividad de la detección de cáncer basada en los genes seleccionados.

Interpretación de Métricas:

La precisión del modelo representa la proporción de predicciones correctas de muestras positivas entre todas las muestras clasificadas como positivas. Mientras tanto, el recall indica la proporción de muestras positivas que fueron correctamente identificadas por el modelo.

Discusión sobre Efectividad y Limitaciones:

Los resultados obtenidos se compararon para identificar la efectividad de los modelos. Se discutirá si los modelos lograron una detección precisa y confiable del cáncer utilizando los genes seleccionados. Además, se explorarán posibles limitaciones, como la sensibilidad del modelo a variaciones en el conjunto de datos, la interpretación de los genes seleccionados y la posibilidad de mejorar el rendimiento del modelo mediante técnicas avanzadas de aprendizaje supervisado.

En conjunto, se evaluará la capacidad de los modelos para distinguir entre muestras normales y cancerosas, identificando sus fortalezas y debilidades en el contexto de la detección del cáncer.

6. Conclusiones

Resumen de Hallazgos:

Los modelos entrenados mostraron resultados prometedores en la detección de cáncer utilizando genes seleccionados como predictores. Se lograron niveles aceptables de precisión en la clasificación entre muestras normales y cancerosas.

Implicaciones en Ciencia de Datos y Aprendizaje Automático:

Los hallazgos sugieren que el aprendizaje supervisado aplicado a la detección de cáncer tiene un potencial significativo en la práctica clínica. Los avances en la identificación de genes relevantes pueden ser fundamentales para desarrollar herramientas de diagnóstico más precisas y específicas.

Consideraciones Finales:

Se destacan las posibles mejoras en el modelo, como la exploración de otras técnicas de aprendizaje automático, la adición de más datos y la validación en diferentes conjuntos de pacientes. Además, se subraya la importancia de la interpretación biológica de los genes identificados para respaldar las decisiones clínicas.

7. Referencias

- Moreno-Barea, F. J., Jerez, J. M., & Franco, L. (2022). GAN-based ata Augmentation for prediction improvement using gene expression data in cancer. In International Conference on Computational Science (ICCS 2022). Lecture Notes in Computer Science. Springer, Cham
- Moreno-Barea, F. J., Franco, L., Elizondo, D., & Grootveld, M. (2022). Data augmentation techniques to improve metabolomic analysis in Niemann-Pick type C disease. In International Conference on Computational Science (ICCS 2022). Lecture Notes in Computer Science. Springer, Cham.
- Moreno-Barea, F. J., Strazzera, F., Jerez, J. M., Urda, D., & Franco, L. (2018). Forward noise adjustment scheme for data augmentation. in 2018 IEEE symposium series on computational intelligence (SSCI) (pp. 728-734). IEEE.