Técnicas de Aprendizaje de Máquina
Proyecto Final

Pontificia Universidad Javeriana
Cristian Javier Diaz Alvarez

12 de mayo de 2025

Resumen

Este proyecto final en el curso de **Técnicas de Aprendizaje de Máquina** tiene como objetivo aplicar diversos modelos de clasificación para predecir la probabilidad de que una persona padezca una o varias enfermedades, basándose en sus respuestas a una encuesta de salud. Los estudiantes emplearán técnicas de preprocesamiento, selección de variables y múltiples algoritmos de aprendizaje de máquina, como **Redes Neuronales, Árboles de Decisión, SVM, K-means, Bagging y Boosting**, para comparar su rendimiento en términos de precisión, recall y F1-score. La meta es que los estudiantes comprendan las fortalezas y limitaciones de cada técnica en un contexto de clasificación multiclase y multietiqueta.

1. Introducción

El curso de **Técnicas de Aprendizaje de Máquina** ha cubierto una variedad de algoritmos y técnicas fundamentales en el área de machine learning. A lo largo del semestre, los estudiantes han estudiado algoritmos supervisados y no supervisados, técnicas de ensamble y métodos de optimización, aplicando estos conocimientos en talleres y proyectos prácticos.

Este proyecto final integra todos estos conceptos y se centra en la aplicación de técnicas de clasificación para analizar datos de salud. El objetivo es **clasificar la presencia de seis enfermedades** en personas, en función de sus respuestas a una encuesta detallada de salud y antecedentes personales. Mediante el uso de diferentes modelos de aprendizaje de máquina, los estudiantes evaluarán la precisión y relevancia de cada técnica, destacando las ventajas y limitaciones de cada una en este contexto.

2. Descripción de los Datos

El conjunto de datos contiene las respuestas de 20000 personas a una encuesta sobre aspectos de vida, salud y antecedentes familiares. La encuesta está compuesta por 30 preguntas de diferentes tipos (opción múltiple, respuestas de "Sí.º "No", y datos numéricos) que buscan proporcionar un perfil integral de cada individuo.

Las enfermedades objetivo son:

- Enfermedad cardiovascular
- Diabetes
- Asma
- Cáncer
- Obesidad
- Depresión/Ansiedad

Preguntas realizadas a los pacientes:

Pregunta	Tipo de Respuesta
1. Edad	Numérica
2. ¿Fuma actualmente?	Sí/No
3. ¿Realiza actividad física regularmente?	Sí/No
4. Nivel de actividad física	Sedentario/Moderado/Activo
5. ¿Tiene antecedentes familiares de diabetes?	Sí/No
6. ¿Ha sido diagnosticado con hipertensión?	Sí/No
7. Frecuencia de ejercicio físico semanal	0, 1-2, 3-4, 5 o más veces
8. ¿Tiene antecedentes familiares de enfermedades cardiovasculares?	Sí/No
9. Nivel de colesterol	Normal/Alto
10. ¿Consume alcohol?	Sí/No
11. ¿Sufre de estrés con frecuencia?	Sí/No
12. ¿Sufre de ansiedad o depresión?	Sí/No
13. ¿Ha tenido dificultades para dormir en el último mes?	Sí/No
14. ¿Está tomando algún medicamento regularmente?	Sí/No
15. ¿Le han realizado alguna cirugía importante?	Sí/No
16. ¿Tiene antecedentes familiares de cáncer?	Sí/No
17. ¿Ha sido diagnosticado con alguna enfermedad respiratoria?	Sí/No
18. ¿Tiene antecedentes familiares de asma?	Sí/No
19. ¿Ha tenido alguna vez problemas de tiroides?	Sí/No
20. ¿Sufre de dolores articulares frecuentes?	Sí/No
21. ¿Ha sido diagnosticado con obesidad?	Sí/No
22. ¿Tiene antecedentes familiares de obesidad?	Sí/No
23. ¿Sufre de dolores de cabeza frecuentes?	Sí/No
24. ¿Está tomando algún tipo de vitamina o suplemento?	Sí/No
25. ¿Realiza chequeos médicos periódicamente?	Sí/No
26. ¿Cuál es su nivel de satisfacción con la vida?	Bajo/Medio/Alto
27. ¿Consume alimentos ricos en azúcares regularmente?	Sí/No
28. ¿Cuántas horas de sueño tiene por noche, en promedio?	Numérica
29. ¿Ha tenido antecedentes de enfermedades gastrointestinales?	Sí/No
30. ¿Está tomando algún tipo de medicamento para el control del colesterol?	Sí/No

Tabla 1: Preguntas de la Encuesta de Salud

3. Actividades del Proyecto

El proyecto se divide en varias actividades para guiar a los estudiantes en el proceso de construcción y evaluación de modelos de clasificación.

3.1. Preprocesamiento y Exploración de Datos

- Realizar un análisis exploratorio para comprender la distribución de respuestas y detectar relaciones entre variables y enfermedades.
- Ejecutar limpieza de datos, imputación de valores y codificación de variables categóricas, según sea necesario.

3.2. Selección de Variables Relevantes

- Utilizar técnicas como análisis de correlación o reducción de dimensionalidad (por ejemplo, PCA o Algoritmo Genético) para identificar preguntas relevantes para cada enfermedad.
- Justificar la elección de variables clave en el contexto de cada enfermedad objetivo.

3.3. Construcción de Modelos de Clasificación

- Aplicar al menos cinco modelos de clasificación estudiados, como Redes Neuronales, Árboles de Decisión, SVM, K-means, Bagging y Boosting.
- Realizar el ajuste de cada modelo, evaluando los hiperparámetros más adecuados.

3.4. Evaluación Comparativa

- Evaluar el rendimiento de cada modelo mediante métricas como precisión, recall, F1-score y tiempos de entrenamiento para cada enfermedad.
- Incluir una tabla comparativa que resuma el rendimiento de cada técnica en términos de las métricas elegidas.

3.5. Análisis de Resultados y Conclusiones

- Discutir las ventajas y desventajas de cada modelo en el contexto de este problema, teniendo en cuenta la cantidad de variables, distribución de clases y características del conjunto de datos.
- Concluir cuál modelo fue más efectivo para cada enfermedad y por qué, basándose en los resultados cuantitativos y cualitativos.

4. Entregables

- **Informe escrito** (40 %): Incluye análisis de datos, selección de variables, construcción de modelos, evaluación y conclusiones comparativas.
- **Código en Python** (40 %): Incluye el código para el análisis y construcción de modelos, con explicaciones detalladas. (Jupyter Notebook)
- **Presentación** (20 %): Resumen de los hallazgos y conclusiones. Se debe subir un video de no mas de 5 minutos a Youtube con los hallazgos y conclusiones.

Este proyecto permite a los estudiantes integrar y aplicar todas las técnicas de aprendizaje de máquina estudiadas durante el curso, explorando sus aplicaciones, limitaciones y efectividad en un problema de clasificación complejo y relevante para la salud.

Es importante que validen que el notebook entregado contenga los recursos que consideren relevantes (es decir, no se borren al guardar el notebook). El video en Youtube debe ser accesible al menos con el enlace que sea compartido (no debe solicitarse acceso para poder verlo, no debe estar privado). El informe idealmente no debe tener más de 10 páginas, sea conciso en sus explicaciones, argumentos y conclusiones. El detalle debe estar en el notebook. Recuerde, no es solo ejecutar ajustes de modelos, no es solo ejecutar código.