Análisis de datos de actividad física con aprendizaje supervisado

Martín Alejandro Hermosillo García

A01634552

Campus GDL

Modelación del aprendizaje con inteligencia artificial

Grupo 301

ITESM

Lunes 06 de junio del 2022

Análisis de datos de actividad física con aprendizaje supervisado

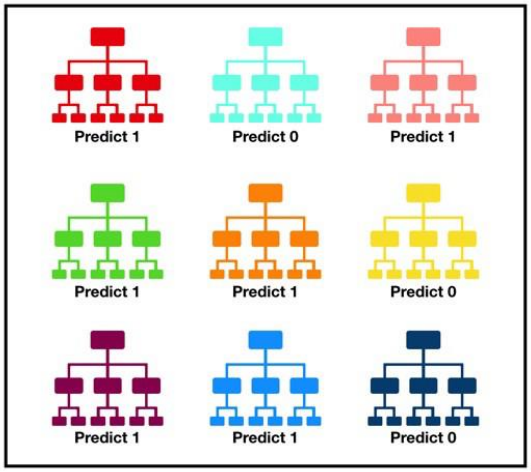
Para la evaluación de los distintos clasificadores probados, es importante primero entender el formato de los datos con los cuales se entrenaron. Con el código proveído se realizaron varios trials con dos distintos sujetos de prueba. En total se realizaron 3 trials por sujeto y en cada trial se debían realizar 4 actividades que consistían en hacer extensiones de pecho, trotar, abdominales y tocarse los dedos de los pies. En consecuencia, se leyeron 6 archivos los cuales fueron combinados para tener un solo dataframe para el entrenamiento de los modelos. La recolección de estos datos tuvo sus complicaciones debido a que en ocasiones el programa devolvía archivos txt con distintas dimensiones en cuanto a las features involucradas, lo cual imposibilitaba el análisis. Posteriormente, estos datos fueron evaluados con los modelos que se enseñan en la siguiente tabla y reportaron exactitudes que también se enseñan a continuación:

	Scores
Multi-Layer Perceptron	0.799797
KNN	0.789642
Random Forest	0.788802
RBF SVC	0.754853
Linear SVC	0.730277
Logistic Regression	0.707381

Se puede observar que los clasificadores lineales no fueron dieron tan buenos resultados como aquellos que no son lineales. Por eso, en caso de tener que entrenar más modelos para buscar alguno más óptimo sería conveniente buscar modelos no lineales. Como se mencionó en el programa efectuado en Jupyter Notebook, todos los modelos utilizados tienen hiperparámetros a seleccionar. Para ver los valores óptimos de los hiperparámetros de algún modelo, se seleccionó el perceptrón multicapa para poder obtener los mejores hiperparámetros para el modelo que ya presentaba los mejores resultados y ver si se podía mejorar. Se utilizó el GridSearchCV para la obtención de hiperparámetros y luego se volvió a hacer la evaluación del modelo. El score de la exactitud sí presentó una mejora, llegando a un valor de aproximadamente 0.8109. Al analizar la cuestión de selección de características, se decidió usar el método de RFE (Recursive Feature Elimination). Al hacer esto, se vio que para llegar a los scores de exactitud reportados, basta con

tener 8 de las 19 características que la base de datos poseía. Esto hubiera sido muy conveniente para reducir los tiempos computaciones de algunos métodos que si eran más tardados.

Para poder entender el funcionamiento de un modelo Random Forest, primero es importante entender el funcionamiento de Decision Trees. El funcionamiento de este último tipo de clasificadores se vio en clase y por ende, no será explicado en este reporte. Un Random Forest, como se podría entender por el nombre, es la combinación de muchos árboles de decisión individuales que operan en conjunto y bajo un solo modelo. Las predicciones de un decisión tree (de manera individual) puede que no sean muy precisos, pero combinados las predicciones serán más precisas. Se dice que es aleatorio o random debido a que considera un subset de features/características aleatorios y porque también accede a un training set aleatorio. Básicamente, cada árbol daría una predicción de la clase y la clase con más votos se vuelve la predicción del modelo como se muestra en la figura:



Tally: Six 1s and Three 0s
Prediction: 1

En pocas palabras, el concepto fundamental del modelo de clasificación de Random Forest es la sabiduría de los grupos. Tal y como menciona Yiu (2019), “Una gran cantidad de modelos relativamente no correlacionados (árboles) que funcionan como un comité superarán a cualquiera de los modelos constituyentes individuales”. La clave al éxito de este concepto es la poca correlación entre los distintos decision trees. La razón de esto es que cada árbol se protege uno al otro de los errores individuales. Es decir, a lo mejor algunos árboles tienen una predicción de la

clasificación errónea, pero muchos otros habrán acertado, así que como grupo los árboles se inclinarán hacia la dirección correcta. (A menos que todos los árboles cometan el mismo error).
A mayor cantidad de árboles de decisión con poca correlación entre ellos, mejores predicciones hará el modelo de Random Forest.

De acuerdo con el Supervisor de la Protección de Datos Europeo (EDPS por sus siglas en inglés), los principales problemas en cuanto a datos relacionados con dispositivos móviles son la responsabilidad de aquellos que recolectan datos, el derecho de información y la seguridad de los datos. Estos tres puntos me parecen muy pertinentes ya que siento que las empresas que permiten que se use la recolección de datos a través de sus dispositivos no analiza los riesgos de esto y solo le interesa el dinero que obtienen al permitir esto. A su vez creo que las empresas que recolectan dichos datos se aprovechan de la ignorancia de la gente o de la flojera de leer los avisos de privacidad en los cuales se indica el uso de datos de los usuarios y esto debe cambiar. Las empresas que recolectan estos datos también deben hacerse responsables y cuidar dichos datos, analizando los riesgos de poseerlos para así garantizar un nivel adecuado de protección. GLS Group (2020) mencionaba que los riesgos principales asociados con la colección de datos personales son el alto costo de cumplir con todos las regulaciones de privacidad de datos del mundo, como la mayoría de las empresas no son transparentes con los datos, las fugas de datos, una definición ambigua de datos personales y que los riesgos jamás serán eliminados en su totalidad. De acuerdo con Forbes (2018), el consumo de aplicaciones móviles abre puertas ya que tantos datos nos permiten un nuevo nivel de precisión en personalización de contenido, pero el hecho de que recolectar datos se vuelva tan cotidiano solo nos indica que también debe haber un gran avance en ciberseguridad para garantizar el uso adecuado de dicha información personal.

En conclusión, creo que este ha sido el proyecto que más me ha llamado la atención, porque sentí que iba muy enfocado hacia mi carrera y aparte me dejó aplicar los conocimientos adquiridos durante este curso y el otro curso que he llevado de ciencia de datos. A su vez, considero que su aplicación va muy de la mano con el contexto del mundo moderno. Monitorear la actividad física de un individuo es benéfico para la salud del individuo, a pesar de que ser físicamente activo debería ser natural y una parte del día. Pero monitorear la actividad física permite proponerse metas más realistas. Greene (2019), menciona que proponerse metas realistas y anotarlas (cosa que al

tenerlas monitoreadas está hecha en automático), incrementa las probabilidades de cumplir dicha meta por un factor de tres. El fenómeno del Internet of Things solo seguirá creciendo y su uso en el monitoreo de la actividad física es algo que ya existe. La posibilidad de poder monitorear señales como el ritmo cardíaco y el ritmo respiratorio con los sensores, permite hacer muchas medidas que nos sirven para saber el estado de salud sin necesidad de ir a un consultorio médico. A su vez causa interés y sirve como motivador para mantenerse saludable. Esta es otra razón por la cual si considero el monitoreo benéfico. Hay una gran cantidad de aplicaciones que también involucran monitoreo que son benéficos para la salud. Entre estas se incluye el monitoreo de la presión arterial para prevenir enfermedades cardiovasculares como la hipertensión que es una de las enfermedades no transmisibles más prevalentes en el mundo, y asimismo monitorear a aquellos que ya padecen una enfermedad de estas para checar que sus niveles queden dentro de rango y no empeore su estado de salud, el monitoreo de glucosa para aquellos con diabetes y/o aquellos con riesgo de padecerlo y el monitoreo de consumo de alimentos y bebidas para llevarlo a su médico/nutriólogo para mantener una dieta balanceada y rica en nutrientes. Todas esta información, variables y señales se pueden monitorear con dispositivos móviles lo cual simplemente nos enseña el enorme avance médico que ha habido en el sector tecnológico en las últimas décadas.

Bibliografía

Garnett, O. (2018). Beware the Power and Pitfalls of Mobile Data Collection. *Forbes*. Recuperado de https://www.forbes.com/sites/forbestechcouncil/2018/04/18/beware-the-power-and-pitfalls-of-mobile-data-collection/?sh=15f8e9340dc4

Greene, B. (2019). The Pyschology of Writing Down Goals. *New Tech Northwest.* Recuperado de https://www.newtechnorthwest.com/the-psychology-of-writing-down-goals/

Mobile Devices. (2022). *European Data Protection Supervisor*. Recuperado de https://edps.europa.eu/data-protection/data-protection/reference-library/mobile-devices_en

What are the risks associated with collecting personal data? (2020). *GLS Group*. Recuperado de https://www.gls.global/en/startupresources/what-are-the-risks-with-collecting-personal-data

Yiu, T. (2019). Understanding Random Forest. *TowardsDataScience.* Recuperado de https://towardsdatascience.com/understanding-random-forest-58381e0602d2