Proyecto: Ciencia de datos aplicada al análisis de datos fisiológicos

"A Wearable Exam Stress Dataset for Predicting Cognitive Performance in Real-World Settings"

Ayline Sánchez Canales

Taller integrador de ciencias de datos

14 de noviembre de 2024

Resultados del análisis exploratorio de los datos (EDA).

En la fase de análisis exploratorio de datos (EDA), se analizaron los datos fisiológicos de los estudiantes recopilados durante los exámenes académicos para identificar patrones, tendencias y posibles anomalías. Los datos incluyen diversas señales fisiológicas como la temperatura corporal (TEMP), la actividad electrodérmica (EDA), la frecuencia cardíaca (HR), el volumen de pulso sanguíneo (BVP), los intervalos entre latidos (IBI) y el movimiento registrado por el acelerómetro (ACC). Además, se consideraron las calificaciones obtenidas por los estudiantes en los exámenes como medida de su rendimiento académico.

1. Análisis de Patrones y Tendencias

Durante el EDA, se observaron patrones significativos en los datos fisiológicos de los estudiantes:

- **Temperatura Corporal (TEMP)**: Se identificaron cambios en la temperatura corporal durante los exámenes, que variaban entre los diferentes momentos del examen (inicio, medio y final). En general, algunos estudiantes presentaron un leve aumento de la temperatura a medida que avanzaba el examen, lo cual podría estar relacionado con una mayor respuesta al estrés.
- **Actividad Electrodérmica (EDA)**: Se observó un incremento significativo en la actividad electrodérmica durante ciertos momentos del examen, lo cual se asocia con niveles elevados de estrés. Estos picos tienden a coincidir con eventos clave, como el inicio del examen o preguntas particularmente difíciles.
- **Frecuencia Cardíaca (HR) y Volumen de Pulso Sanguíneo (BVP)**: Ambas señales mostraron variaciones que reflejan la respuesta emocional de los estudiantes ante situaciones de presión. Se observaron aumentos abruptos en la frecuencia cardíaca que coincidieron con los picos de la actividad electrodérmica, sugiriendo momentos de alto estrés.
- **Acelerómetro (ACC)**: El movimiento registrado indicó que algunos estudiantes presentaban mayor inquietud durante los exámenes, lo cual podría estar relacionado con la ansiedad. Los patrones de movimiento variaban significativamente entre estudiantes, con algunos mostrando comportamientos de inquietud durante la mayor parte del examen.
- **Intervalos entre Latidos (IBI)**: Los cambios en los intervalos entre latidos mostraron cómo variaba la variabilidad del ritmo cardíaco (HRV) a lo largo del examen. Los momentos de menor HRV coincidieron con los picos de actividad

electrodérmica, lo cual indica una menor capacidad de regulación autonómica durante los momentos más estresantes.

2. Relación entre Variables Fisiológicas y Rendimiento Académico

Para evaluar la relación entre las señales fisiológicas y el rendimiento académico, se realizaron gráficos de dispersión entre las variables fisiológicas promedio y las calificaciones obtenidas por los estudiantes. Los resultados mostraron lo siguiente:

- **Relación entre la Frecuencia Cardíaca y Calificaciones**: Los estudiantes con mayores picos de frecuencia cardíaca durante el examen tendieron a obtener calificaciones más bajas, lo cual sugiere que el alto nivel de estrés afectó negativamente su desempeño.
- **Relación entre la Actividad Electrodérmica y Calificaciones**: Se observó una correlación entre los altos niveles de actividad electrodérmica y el bajo rendimiento académico, indicando que los estudiantes con mayor respuesta al estrés emocional tendieron a tener más dificultades para concentrarse y rendir adecuadamente.

3. Anomalías y Hallazgos Relevantes

Durante el análisis, se identificaron algunas anomalías, como lecturas inconsistentes en los datos del acelerómetro de ciertos estudiantes, posiblemente debido a errores de medición o interferencias. Estas lecturas fueron filtradas y corregidas durante el preprocesamiento para asegurar la calidad de los datos utilizados en el modelado predictivo.

4. Visualizaciones y Conclusiones Iniciales

Las visualizaciones generadas durante el análisis incluyeron gráficos de tendencia para cada señal fisiológica, así como gráficos de barras para representar el rendimiento académico de cada estudiante. Estas visualizaciones ayudaron a identificar las relaciones entre las señales fisiológicas y el desempeño en los exámenes, destacando la importancia del control del estrés para mejorar el rendimiento académico.

Justificación y selección de los modelos utilizados.

Para abordar el desafío de predecir el rendimiento académico de los estudiantes a partir de los datos fisiológicos, se seleccionaron varios modelos de aprendizaje automático y se justificó su elección con base en las características de los datos y el objetivo del estudio.

1. Selección de Modelos

Se eligieron cuatro modelos principales para el análisis predictivo: Regresión Lineal, K-Nearest Neighbors (KNN), Support Vector Regressor (SVR) y Random Forest Regressor. Cada uno de estos modelos tiene características específicas que los hacen adecuados para el problema planteado:

- Regresión Lineal: Este modelo fue seleccionado debido a su simplicidad y capacidad para establecer una línea base en el rendimiento predictivo. La regresión lineal permite identificar relaciones lineales entre las señales fisiológicas y las calificaciones, proporcionando una primera aproximación para entender qué tan bien estas variables se correlacionan.
- K-Nearest Neighbors (KNN): KNN fue seleccionado por su capacidad para capturar relaciones no lineales entre las variables fisiológicas y el rendimiento académico. Este modelo es particularmente útil cuando la estructura de los datos no sigue un patrón lineal claro, lo cual es común en los datos fisiológicos debido a la variabilidad individual entre los estudiantes.
- Support Vector Regressor (SVR): SVR se utilizó debido a su efectividad en manejar relaciones complejas y encontrar el mejor hiperplano que minimice el error. Este modelo es adecuado para trabajar con datos de alta dimensionalidad y es capaz de manejar la no linealidad presente en las señales fisiológicas.
- Random Forest Regressor: El modelo de Random Forest fue seleccionado por su capacidad para manejar grandes cantidades de datos y capturar relaciones complejas entre las variables. Además, su capacidad para estimar la importancia de cada variable nos permitió identificar cuáles señales fisiológicas tenían un mayor impacto en el rendimiento académico de los estudiantes.

2. Justificación de la Elección de los Modelos

La selección de estos modelos se basó en la necesidad de capturar tanto relaciones lineales como no lineales presentes en los datos fisiológicos. La Regresión Lineal proporcionó una línea base para comparar el desempeño de otros modelos más complejos. KNN y SVR fueron elegidos para explorar la capacidad predictiva de relaciones no lineales, considerando la naturaleza altamente variable de las respuestas fisiológicas al estrés. Random Forest, por otro lado, fue seleccionado por su capacidad de manejar la complejidad y proporcionar información sobre la importancia relativa de cada característica fisiológica.

Estos modelos se entrenaron y evaluaron utilizando técnicas de validación cruzada para asegurar la generalización de los resultados. Además, se utilizaron métricas como el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2) para comparar el rendimiento de los modelos y seleccionar el más adecuado para el problema planteado.

Evaluación de los modelos en función de las métricas apropiadas.

Para evaluar el rendimiento de los modelos de aprendizaje automático utilizados en este estudio, se seleccionaron varias métricas adecuadas. A continuación, se describen cada una de estas métricas y cómo se aplicaron en el código:

- **Error Cuadrático Medio (MSE)**: El MSE mide el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales. Se utiliza para evaluar la precisión del modelo, penalizando fuertemente los errores grandes. En el código, se implementó utilizando la función mean_squared_error() de sklearn.metrics. Esta métrica se calculó después de entrenar cada modelo para medir qué tan bien se aproximaban las predicciones a los valores reales de calificación.

  Fórmula:

  MSE = (1/n) * Σ(predicción_i - valor_real_i)^2

- **Error Absoluto Medio (MAE)**: El MAE es una métrica que calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales. A diferencia del MSE, el MAE no penaliza de manera desproporcionada los errores grandes, lo cual lo hace útil para interpretar el error promedio de las predicciones. En el código, el MAE se calculó usando la función mean_absolute_error() de sklearn.metrics y se aplicó para tener una idea clara de cuán lejos, en promedio, estaban las predicciones de los valores reales.

  Fórmula:

  MAE = (1/n) * Σ|predicción_i - valor_real_i|

- **Coeficiente de Determinación (R^2)**: El R^2 indica el porcentaje de varianza en la variable dependiente que es explicada por las variables independientes del modelo. Un valor de R^2 cercano a 1 indica un buen ajuste del modelo. En el código, se utilizó la función r2_score() de sklearn.metrics para evaluar cada modelo y entender qué tan bien los datos fisiológicos lograban predecir el rendimiento académico. Esta métrica fue esencial para evaluar qué porcentaje

de la variación en las calificaciones se podía explicar por los datos fisiológicos recopilados.

Fórmula:

$R^2 = 1 - (\Sigma(error\_residual)^2 / \Sigma(error\_total)^2)$

Aplicación en el Código

Después de entrenar cada modelo, se dividieron los datos en conjuntos de entrenamiento y prueba para evitar sobreajuste y asegurar una buena generalización. Para cada modelo, se calcularon las métricas de rendimiento mencionadas utilizando el conjunto de datos de prueba:

```
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Predicciones del modelo
y_pred = model.predict(X_test)

# Calcular MSE, MAE y R^2
y_mse = mean_squared_error(y_test, y_pred)
y_mae = mean_absolute_error(y_test, y_pred)
y_r2 = r2_score(y_test, y_pred)

print(f"MSE: {y_mse}")
print(f"MAE: {y_mae}")
print(f"R^2: {y_r2}")
```

Estas métricas ayudaron a determinar qué modelo ofrecía la mejor capacidad predictiva y cuál era el más adecuado para entender las relaciones complejas entre las señales fisiológicas y el rendimiento académico de los estudiantes.

Resultados con relación al problema original.

En la implementación de los modelos de aprendizaje automático, se logró entrenar y evaluar algunos de los modelos seleccionados. Sin embargo, debido a la gran cantidad de datos y las limitaciones de recursos computacionales, no fue posible realizar una prueba completa y concluyente de todos los modelos. Mi equipo no soportó la carga de trabajo necesaria para procesar los datos de manera eficiente, lo cual dificultó obtener resultados definitivos sobre el rendimiento predictivo.

A pesar de estas limitaciones, el análisis exploratorio de datos (EDA) proporcionó información valiosa sobre la relación entre las señales fisiológicas y el rendimiento

académico. Se pudieron identificar patrones significativos que indican cómo las respuestas fisiológicas al estrés influyen en el desempeño de los estudiantes durante los exámenes. Estos hallazgos preliminares sientan una base importante para futuras investigaciones y la mejora en el modelado predictivo.

Sigo investigando esta área, buscando optimizar los procesos y trabajar con herramientas que me permitan manejar grandes volúmenes de datos de manera más efectiva. Lo obtenido en el EDA me ayudó a tener algunas conclusiones importantes sobre la influencia del estrés fisiológico en el rendimiento académico, y continuaré explorando métodos para mejorar las predicciones con recursos más adecuados.