Documentación del Proyecto: Redes Sociales vs Productividad

Documentación del Proyecto: Redes Sociales vs Productividad

1. Descripción del Caso de Uso Elegido

Temática

El proyecto "Redes Sociales vs Productividad" se basa en un dataset de Kaggle (Redes Sociales vs Productividad) con 30.000 filas y 19 columnas, que incluye datos sobre uso de redes sociales, productividad, salud mental, sueño, estrés y hábitos de estudio (e.g., edad, género, tiempo en redes, notificaciones).

Qué Queremos Resolver

Se desarrolla un modelo de regresión basado en inteligencia artificial para predecir la productividad (actual_productivity_score, 0-10) usando variables como:

- daily_social_media_time (minutos).

- number_of_notifications.
- breaks_during_work (interrupciones).
- screen_time_before_sleep.
- uses_focus_apps.

Este modelo se integra con un cuadro de mando en Power BI para análisis interactivo por género, tipo de trabajo y plataforma preferida.

Aplicación Práctica

- **Autoconocimiento**: Ayudar a usuarios a entender el impacto de sus hábitos digitales.
- **Optimización personal**: Sugerir ajustes para mejorar productividad.
- **Herramientas de bienestar digital**: Desarrollar apps con predicciones de IA.
- **Investigación organizacional**: Analizar patrones en equipos.

La IA ofrece predicciones personalizadas, mientras Power BI visualiza los datos para decisiones informadas.

---

2. Esquema Funcional/Técnico de los Desarrollos Realizados (Aprox. 1 página)

Elementos Tecnológicos

- **Herramientas**: Power BI Desktop (junio 2025) para visualización; Google Colab/Python con pandas, numpy, scikit-learn, shap, y wandb para IA.
- **Fuentes de datos**:
  - DimUser (demográficos: age, gender, IdUser).
  - DimJob (categorías: Education, Student, Finance, Health, IT, Unemployed).
  - DimSocialMedia (plataformas: TikTok, Instagram, Twitter, Telegram, Facebook).
  - FactProductivity (métricas: actual_productivity_score, number_of_notifications, daily_social_media_time).

- **Medidas DAX**:
  - AverageProductivity = AVERAGE(FactProductivity[actual_productivity_score])
  - AverageTimeRRSS = AVERAGE(FactProductivity[daily_social_media_time])
  - AgeRange = SWITCH(TRUE(), DimUser[age] <= 30, "20-30", DimUser[age] <= 40, "31-40", "41+")
- **Visualizaciones**: KPIs, barras, líneas, donas.
- **Filtros**: job_type, gender, social_platform_preference.
- **Inteligencia Artificial**: Procesamiento y modelos (ver punto 3)

Relaciones entre Elementos

- **Flujo**: Datos de dfProductivity_cleaned.csv se preprocesan, entrenan modelos IA, y se integran en Power BI vía DimUser/FactProductivity.
- **Interacciones**: Slicers filtran datos, actualizando visuales; predicciones IA (pendientes) se integrarán como KPI.
- **Dependencias**: Visuales dependen de DAX; IA depende de datos exportados de Colab.

Páginas Desarrolladas

- **Dashboard Principal**: KPIs (Productividad, Estrés, Sueño, Tiempo RRSS), gráficos (Horas vs. Sueño, Notificaciones, Productividad por Edad, Estrés).
- **Productivity VS Social media**: KPI Tiempo RRSS, Notificaciones, Horas Desconectado.
- **Productividad Máxima**: KPI combinado, KPIs secundarios, Notificaciones, Tiempo Offline.

---

3. Resolución de los diferentes aspectos del proyecto

1. Ingesta de Datos

El proceso comienza con la ingesta del dataset “Redes Sociales vs Productividad” desde un csv (’social_media_vs_productivity.csv’) usando PySpark en Google Colab.

Se crea una sesión Spark (SparkSession) para gestionar el procesamiento distribuido, asegurando escalabilidad. Los datos incluyen variables demográficas (e.g., age, gender), hábitos digitales (e.g., daily_social_media_time, number_of_notifications), y métricas de bienestar (e.g., actual_productivity_score). Se verifica el tamaño (30.000 filas, 19 columnas) y el esquema, confirmando tipos iniciales (integer, string, double, boolean).

2. Calidad del Data

Garantizar la calidad de los datos es esencial para asegurar que los resultados del modelo predictivo sean fiables y representativos.

Limpieza y Conversión

Numéricas a DecimalType(4,2)/IntegerType, booleanos a 0/1.

Tratamiento de Nulos

Para abordar este problema, se realizaron dos tipos de imputación:

1. Imputación por KNN (K-Nearest Neighbors): Este método consiste en buscar los registros más similares por distancia euclidiana (vecinos cercanos) y utilizar sus valores para completar los datos faltantes.

   En este caso:

   - Para las variables numéricas: calcula la media de los valores de esos tres vecinos y usa ese promedio para rellenar el valor faltante.
   - Los posibles nulos en variables categóricas se completaron con el valor más frecuente (moda).

2. Imputación estadística (Media o Mediana)

   Teniendo en cuenta outliers se eligión entre mediana y media.

   Imputación por Mediana:

   La **mediana es más robusta** frente a estos valores extremos y refleja mejor el comportamiento "típico" de la mayoría.

Imputación por Media:

La **media es una buena representación del valor central** cuando no hay muchos outliers y permite mantener la variabilidad natural del dato.

Validación

Categóricas válidas, rangos numéricos (0-10), coherencia temporal (≤24h), outliers (IQR) conservados.

Exportación

Se generan dfProductivity_cleaned.csv (PySpark) y dfProductivity_KNN_cleaned.csv (Pandas) para comparación.

Modelado

Modelo en Power BI con esquema estrella:

- **Fuentes**: DimUser, DimJob, DimSocialMedia, DimCalendar, FactProductividad.
- **Relaciones**: 1:N entre dimensiones y hechos.
- **Visuales**: KPIs y gráficos filtrados por dimensiones.

Inclusión de la Inteligencia Artificial

- **Preprocesamiento**: StandardScaler, LabelEncoder, SelectKBest, PCA.
- **Modelos**: LinearRegression, RandomForestRegressor, GradientBoostingRegressor, MLPRegressor.
- **Validación**: Cross-validation anidada (5 outer, 3 inner folds) con GridSearchCV.
- **Métricas**: R^2, RMSE, MAE, análisis de overfitting.
- **Explicabilidad**: SHAP para importancia de características.
- **Seguimiento**: wandb y guardado con joblib/pickle.

---

4. Conclusiones

Nuestra solución es adecuada por las siguientes razones:

Relevancia del Problema

El impacto de las redes sociales en la productividad es un desafío global, respaldado por un dataset robusto de 30.000 registros que captura variables clave (tiempo en redes, notificaciones, bienestar). Al abordar esta temática, la solución responde a necesidades actuales de autoconocimiento y optimización personal, alineándose con tendencias de bienestar digital en 2025.

Efectividad del Modelo de IA

El uso de modelos como RandomForestRegressor y LinearRegression, validados con cross-validation anidada y optimizados con GridSearchCV, ofrece alta precisión (R^2 competitivo, RMSE bajo). SHAP proporciona explainability, permitiendo a los usuarios entender las influencias (e.g., daily_social_media_time) en sus scores, lo que refuerza la confianza en las predicciones.

Usabilidad y Valor Añadido

La integración en Power BI con un esquema estrella (dimensiones como DimJob, DimSocialMedia) y visuales interactivos (KPIs, gráficos) permite análisis detallados por segmentación (género, trabajo). Esto apoya decisiones informadas, desde ajustes personales hasta estrategias organizacionales, con un potencial inmediato para apps de bienestar digital.

Adaptabilidad Futura

La solución, desarrollada con herramientas abiertas (Colab, Power BI) y persistencia (joblib), es adaptable a nuevos datos o variables, asegurando relevancia a largo plazo. Esto la posiciona como una herramienta versátil frente a la evolución de hábitos digitales.

En resumen, nuestra solución combina datos sólidos, modelos precisos y una interfaz práctica, ofreciendo un enfoque integral y escalable para mejorar la productividad en el contexto actual de junio 2025.