**Guía de Proceso de Desarrollo para el Proyecto Final: Análisis Integral de Datos sobre el Vino**

**1. Definición del Proyecto**

* **Objetivo principal:** Desarrollar una solución integral que analice datos del vino (producción, exportación, importación, maridaje) para extraer insights relevantes y construir un modelo predictivo de Machine Learning, además de integrar un componente de IA generativa.
* **Alcance:**
  + Ingeniería de datos.
  + Exploración y visualización.
  + Modelado predictivo (regresión, clasificación, clustering, etc.).
  + Implementación de IA generativa para maridaje.
  + Despliegue en la nube o entorno local.
  + Reporte y documentación final.
* **Roles y responsabilidades (equipo de 4 personas):**
  + **Data Engineer:** Recopilación e ingesta de datos, diseño de pipelines de datos.
  + **Data Analyst:** Limpieza, exploración, y visualización de datos.
  + **Machine Learning Engineer:** Diseño, entrenamiento y validación de modelos.
  + **Software Engineer/Cloud Specialist:** Despliegue en la nube y optimización.

**2. Recopilación e Ingesta de Datos**

* **Identificación de fuentes:**
  + Bases de datos públicas: OIV (Organización Internacional del Vino), FAO, Kaggle.

<https://www.oiv.int/es/what-we-do/data-discovery-report?oiv>

[OEMV - Resumen ejecutivo del Proyecto OEMV "Exportaciones de vinos europeos por colores 2023"](https://www.oemv.es/resumen-ejecutivo-del-proyecto-oemv-exportaciones-de-vinos-europeos-por-colores-2023) [Microsoft Power BI](https://app.powerbi.com/view?r=eyJrIjoiNTJlNDZlMzktMmI3OS00ODUwLTliNjAtYmExZTI3MWE0Y2ZiIiwidCI6ImE1MzEzNDFkLTMwY2ItNGEyNS05NTUyLWVjM2FhMWNmMmRkNSIsImMiOjl9&pageName=ReportSection46371864f87e3b5ac03f)

<https://www.kaggle.com/code/kabure/wine-review-s-eda-recommend-systems>

<https://www.kaggle.com/code/sudhirnl7/wine-recommender#Missing-values>

<https://www.kaggle.com/datasets/fredericqiu/vivinoallwineexportfrance/data>

<https://www.kaggle.com/datasets/carlmcbrideellis/calidad-de-vinho-verde-portugus>

<https://www.kaggle.com/datasets/nikitatkachenko/vivinoredwine>

<https://www.kaggle.com/datasets/joshuakalobbowles/vivino-wine-data-top-10-countries-exchina/data>

<https://www.kaggle.com/datasets/zynicide/wine-reviews/data>

<https://winefolly.com/> videos

* + APIs: Datos sobre exportación/importación por zonas geográficas.
  + Datos web scraping para precios o maridajes.
* **Estrategia de ingesta:**
  + Automatizar con herramientas como Python (librerías pandas, requests, BeautifulSoup) o herramientas de ETL (Airflow, Apache NiFi).
  + Almacenar los datos en un formato estructurado (bases de datos SQL/NoSQL como PostgreSQL o MongoDB).

**3. Limpieza y Exploración de Datos (EDA)**

* **Limpieza:**
  + Tratamiento de valores nulos, duplicados y outliers.
  + Normalización y estandarización de columnas (unidades, formatos de fecha).
  + Análisis de datos faltantes con técnicas como imputación o eliminación.
* **Exploración inicial:**
  + Identificar correlaciones, tendencias y patrones mediante visualizaciones.
  + Uso de herramientas como Matplotlib, Seaborn, y Plotly para gráficos avanzados.
  + Estadísticas descriptivas (media, mediana, moda, percentiles).
* **Resultados esperados del EDA:**
  + Insights sobre exportación/importación por región y tipo de uva.
  + Identificación de patrones de maridaje.

**4. Desarrollo de Modelos de Machine Learning**

* **Selección del modelo:**
  + **Regresión:** Predicción de precios del vino en función de sus características.
  + **Clasificación:** Identificación de tipos de vino (ej. tinto, blanco) basado en propiedades químicas.
  + **Clustering:** Agrupamiento por perfiles de sabor o regiones geográficas.
  + **Series temporales:** Predicción de demanda o precios por estación.
* **Pipeline de desarrollo:**
  + División de datos en conjuntos de entrenamiento y prueba.
  + Feature engineering: Creación de nuevas variables relevantes.
  + Selección de modelos (Regresión Lineal, Random Forest, XGBoost, etc.).
  + Evaluación y validación (métricas como RMSE, precisión, F1-score).
* **Automatización:** Uso de scikit-learn, PyCaret, o MLFlow para pruebas iterativas.

**5. Integración de IA Generativa**

* **Uso propuesto:**
  + Un modelo de lenguaje como GPT para sugerir maridajes basados en las características del vino y platos disponibles.
* **Pipeline de integración:**
  + Entrenamiento adicional (fine-tuning) de un modelo base como GPT en datos específicos de maridaje.
  + Implementación con APIs como OpenAI o Hugging Face.
  + Flujo de interacción: Entrada de características del vino y sugerencias generativas como salida.

**6. Despliegue**

* **Entorno:**
  + Nube: AWS, Google Cloud, o Azure (uso de S3, EC2, y RDS para almacenamiento y procesamiento).
  + Local: Docker para crear entornos reproducibles.
* **Herramientas:**
  + Backend: Flask, FastAPI.
  + Dashboard: Streamlit, Power BI o Tableau.
  + Configuración de CI/CD: GitHub Actions, Jenkins.

**7. Reporte y Documentación Final**

* **Contenido:**
  + Resumen ejecutivo.
  + Metodología del proyecto.
  + Insights clave obtenidos del análisis.
  + Resultados del modelo de ML y métricas.
  + Ejemplo práctico del modelo generativo.
  + Recomendaciones y conclusiones.
* **Formato:**
  + Presentación: PowerPoint, con visualizaciones interactivas.
  + Documento técnico: Markdown, Jupyter Notebook o formato PDF detallado.
  + Código: Documentado y almacenado en un repositorio público (GitHub).

**8. Buenas Prácticas**

* **Gestión de proyectos:** Uso de metodologías ágiles (Scrum o Kanban).
* **Control de versiones:** Git.
* **Reproducibilidad:** Documentar dependencias (requirements.txt o environment.yml).
* **Evaluación:** Iteración basada en feedback continuo.

**1. Fuentes de Información para Entrenamiento**

**Bases de datos públicas y especializadas**

1. **Wine & Food Pairing Databases:**
   * **Open Food Facts:** Contiene datos abiertos de alimentos que puedes usar para correlacionar con tipos de vino.
   * **Vivino:** Información sobre vinos, tipos, características, y maridajes populares.
   * **Wine.com API:** Ofrece información sobre vinos y recomendaciones de maridaje.
   * **Wine Enthusiast:** Listas de maridaje recomendadas basadas en tipos de vinos.
2. **Bases de datos culinarias:**
   * **Edamam Food and Recipe API:** Proporciona información sobre recetas, ingredientes y maridajes.
   * **Spoonacular API:** Incluye datos sobre recetas, ingredientes, y posibles combinaciones.
   * **Epicurious y AllRecipes:** Con datasets de recetas que puedes extraer mediante web scraping.
3. **Libros y publicaciones digitales:**
   * "The Wine Bible" de Karen MacNeil.
   * "What to Drink with What You Eat" de Andrew Dornenburg y Karen Page.
   * "Wine Folly" (sitio web y libros).

**Web scraping de sitios especializados**

* **Guías de maridaje online:** Sitios como [Wine Folly](https://winefolly.com) o [Food and Wine](https://www.foodandwine.com) son excelentes recursos.
* **Blogs y foros especializados:** Reddit (r/wine, r/foodpairing), Wine Spectator.

**Datasets ya estructurados:**

* **Kaggle:**
  + Dataset de análisis sensorial de vinos.
  + Dataset de reseñas y calificaciones de vinos con notas de sabor.
  + Datos de recetas y sabores.

**IA y bases preentrenadas:**

* **Hugging Face Datasets:**
  + Busca datasets relacionados con alimentos, recetas y vinos.
  + Datasets de lenguaje culinario para enriquecer vocabulario y patrones.

**2. Estrategias para el Fine-Tuning**

**Preparación de los datos:**

1. **Estructura de los datos:**
   * Crear pares de entrada-salida:
     + Entrada: Descripción del vino (ej. "Vino tinto, con notas de frutos rojos y toques de vainilla").
     + Salida: Recomendación de plato (ej. "Ideal para acompañar carnes rojas asadas o quesos maduros como el gouda").
   * Categorías sugeridas:
     + Tipo de vino: Tinto, blanco, rosado, espumoso.
     + Notas de sabor: Frutales, especiadas, terrosas.
     + Atributos del plato: Ingredientes principales, método de preparación.
2. **Curación de datos:**
   * Elimina duplicados o datos irrelevantes.
   * Uniformiza los formatos (tipos de vino, categorías de platos).
3. **Etiquetado manual:** Si es necesario, usar herramientas como Label Studio para etiquetar combinaciones personalizadas.

**Proceso de fine-tuning:**

1. **Modelo base:** Usa un modelo preentrenado como GPT-3.5 o GPT-J.
2. **Herramientas:**
   * **OpenAI Fine-tuning:** Carga tu dataset en formato JSONL y utiliza las guías de OpenAI.
   * **Transformers de Hugging Face:** Usa un modelo preentrenado como GPT-2 o T5 y realiza el fine-tuning con PyTorch o TensorFlow.
3. **Configuración:**
   * Ajusta hiperparámetros para minimizar el overfitting.
   * Divide el dataset (80% entrenamiento, 20% validación).

**3. Creación de la API**

**Arquitectura:**

* **Backend:**
  + Usa **FastAPI** o **Flask** para construir endpoints.
  + Endpoint sugerido: /suggest-pairing
    - Entrada: Características del vino (tipo, notas de sabor, región).
    - Salida: Plato recomendado y descripción del maridaje.

**Pipeline de predicción:**

1. Recibe la entrada del usuario (JSON).
2. Procesa las características y genera un prompt para el modelo de lenguaje.
3. Llama al modelo de GPT (fine-tuned) para obtener la salida generativa.
4. Devuelve el resultado en formato JSON.

**Herramientas recomendadas:**

* **Implementación local:** Usa Docker para empaquetar la aplicación.
* **Despliegue en la nube:**
  + **AWS Lambda** o **Google Cloud Functions** para APIs serverless.
  + **Streamlit** para un frontend interactivo.
  + **Hugging Face Spaces** para hospedar el modelo directamente.

**Ejemplo de estructura de API con FastAPI:**

python

CopiarEditar

from fastapi import FastAPI

from transformers import pipeline

app = FastAPI()

model = pipeline("text-generation", model="fine-tuned-wine-pairing")

@app.post("/suggest-pairing")

def suggest\_pairing(wine: dict):

wine\_description = wine.get("description", "")

suggestion = model(f"Suggest a food pairing for: {wine\_description}")

return {"wine": wine\_description, "suggestion": suggestion[0]['generated\_text']}

**4. Evaluación del Modelo Generativo**

* **Métricas cualitativas:**
  + Revisión por expertos en vinos y maridajes.
  + Encuestas a usuarios finales para validar la utilidad de las sugerencias.
* **Pruebas A/B:**
  + Comparar el modelo generativo con reglas predefinidas de maridaje.

**5. Recursos Complementarios**

* **Curso de OpenAI Fine-tuning:** Aprende los pasos para entrenar modelos personalizados.
* **Comunidades y foros:** Participa en comunidades de datos y vinos para obtener feedback y más datasets.