Imagen que contiene Forma 
 
Descripción generada automáticamente

**PROYECTO SGE**

|  |
| --- |
| CFGS Desarrollo de Aplicaciones Multiplataforma  Informática y Comunicaciones |

|  |
| --- |
| **APLICACIÓN CON FASTAPI** |
| ***Año:*** *2025*  ***Fecha de presentación:*** *30/05/2025* |

|  |  |
| --- | --- |
| **Nombre y Apellidos: Mario Gangoso Ferreras**  **Email: mario.ganfer@educa.jcyl.es** |  |

|  |
| --- |
| 1. **Investigación conceptos despliegue aplicaciones:**  * Utilizando este enlace, indicar las definiciones que se piden a continuación:   <https://raul-profesor.github.io/Despliegue/debian_teoria/>   * + VPS – Sus siglas significan Virtual Private Server y es un tipo de alojamiento web que proporciona una partición privada y recursos dedicados dentro de un servidor compartidos con otros usuarios.   + SSH – Es un protocolo de red que permite acceder y administrar sistemas de forma remota a través de una conexión cifrada.   + Servidor web – Sistema informático que aloja y sirve contenido web a través de internet.   + Protocolo HTTP, métodos - El protocolo de transferencia de hipertexto (HTTP, Hypertext Transfer Protocol) es el motor que da vida a Internet, ya que es la base para la web.   Metodos:   * **GET**: se utiliza para solicitar cualquier tipo de información o recurso al servidor. Cada vez que se pulsa sobre un enlace o se teclea directamente a una URL se usa este comando. Como resultado, el servidor HTTP enviará el recurso correspondiente. * **HEAD**: se utiliza para solicitar información sobre el recurso: su tamaño, su tipo, su fecha de modificación… Es usado por los gestores de cachés de páginas o los servidores proxy, para conocer cuándo es necesario actualizar la copia que se mantiene del recurso. Con HEAD se podrá comprobar la última fecha de modificación de un recurso antes de traer una nueva copia del mismo. * **POST**: sirve para enviar información al servidor, por ejemplo, los datos contenidos en un formulario. El servidor pasará esta información a un proceso encargado de su tratamiento. * **OPTIONS**: Devuelve los métodos HTTP que el servidor soporta para una URL específica. Esto puede ser utilizado para comprobar la funcionalidad de un servidor web mediante petición en lugar de un recurso específico. * **DELETE**: sirve para eliminar un recurso especificado en la URL, aunque pocas veces sera permitido por un servidor web. * **TRACE**: comando que permite hacer un sondeo para saber todos los dispositivos de la red por los que pasa nuestra petición. Así podremos descubrir si la petición pasa a través dispositivos intermedios o proxys antes de llegar al servidor Web. * **PUT**: puede verse como el comando inverso a GET. Nos permite escribir datos en el servidor o, lo que es lo mismo, poner un recurso en la URL que se especifique. Si el recurso no existe lo crea sino lo reemplaza. La diferencia con POST puede ser algo confusa; mientras que POST está orientado a la creación de nuevos contenidos, PUT está más orientado a la actualización de los mismos (aunque también podría crearlos).   + Apache, nginx – Son dos servidores de código abierto que se utilizan para entregar contenido web a los usuarios a través de internet.   + Indicar 3 lenguajes de programación del lado del servidor   PHP, Python, Ruby.   * Utilizando este enlace, indicar las siguientes definiciones:   <https://openwebinars.net/blog/tensorflow-keras-fundamentos/>   * + Tensorflow - Es una biblioteca open source desarrollada por Google brain para llevar a cabo tareas de machine learning y deep learning. Es una de las librerías de aprendizaje automático más utilizadas a nivel industrial debido a su flexibilidad y escalabilidad.   + Keras - Es una biblioteca de código abierto para el aprendizaje automático y el deep learning. Fue desarrollada por François Chollet y actualmente es mantenida por Google.   Se caracteriza por ser una biblioteca de alto nivel, fácil de usar y flexible, que permite a los desarrolladores crear y entrenar modelos de aprendizaje automático de manera rápida y eficiente.   * Buscar las siguientes definiciones en este enlace:   <https://la.mathworks.com/discovery/lstm.html>   * + Modelo LSTM - Son redes neuronales recurrentes (RNN) especializadas diseñadas para procesar y recordar datos secuenciales, como series de tiempo o texto. * Buscar las definiciones de:   + numpy - Es una biblioteca de Python de código abierto fundamental para la computación científica, especialmente en áreas como ciencia de datos, ingeniería y matemáticas. Permite trabajar con objetos llamados arreglos multidimensionales, que son como matrices, y ofrece una gran cantidad de funciones matemáticas de alto nivel para operar con ellos de manera eficiente.   + matplotlib - Es una biblioteca de visualización de datos en Python que permite crear gráficos de diferentes tipos, como histogramas, diagramas de barras, gráficos de dispersión, entre otros, con unas pocas líneas de código.   + Google Colab - Es un servicio gratuito basado en la nube que permite escribir y ejecutar código Python, especialmente útil para tareas de aprendizaje automático y análisis de datos. |
| 1. **Creación de un modelo LSTM con TensorFlow:**  * Utilizando este enlace, realizar lo que se pide a continuación: <https://github.com/josejuansanchez/modelo-lstm-tensorflow> * Leer el apartado: **modelo-lstm-tensorflow** * Acceder al **notebook** del **contenido del repositorio** y comprender cómo funciona Google Colab * En otra pestaña del navegador, abrir Google Colab: <https://colab.research.google.com/> y realizar lo siguiente:   + Crear un cuaderno nuevo y cambiarle el nombre: modeloLSTM\_SGE2425.ipynb   + El siguiente enlace lleva a mi cuaderno.   <https://colab.research.google.com/drive/1kSQfcLPCpA-oB6knD8-Y8ursZFIzHJSi?usp=sharing> |

**Segunda parte del proyecto**

1. **Descripción general del proyecto**

Desarrollar una API REST que permita clasificar preguntas automáticamente en distintas categorías predefinidas: OK, AYUDA y SERVICIO\_TECNICO, utilizando un modelo de machine learning entrenado previamente con TensorFlow.

Además de la clasificación, se ha integrado un sistema de autenticación mediante tokens JWT, lo que permite proteger determinados endpoints y garantizar que solo usuarios autenticados puedan acceder a ciertas funcionalidades, como registrar nuevas predicciones en la base de datos.

El proyecto también ha incluido:

* Conexión e integración con una base de datos PostgreSQL.
* Persistencia de las predicciones realizadas.
* Implementación de pruebas unitarias para asegurar la funcionalidad básica del sistema.
* Uso de Swagger UI para la documentación y prueba interactiva de la API.
  1. **Entorno de trabajo (tecnologías de desarrollo y herramientas)**
* Lenguaje de programación: Python
* Framework backend: FastAPI (para construcción de la API REST)
* Machine Learning: TensorFlow (modelo entrenado previamente y cargado en tiempo de ejecución)
* ORM y Base de datos: SQLAlchemy para la conexión con una base de datos PostgreSQL
* Autenticación: JWT (JSON Web Token), usando la librería python-jose
* Pruebas: Pytest
* Contenedores y despliegue: Docker y Docker Compose para contenerizar el proyecto y los servicios (como la base de datos)
* Editor de código: Visual Studio Code
* Documentación automática: Swagger UI, generado automáticamente por FastAPI
* Control de versiones: Git

1. **Documentación técnica: análisis, diseño, implementación, pruebas y despliegue**
   1. **Análisis del sistema (funcionalidades básicas de la aplicación)**

La API ofrece las siguientes funcionalidades principales:

* POST /predict: Clasifica una pregunta enviada en el cuerpo del request y la guarda en la base de datos. Esta ruta está protegida por autenticación JWT.
* GET /ping: Ruta pública para comprobar si la API está en funcionamiento.
* GET /prueba: Ruta de prueba sin lógica adicional, útil para testear conectividad.
* POST /login: Permite a un usuario autenticarse (con credenciales fijas o desde base de datos, según implementación) y obtener un token JWT para acceder a rutas protegidas.
* Las predicciones realizadas se almacenan con la pregunta original y la categoría asignada en la tabla predicciones de la base de datos PostgreSQL.
  1. **Diseño de la base de datos**

Gráfico

El contenido generado por IA puede ser incorrecto.

* 1. **Implementación**

**Requirements**

El proyecto ha sido desarrollado en Python utilizando el framework FastAPI para la creación de una API REST que permite clasificar preguntas mediante un modelo de machine learning. Se empleó TensorFlow para la carga y ejecución del modelo, y SQLAlchemy junto a PostgreSQL para la gestión de la base de datos. La validación de datos se realizó con Pydantic, y la autenticación de usuarios mediante tokens JWT con la librería python-jose. Para las pruebas unitarias se usaron pytest y httpx. El servidor Uvicorn permite ejecutar la API localmente, y todo el entorno de trabajo ha sido gestionado en Visual Studio Code sobre un entorno virtual.

Texto

El contenido generado por IA puede ser incorrecto.

**docker-compose.yml**

Para facilitar la gestión del entorno de desarrollo y despliegue, se utilizó Docker mediante un archivo docker-compose.yml. Este define dos servicios: una base de datos PostgreSQL que actúa como backend de almacenamiento, y PgAdmin como herramienta de administración visual para PostgreSQL. Ambos servicios están conectados a través de una red personalizada y configurados con variables de entorno para facilitar el acceso y la persistencia de datos mediante volúmenes.

Texto

El contenido generado por IA puede ser incorrecto.

**Main**

El archivo main.py actúa como punto de entrada principal de la API. Se encarga de configurar y lanzar la aplicación FastAPI, estableciendo metadatos como el título, la descripción y la versión del servicio. También crea las tablas necesarias en la base de datos a partir de los modelos definidos con SQLAlchemy, e incluye los routers de las rutas principales: predict para realizar predicciones de clasificación de preguntas, y auth para el sistema de autenticación mediante JWT. Además, define una ruta raíz (/) que devuelve un mensaje simple, y configura el servidor de desarrollo con Uvicorn para ejecutar la API en el puerto 8000.

Texto

El contenido generado por IA puede ser incorrecto.

**App/db/Database**

El archivo database.py configura la conexión con la base de datos PostgreSQL usando SQLAlchemy. Define la URL de conexión con las credenciales y parámetros necesarios para acceder a la base de datos, y crea un motor (engine) que gestiona la conexión con el servidor. Además, se define SessionLocal, que permite crear sesiones para interactuar con la base de datos dentro de las rutas de la API. Finalmente, se declara Base, la clase base de la que heredarán todos los modelos de datos del proyecto para que puedan ser registrados correctamente en la base de datos.

Captura de pantalla de un celular

El contenido generado por IA puede ser incorrecto.

**App/db/Models**

El archivo models.py define el modelo de datos Prediccion, que representa la tabla predicciones en la base de datos PostgreSQL. Este modelo hereda de Base, que está configurada en database.py. La tabla contiene tres columnas: id (clave primaria y autoincremental), pregunta (texto de la consulta enviada por el usuario) y categoria (la clase o tipo al que pertenece la pregunta, como OK, AYUDA o SERVICIO\_TECNICO). Este modelo es fundamental para almacenar las predicciones generadas por el sistema de clasificación.

Texto

El contenido generado por IA puede ser incorrecto.

**App/auth/ jwt\_handler**

Este código implementa la creación y verificación de tokens JWT (JSON Web Tokens) para autenticar usuarios en una aplicación. Primero, define parámetros esenciales como la clave secreta (SECRET\_KEY), el algoritmo de cifrado (HS256) y el tiempo de expiración del token (30 minutos). La función create\_access\_token genera un token tomando un diccionario con datos del usuario, añade una fecha de expiración calculada con la hora actual en UTC más el tiempo configurado, y luego codifica esta información en un token seguro usando la clave y el algoritmo definidos. Por otro lado, la función verify\_token recibe un token, intenta decodificarlo y validar su integridad y autenticidad; si es válido, devuelve el contenido del token, y si no (por ejemplo, si el token es inválido o expiró), retorna None. Este mecanismo es fundamental para controlar el acceso a recursos protegidos en aplicaciones web modernas.

Texto

El contenido generado por IA puede ser incorrecto.

**App/routers/auth**

El archivo auth.py define un endpoint /login para la autenticación de usuarios utilizando FastAPI y Pydantic. La autenticación se basa en un diccionario simulado de usuarios con sus contraseñas en memoria (fake\_users\_db). Cuando un usuario envía sus credenciales, se verifica la correspondencia y, si son válidas, se genera un token JWT mediante la función create\_access\_token del módulo jwt\_handler. Este token es devuelto como respuesta para que el cliente lo use en peticiones autenticadas. En caso de credenciales incorrectas, se devuelve un error 401 Unauthorized

Texto

El contenido generado por IA puede ser incorrecto.

**App/router/predict**

Este módulo define el router /predict usando FastAPI. Aquí se implementa la lógica de autenticación con JWT mediante el esquema HTTPBearer para proteger la ruta de predicción.

* Autenticación:  
  La función get\_current\_user extrae y verifica el token JWT recibido en la cabecera Authorization. Si el token es inválido o ha expirado, lanza un error 401. Esto asegura que solo usuarios autenticados puedan usar la API.
* Dependencias:  
  Se usa Depends para inyectar la sesión de base de datos y la validación del usuario en el endpoint.
* Endpoint /predict:  
  Recibe un JSON con la pregunta, valida el token, y llama a la función predecir\_categoria para obtener la categoría usando un modelo TensorFlow previamente entrenado. Después, guarda la predicción en la base de datos. En caso de error en la predicción o falta de clasificación, responde con errores HTTP apropiados.
* Rutas adicionales:  
  /ping para verificar que la API está activa, y /prueba como ruta auxiliar de prueba.

Este diseño modular con dependencias y manejo de errores facilita la extensión y mantenimiento de la API, mientras protege las operaciones sensibles con autenticación.

Texto

El contenido generado por IA puede ser incorrecto.

**App/squemas/predict**

El esquema PreguntaRequest está definido con Pydantic y representa la estructura de los datos que la API espera recibir para la clasificación de preguntas. Contiene un único campo pregunta de tipo cadena (str), con validación incorporada que exige que la pregunta tenga entre 5 y 200 caracteres. Esta validación ayuda a asegurar que las entradas sean apropiadas para el modelo de clasificación.

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

**App/model**

Esta sección del código se encarga de la carga y uso del modelo de clasificación basado en TensorFlow. Primero, se carga un modelo previamente entrenado desde el directorio saved\_model/modelo.keras. Luego, se prepara un tokenizer con la clase TextVectorization de TensorFlow, que convierte las preguntas en secuencias numéricas basadas en un vocabulario cargado desde vocabulary.txt. Además, se leen las etiquetas de clasificación desde un archivo JSON (etiquetas.json) y se crea un diccionario inverso para mapear índices a etiquetas.

La función predecir\_categoria recibe una pregunta en formato texto, la preprocesa usando el tokenizer, realiza una predicción con el modelo y devuelve la etiqueta correspondiente a la categoría más probable. Si no encuentra una etiqueta válida, devuelve "NO\_ENTIENDO".

Texto

El contenido generado por IA puede ser incorrecto.

**Test/test\_predict**

Este código sirve para verificar el correcto funcionamiento del endpoint /predict. Primero, se ajusta la ruta del sistema (sys.path) para asegurar que se puedan importar correctamente los módulos del proyecto desde cualquier ubicación. Luego, se crea un cliente de prueba (TestClient) que simula peticiones HTTP hacia la aplicación. Para realizar peticiones autenticadas, se define la función obtener\_token, que envía una solicitud al endpoint /login con credenciales predefinidas y extrae el token JWT de la respuesta. Este token se utiliza en los encabezados (headers) de las peticiones de prueba. A continuación, se definen tres funciones de test: test\_predict\_success, que verifica que una pregunta válida retorna una respuesta correcta con categoría y texto; test\_predict\_empty, que prueba el comportamiento cuando se envía una pregunta vacía (esperando un error de validación 422); y test\_predict\_missing\_field, que evalúa la respuesta cuando el campo pregunta no está presente (también esperando un 422). Estas pruebas ayudan a asegurar la calidad y estabilidad del endpoint /predict.

Texto

El contenido generado por IA puede ser incorrecto.

* 1. **Pruebas**

Realizar y documentar un mínimo de 2 pruebas: una con swagger y otra con pgadmin para ver el contenido de la tabla de la base de datos

Prueba de login

<http://127.0.0.1:8000/docs>

Busca la ruta POST /login.

Pulsa sobre ella y haz clic en Try it out.

Rellena el formulario con las credenciales correctas: Admin y admin123

Haz clic en Execute.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

En la respuesta, copiar el campo "access\_token" que te devuelve el servidor.

Texto

El contenido generado por IA puede ser incorrecto.

Hacer clic en el botón Authorize

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

En el campo que aparece, escribe el token con el formato: Bearer y el token copiado anteriormente.

Pulsa Authorize y luego Close.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

El contenido generado por IA puede ser incorrecto.

Aquí podemos ver que ha ido todo bien. Ahora podemos hacer cualquier pregunta al predict.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Abrir en el navegador <http://localhost:5050>

Iniciar sesión con los datos que definiste en el docker-compose.yml

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

Agregar un nuevo servidor

* Haz clic derecho sobre "Servidores" (barra izquierda).
* Selecciona "Crear" → "Servidor...".
* Nombre: fastapi\_postgres

##### Pestaña Conexión

##### Rellena con los datos de tu base de datos definidos en docker-compose.yml:

* + 1. **Host / nombre del servidor**: db
    2. **Puerto**: 5432
    3. **Mantenimiento / Base de datos**: fastapi\_db
    4. **Nombre de usuario**: fastapi\_user
    5. **Contraseña**: fastapi\_pass
    6. Marca la opción **"Guardar contraseña"**

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Ya conectado:

Desplegar Servidores → fastapi\_postgres → Bases de datos → fastapi\_db → Esquemas → public → Tablas.

Ahí está la tabla predicciones y sus columnas.

Imagen que contiene Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

* 1. **Despliegue de la aplicación**

Este proyecto se ejecuta en un servidor local, utilizando contenedores Docker para levantar tanto la base de datos PostgreSQL como la herramienta de administración pgAdmin. La API desarrollada con FastAPI se ejecuta localmente a través de uvicorn, permitiendo pruebas y desarrollo en un entorno controlado. Todo el sistema está preparado para ser desplegado fácilmente en un servidor remoto si se desea escalar o publicar la solución en producción.

Manuales

* 1. **Manual de usuario:**

Breve explicación de cómo se usa la API con capturas de pantalla

Aquí están las rutas principales que aparecen en Swagger.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

Endpoints de prueba abiertos para verificar que la API funciona tanto /ping como /prueba.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

El /login permite simular un inicio de sesión con un usuario y contraseña que hemos insertado en el código y devuelve un Access\_token.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Aquí vemos el access\_token

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

Ahora debemos copiar el Access\_token anterior e ir a la parte superior derecha que pone Authorize.

Interfaz de usuario gráfica, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

En esta pantalla vamos a poner Bearer y el Access token que tenemos copiado, damos a authorize y vemos que si funciona.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

El contenido generado por IA puede ser incorrecto.

Ahora nos deja hacer preguntas y nos devuelve la categoría en /predict cuando antes no nos dejaba porque no teníamos acceso autorizado.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

Nos mostraría este mensaje si no está autorizada la sesión

Rectángulo

El contenido generado por IA puede ser incorrecto.

Ahora para ver si funcionan los test vamos a ir a la terminal de visual studio code y vamos a introducir el comando de ‘pytest test/test\_predict.py’ y se nos mostrará si los test funcionan correctamente.

Texto

El contenido generado por IA puede ser incorrecto.

* 1. **Manual de instalación:**

Crear y activar el entorno virtual

Python -m venv venv

Venv\Scripts\activate

Installar dependencias

Pip install -r requirements.txt y después de requirements me pide pip install psycopg2-binary

Levantar contenedores de docker

docker-compose up -d

Ejecutar la API

python main.py

1. **Conclusiones**

* Durante el desarrollo de esta aplicación se presentaron algunas dificultades, especialmente en la integración del sistema de autenticación con JWT y la protección de rutas privadas, como la del endpoint /predict. También resultó algo complejo configurar correctamente la base de datos PostgreSQL con Docker y lograr que FastAPI pudiera comunicarse con ella sin problemas.
* Otra dificultad fue la gestión de permisos y cabeceras a la hora de realizar pruebas en Swagger, ya que requería entender bien cómo se pasaban los tokens Bearer para que la API autorizara las peticiones correctamente.
* A pesar de estos retos, el grado de satisfacción con el trabajo realizado es alto, ya que se logró implementar una API funcional y segura que permite clasificar preguntas usando un modelo de machine learning entrenado previamente, además de dejar la aplicación lista para futuras mejoras o despliegues.
* Durante este proyecto se ha afianzado el uso de FastAPI, SQLAlchemy, JWT y la integración de modelos de TensorFlow en APIs REST. Además, se ha aprendido a trabajar con entornos Dockerizados y a realizar pruebas automatizadas con pytest.

**Ampliación**

Una posible ampliación del proyecto sería permitir el registro de nuevos usuarios, añadiendo persistencia en base de datos y roles de usuario.

1. **GitHub**

<https://github.com/MarioGangosoRibera/ApiChatbotPython.git>

1. **Bibliografía**

* **Documentación oficial de FastAPI**  
  https://fastapi.tiangolo.com/
* **Documentación de SQLAlchemy**  
  https://docs.sqlalchemy.org/
* **Documentación de Pydantic**  
  https://docs.pydantic.dev/
* **Documentación de TensorFlow**  
  https://www.tensorflow.org/api\_docs
* **Documentación de PostgreSQL**  
  <https://www.postgresql.org/docs/>
* **JWT (JSON Web Tokens) - jose package**  
  <https://python-jose.readthedocs.io/en/latest/>
* **Chatgpt**