1. **Propuesta para la Clasificación Automática de Correos en GA**

**Metodología de Desarrollo:**

* Utilizaremos metodología ágil con sprints menores a 4 semanas
* Enfoque iterativo con validación continua del negocio
* Desarrollo basado en datos con retroalimentación constante

|  |  |  |
| --- | --- | --- |
| Fase | Actividad | Duración Semanas |
| 1. Preparación | Recopilación de datos históricos | 2 |
|  | Limpieza inicial de datos | 1 |
| 2. Desarrollo | Implementación del modelo ML | 3 |
|  | Desarrollo de interfaz | 2 |
| 3. Pruebas | Validación del modelo | 2 |
| 4. Implementación | Despliegue | 1 |

**Fase 1: Preparación**

**Recopilación y Limpieza de Datos**

* **Objetivo**: Consolidar dataset histórico de correos
* Tareas:
  + Extracción de correos de fuentes múltiples
  + Limpieza de datos
  + Eliminación de duplicados

**Preprocesamiento de Datos**

* **Objetivo**: Transformar datos para entrenamiento de ML
* Tareas:
  + Tokenización de texto
  + Lematización
  + Vectorización (TF-IDF, Word Embeddings)
  + Balanceo de clases

**Fase 2: Desarrollo**

**Implementación Modelo ML**

* **Objetivo**: Desarrollar modelo de clasificación
* Modelos a evaluar:
  + Random Forest
  + Gradient Boosting
  + Support Vector Machines
  + Redes Neuronales

**Desarrollo de Interfaz**

* **Objetivo**: Crear interfaz de usuario para clasificación
* Componentes:
  + Dashboard de clasificación
  + Módulo de retroalimentación
  + Integración con sistemas existentes

**Fase 3: Pruebas**

**Validación y Refinamiento**

* **Objetivo**: Evaluar rendimiento del modelo
* Métricas de evaluación:
  + Precisión por categoría
  + Recall
  + F1-Score
  + Matriz de confusión

**Fase 4: Implementación (1 semana)**

**Despliegue**

* **Objetivo**: Puesta en producción
* Actividades:
  + Configuración ambiente cloud
  + Despliegue modelo
  + Monitoreo inicial
  + Capacitación de usuarios

**Recursos Requeridos**

**Equipo**

* 1 Científico de Datos Senior
* 1 Analista de negocio
* 1 Desarrollador Backend
* 1 Especialista en UI/UX

**Infraestructura Tecnológica**

* Computación en la nube
* Herramientas de ML
* Licencias de software
* Infraestructura de procesamiento

**Métricas de Éxito**

* Precisión de clasificación: >85%
* Reducción tiempo manual: 70%
* Satisfacción del usuario: >90%

**Cronograma de Retroalimentación**

* Sprint review: Semanal
* Revisión con stakeholders: Quincenal
* Ajustes de modelo: Mensual

**Ver** Sistema de Clasificación de Correos:

1. **Validación del Drift:**

La disminución en la precisión de las predicciones luego de seis meses que degrada el rendimiento del modelo sugiere que puede estar sufriendo drift. Sin embargo, es necesario realizar un análisis profundo para confirmar esta hipótesis y determinar su causa.

**Diagnóstico**

Aplicación de métodos estadísticos como prueba de hipótesis que permitan comparar distribuciones de las variables en los datos de entrenamiento como Prueba de Kolmogorov-Smirnov y si hay diferencias significativas, se puede inferir un drift.

Ver: TEORIA 2. PCA

Como es recomendado en estos casos aplicaría otros métodos como el de análisis de componentes PCA donde en una estructura de los datos en un espacio de menor dimensión y si los nuevos datos se proyectan en un espacio diferente al de los datos de entrenamiento, puede indicar un drift.

**Ver:** TEORIA 2. PCA

1. **Chatbot**

**Control a través del Prompt de Sistema:**

* Mantener la conversación enfocada en obtener esta información
* Si el usuario desvía la conversación, redirige amablemente hacia los datos necesarios
* No proporcionar información no solicitada
* Almacenar la información en un formato estructurado o plantilla
* Confirma los datos antes de finalizar

**Variables de contexto: Recordar, priorizar y rastrear**

* El código mantiene un estado actualizado de la información recopilada
* Prioriza la obtención de información faltante
* Utiliza un sistema de flags para rastrear el progreso

**Aprendizaje continuo**

* A través de ConversationAnalyzer registrar y analizar conversaciones para identificar patrones utilizando clustering.
* Realizar seguimiento del rendimiento y ajustar los prompts basándose en métricas de éxito
* Seguimiento de métricas clave y generación de insights.

**Ver:** TEORIA 3. Código ejemplo promt ChatBot