# Notas del autor para la Universidad Internacional de la Rioja:

Por lo ya debatido con el profesor el proyecto es una solución IA que se modificara/adaptara a futuro en un entorno real. Por la posible sensibilidad de los datos toda mención a grupos de individuos, empresas, equipos se tratará como “REDACTED”. Disculpen las molestias ocasionadas

En todos los apartados principales del apartado “CRISP-DM” que suponen carga laboral más allá de la propia explicación del proyecto irán una estimación extremadamente superficial de la dedicación a esto. Esto se debe a, aunque se tiene extrema flexibilidad en los protocolos del equipo, se debe dar una primera estimación.

Junto con los títulos de cada apartado ira un número, el cual es la estimación en jornadas que tardaría una persona para este proyecto. En los subtítulos estará el desglose de cuanto se tarda esos sub-apartados, es decir, para saber el tiempo total solo sumas los títulos principales. (verdes).

Estas estimaciones tendrán en cuenta que esa persona es junior en el sector de datascience e IA pero con 5+ años en el sector del desarrollo software, y ~4-5 años en el sector de bases de datos, por lo que especialmente en las partes con menos experiencia serán más jornadas de las habituales

# Cambios con referencia al trabajo original

**Estos apartados se subrayará su título para facilitar la búsqueda**

* **Pequeño retoque en el tema legal**
* **Apartado Gestión de tiempo y presupuestos tras la propuesta inicial**
  + **Gestión de personal y costes**
  + **Gestión de licencias y coste de hardware**
  + **Gestión de tiempos**
* **Porque no usar soluciones ya creadas**

# Propuesta inicial

Solución de Análisis de Conversaciones y Feedback en Asistentes virtuales.

## Componentes de IA dentro del proyecto

* Procesamiento de lenguaje natural: Para analizar las conversaciones y entender los tópicos y sentimientos de los usuarios
* Modelos estadísticos: Para analizar e identificar patrones de interacción y feedback

## Componentes de Datascience

* Databases
* Herramientas de visualización de datos para generar una comprensión humana

## Justificación de riesgo

El riesgo del proyecto es considerable dado que se van a tratar con datos de empresas sensibles y en algunos casos con datos de empresas de sectores críticos como administraciones, aeronáutica y militar entre otros, por lo que aun siendo algo relacionado con chatbots podría considerarse de “Alto Riesgo” al trabajar con datos de estos sectores críticos.

Aun asi el proyecto está cubierto por varios frentes:

* Los datos utilizados son internos y ya se cuenta con la autorización para su uso, minimizando riesgos de privacidad y legales.
* En el apartado de ciberseguridad todos los datos usados para estadísticas serán usados en “sub-nubes” internas de la compañía las cuales disponen de múltiples capas de seguridad para evitar accesos indeseados.
* Por la propia infraestructura y protocolos de la empresa los ataques usando técnicas de hacking social tienen un riesgo mínimo.
* Este sistema no realiza acciones críticas que puedan causar daños físicos o psicológicos directos a los usuarios.
* Se enfoca en mejorar un servicio ya existente y optimizar recursos internos sin implicaciones de vigilancia o reconocimiento en espacios públicos.

Expandiendo el apartado legal, en gran parte de los datos sensibles de usuarios para este procedimiento serán directamente eliminados (como direcciones, nombres, apellidos, códigos de empleado…) y en caso de necesitar datos de identificador único se optará por la opción de hashing ya que no supone un riesgo ese extremado bajo % de que dos datos diferentes tengan el mismo hashing.

|  |  |
| --- | --- |
| **Tabla uso de datos personales** | |
| **Lawfulness, fairness and transparency** | Los datos recogidos y su eso están expresamente declarados y aceptados por los trabajadores |
| **Purpose limitation** | Su uso está especificado, en este caso la explotación para estadísticas que se traducen en mejoras |
| **Data minimization** | Solo se recogen los datos necesarios, hasheando o eliminando los personales e identificativos hacia una persona para mantener identificadores únicos de conversación pero que no se tenga conocimiento de quien la mantuvo |
| **Accuracy** | Al tratar con la propia BBDD de los empleados son datos totalmente precisos |
| **Storage limitation** | Solo se almacenaran los datos estrictamente necesarios para la obtención de estas mejoras, no almacenando identificadores personales o similares |
| **Integrity and confidentiality** | Se realizara en nubes privadas e internas minimizando al máximo el riesgo de brechas de seguridad |
| **Accountability** | Disponemos de la documentación adecuada para demostrar todos los anteriores puntos en caso de ser necesario |

En el hipotético caso de que quisiésemos extender este proyecto a múltiples empresas fuera del tejido interno y comenzar a venderlo como un producto necesitaríamos asegurarnos de que en los entornos en los que se va a usar se autoriza uso de los datos personales para el análisis de sentimientos, feedback, mejoras de conversación…

## Retoque al tema legal

Se respetara en todo momento los principios en los artículos de la GDPR

<https://gdpr.eu/article-5-how-to-process-personal-data/>

Asi como la legislación de IA como mencionamos en el apartado anterior  
<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

El proyecto estará en constente mejora para cumplir perfectamente las diferentes legislaciones de la unión europea, en caso de dudas con la misma se podrá contar con compañías como OneTrust para hacr una revisión y auditoria legal de que todo esta correcto

<https://www.onetrust.com/resources/navigating-the-eu-ai-act-ebook/>

# Gestión de tiempos y presupuestos

## Gastos en personal:

Para este proyecto y al ser algo a menor escala vamos a obtar por un equipo reducido de alrededor de 3-4 personas, puede llegar a alargar un poco el proceso de desarrollo pero con esto nos aseguraremos de una eficiencia en la comunicación y desarrollo.

Asi mismo para ahorrar gastos contaremos con auditorias de seguridad y legales externas en las fases intermedias y finales del proyecto.

|  |  |
| --- | --- |
| Personal contratado | |
| **TL** | Líder de equipo encargado de la coordinación de los diferentes miembros y hacer de interlocutor con las diferentes partes de la empresa con las que sea necesario interactuar |
| **Ingeniero Informático con conocimientos devops** | La tarea de este rol seria en un inicio el montaje de la infraestructura de trabajo inicial y tras esto dar apoyo a los otros 2 roles especialmente en la automatización de las tareas |
| **Ingeniero IA / Científico de datos** | Su función inicial será trabajar en las primeras fases como sub-lider del proyecto apoyándose en los otros 2 desarrolladores para montar los diferentes datasets con los que trabajemos, seguido de esto y una vez preparado todo ira haciendo el prototipo que el ingeniero informático mas adelante automatizara |
| **Desarrollador de Software FullStack** | Su función principal será el desarrollo de la página web en la que se montaran las estadísticas y datos relevantes a los clientes y/o directivos. Para asi dar más visibilidad tanto en las fases iniciales como en el proyecto final y su utilidad |

|  |  |
| --- | --- |
| Personal subcontratado | |
| **Auditoria Ciberseguridad** | Se realizarán dos auditorías, la primera se realizará en los entornos de desarrollo una vez se complete el primer prototipo.  La segunda auditoria será una vez la puesta en producción esté acabada aunque aún limitada se hará sobre esta misma para asegurar la correcta seguridad en todo |
| **Auditoria Legal** | La auditoría legal será algo más constante en lugar de tener puntos concretos sobre los que entrara, siendo un apoyo con capacidad de consulta a los miembros del equipo, especialmente al científico de datos para evitar cualquier problema legal en la recolección de los datos.  Esto podría hacerse con empresas como OneTrust |

Con todo esto mencionado los costes aproximadamente en personal serían los siguientes

* Team leader – 18.000€ trimestre
* Ingeniero informático con DevOps – 12.000€ trimestre
* Ingeniero IA / Cientifico de datos – 18.000€ trimestre
* Desarrollador Fullstrack – 12.000€ Trimestre
* Subcontratación Legal – 3000€ Trimestre

Y a lo largo del proyecto se añadirán 7500€ por auditoria de ciberseguridad.

Teniendo en cuenta que la estimación total de este proyecto se espera que se tarde entre 6 y 9 meses podríamos hacer una estimación de coste humano en:

**126.000€ a 189.000€ + 15.000€** en total por ambas auditorias.

## Gestión de licencias y coste de hardware

### Azure Machine Learning

* Azure Machine Learning Service Workspace: $9.9/mes

### Computación

* Virtual Machines (para entrenamiento del modelo):
* Azure DSv3-series (D8s v3) VM: $0.40/hora
* Asumiendo 100 horas de uso al mes: $40

### Almacenamiento

* Azure Blob Storage: $0.0208/GB al mes
* Estimación de 100 GB de almacenamiento: $2.08/mes

### Azure Databricks

* DBU (Databricks Unit) por hora:
* Standard plan: $0.25/DBU/hora
* Asumiendo 100 horas al mes: $25

### Azure Kubernetes Service (AKS)

* AKS Cluster: Los costos dependen de los nodos utilizados.
* Suponiendo un pequeño clúster con 3 nodos: $0.096/nodo/hora
* Asumiendo 100 horas de uso al mes: $28.8

### Azure Cognitive Services (Text Analytics API)

* Text Analytics: $2 por 1,000 llamadas de API
* Estimación de 50,000 llamadas al mes: $100

### Licencias de software

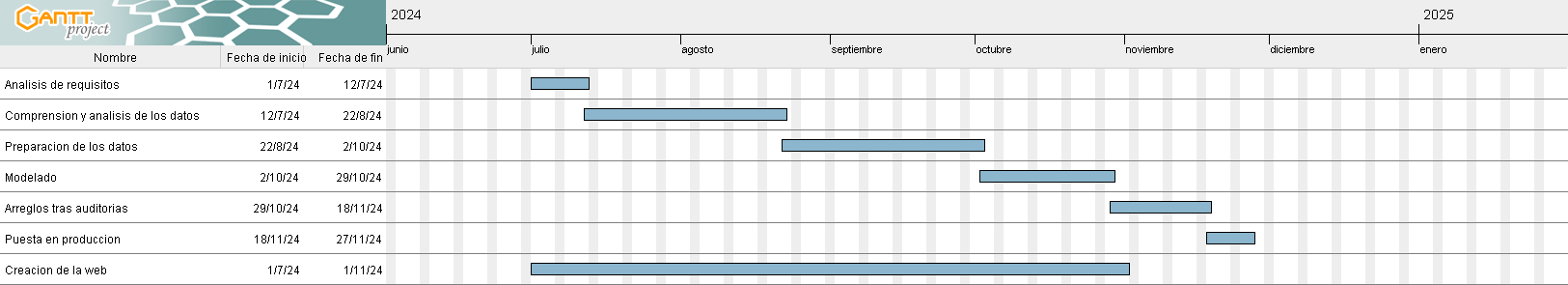
Para facilitar el desarrollo se proporcionara a los desarrolladores las siguientes licencias.

* IntellIdea Package – 30€ por desarrollador
* Github Copilot – 15€ por desarrollador
* ChatGpt 4º - 20€ por integrante (prohibido su uso con datos sensibles)

Sumando un total al costo de software de 215€ mensuales extra, haciendo una suma de **410€ mensuales extra**

## Gestión de tiempos

Los tiempos los hemos planteado para aproximadamente tener un prototipo durante el primer trimestre y a partir de ahí según los retoques y adaptaciones que vayamos necesitando a los gustos de los clientes lo iremos adaptando, algo que estimamos que nos llevara de 1 a 2 trimestres de pequeñas adaptaciones y pulimientos



## Costes totales:

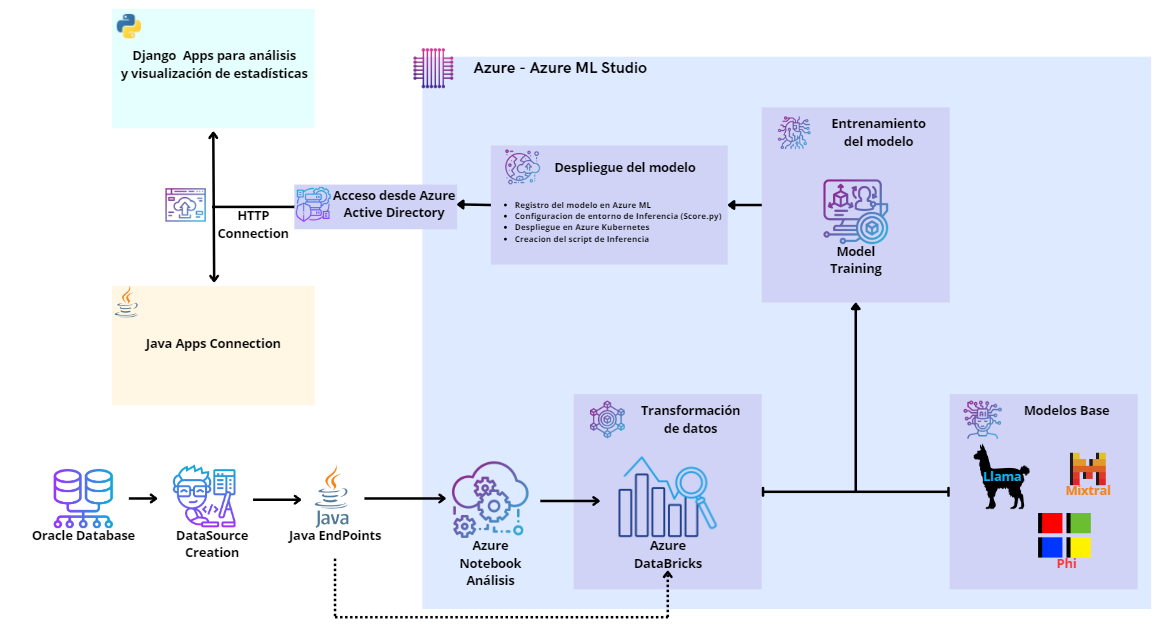
Tras todo lo mencionado los costes totales del proyecto se irían a:

* **141.000 a 204.000** en personal
* **1230** en hardware y software trimestral

Dando un total de:

**142230 a 207690€**

Esquema de infraestructura

****

# Porque no usar soluciones ya creadas

Aunque soluciones como datarobot (65.000$ aproximados de coste anual inicial) o dataiku con su plan de gastos escalables pueden resultar extremadamente tentadoras dada las políticas actuales de la empresa es, en la medida de lo posible, necesario el uso o bien de servicios internos de la compañía o, en su defecto servicios de Microsoft por la relación que tiene con la misma, de ahí que se opte por su desarrollo en todo el entorno Azure.

Asi mismo por la propia sensibilidad de los datos con los que trabajamos es de extrema importancia que estos no abandonen el entorno de la compañía, por lo que usando los servicios de partnetship con Microsoft nos facilitamos el control, customizacion, flexibilidad, escalado asi como el provecho del expertise que ya tienen los miembros de la compañía en los entornos de Azure.

## Herramientras de Analytics

### Data Catalog and Discovery

En este apartado sí que utilizaremos algunas herramientas pero más que para el propio tratado y análisis de nuestros datos para facilitar y agilizar al máximo la búsqueda de información con catálogos/buscadores alimentados de inteligencia artificial como You.com.

Es importante mencionar que ninguno de los desarrolladores expondrá ninguno de los datos en estas herramientas por la sensibilidad de los mismos.

ChatGPT será usado de manera similar a You.com, limitando la alimentación de datos para evitar filtraciones de los mismos

En cambio sistemas recientes como Copilot y Github copilot y por tratos con la empresa madre si se podrán usar dada la conexión con nubes privadas que no guardaran ninguno de los datos dados y tratados .

### Privacidad y Seguridad

Aunque la empresa dispone de un equipo de tanto blue team, como red team propio para estos temas, siempre se podría buscar apoyo de empresas tales como Ketch o Privacera para analizar los posibles riesgos asi como realizar auditorías, aunque esto podría elevar los costes del proyecto entrono a 10.000 o 30000€ dependiendo de la profundidad de esos apoyos y auditorias.

# CRISP-DM

## Business Understanding

### ¿Para qué?

El objetivo de este proyecto son 2, un objetivo principal y un objetivo secundario.

El objetivo principal del proyecto y la tarea que debe cumplir es desarrollar un sistema de entendimiento de lenguaje natural y estadística para analizar 2 cosas:

La primera encontrar aquellos apartados, palabras claves, tópicos… los cuales son altamente preguntados por los usuarios de un asistente virtual ya funcionando los cuales este asistente no tiene los conocimientos para responder, para de esta manera, poder ir optimizando el crecimiento del mismo y enfocando en los tópicos más buscados.

La segunda un análisis de feedback y sentimientos sobre el propio asistente, analizando el feedback dado por los usuarios, las situaciones en las que se dieron los diferentes feedback (el asistente consta de 2 feedback, uno genérico y uno que aparece en cada conversación que se tiene con él al final de la misma). Para asi analizar posibles frustraciones, y puntos débiles del mismo.

El objetivo secundario de este proyecto es introducir a un equipo el cual ha estado trabajando con principalmente tecnologías Java, microservicios… y asistentes virtuales con IBM Watson (no confundir con Watsonx que es la nueva versión) el funcionamiento con las nuevas tecnologías de IA para sumarse a la cresta de esta ola, priorizando servicios de Azure para ello.

### ¿Vale la pena?

Respuesta corta: Sí

Durante años este asistente ha estado ahorrando increíbles cargas de trabajos al departamento de recursos humanos de múltiples empresas entre ellas una de las más grandes que operan en España, lo que supone una reducción cuantiosa de costes. Este proyecto plantea el evolucionar esto a un siguiente nivel y abrir la puerta no solo a la mejora de este proyecto si no a comenzar a adaptar a equipos de la empresa a estas tecnologías.

### ¿Cómo se obtendrán los datos?

Al ser un asistente virtual dentro de la propia empresa ya disponemos de las autorizaciones necesarias por parte de los usuarios (dado que ya se sacaban estadísticas de manera más superficial) para tratar con estos datos de conversación e interacción.

Actualmente están almacenados en formato de tablas.

### ¿Cómo lo hacemos?

Recogiendo los diferentes eventos durante las conversaciones con el asistente virtual.

### Riesgos

Actualmente los riesgos del proyecto serian bastante bajos, dado que principalmente se redirigían recursos de **REDACTED** durante épocas de baja presión haciendo que el impacto sobre el resto de tareas del equipo a penas se vea afectado reduciendo al mínimo el coste humano.

Asi mismo al no ser un proyecto de cara al público no exige una respuesta instantánea lo que nos habilita trabajar con menos recursos de hardware/software reduciendo costes monetarios.

El proyecto se realizará con Azure cosa que también nos permite más agilidad en soporte y mantenimientos más baratos gracias al acceso a un soporte más personalizado gracias a anteriores tratos y proyectos con **REDACTED.**

## Compresión de los datos

### Recolección de los datos

Una vez tenemos los datos se comenzarán a la creación de diferentes datasets sobre estos mismos debido al gran volumen de ellos, evitando la recolecta de datos innecesarios para alimentar a los modelos.

### Examinado y tratado

Actualmente se tienen planteadas varias aproximaciones para el examinado y tratado de estos mismos.

#### Para las conversaciones

Dado a que desde el comienzo y para asegurar una calidad de estos datos se trabajó con tanto identificadores únicos de conversaciones como variables de contexto que identificaban el final de un tópico (ya sea porque el asistente no tenía conocimiento del mismo o por que el usuario cambio de tópico), podemos analizar con seguridad los apartados que más nos interesan, es decir, los finales de las conversaciones y, las situaciones donde el asistente no puedo encontrar esas respuestas.

Las ideas iniciales es realizar diferentes investigaciones con datasets centrados en:

* Cuantos intentos diferentes hace el usuario con respecto un tema que el asistente no entiende antes de abandonarlo, y si tras no encontrar una respuesta lo abandona directamente o pregunta por otro tema
* Análisis más profundo de duración de conversaciones, preguntas… …

En la parte de preparación de datos expandiremos más en profundidad esta sección.

#### Para el feedback

Aunque no fue una opción que se encontraba desde el comienzo del asistente virtual, el feedback lleva ya varios años implementados con gran profundidad en las estadísticas que saca cuando es enviado asi como las vinculaciones suficientes para identificar lo que sucedió durante cada conversación en la que se dio ese feedback, para de esta manera poder analizar las situaciones del usuario al dar este mismo.

***A partir del siguiente punto son ideas generales las cuales son susceptibles a cambios, pero nos sirven como base de la aproximación que se tiene con el proyecto y como se van a plantear de primeras los diferentes apartados.***

## Preparación de los datos - 15

Dado que estos datos irán cambiando, todos los scripts de la preparación de datos serán planteados de tal manera que puedan ser usados para la automatización de estos procesos.

### Preparación base - 10

La idea inicial para el filtrado y creación de datos es establecer 3 datasets iniciales

1. El primero sería más ligero enfocándolo principalmente en la extracción del nodo/nodos (en caso de ser nodo múltiple) de la última interacción que tuvo un usuario con el asistente antes de abandonarlo. Remarcando si esos nodos eran de una respuesta encontrada o no encontrada
2. En caso del segundo dataset sería algo más pesado porque la idea sería poder analizar las conversaciones en sí, es decir en este dataset almacenar ese flujo de nodos de conversación para poder interpretarlo y estudiarlo en los diferentes casos, manteniendo en todo caso las variables de entorno que nos facilitan el trabajo de análisis ya que reducen lo que debe suponer el futuro modelo de análisis.
3. En caso del tercer dataset lo tendríamos orientados más a identificar las diferentes **REDACTED** de la compañía a las que pertenecen, para asi identificar qué áreas lo usan más, datos que nos pueden servir para orientar la carga del asistente en un futuro y hacia donde enfocar los recursos restantes de RRHH

### Limpiado de datos - 5

El limpiado de datos se tendrá en varias fases, ya que varias de ellas las tenemos bastante presentes:

1. Lo primero tenemos que eliminar los datos de los usuarios de testing, ya que, algunos usuarios de los desarrolladores (que no son el público objetivo real del asistente) pueden llegar a ejercer un Bias. Esto se debe a que para testear el correcto funcionamiento de servicios interconectados con el asistente usamos casos controlados que sabemos que actúan sobre esos servicios.  
   Estos usuarios serian principalmente: **REDACTED**
2. Lo segundo habría que tener en cuenta las fechas de cada uno de los datos, ya que, aunque el conocimiento del asistente en algunos sectores este estancado en otros sigue updates constantes, asi que hay que tener en cuenta esos sectores si destacan en esos datos
3. Tratado con usuarios desconocidos o que ya no se encuentren en la compañía, según el nivel de casos habrá que actuar de una manera u otra, ya sea mediante simple eliminación, filtrado o técnicas más avanzadas si fuese necesario.