- Alexander Sánchez Sánchez
- Juan David Cruz García
- Juan Sebastian Pérez Camacho
- Kennet Santiago Sánchez Roldán

Business Understanding

Tarea 1: Determine Business Objectives.

- **Background:** Actualmente plataformas como Crunchyroll cuentan con más de 20 millones de usuarios activos anualmente de los cuales un 10% son habituales mes a mes, a su vez también se registra que la industria creciente del anime cuenta con más de 20 mil títulos disponibles para ser vistos. El sistema de monetización de estas plataformas de streaming se basa en sus usuarios suscritos. Además, los convenios con casas de animación para transmitir sus contenidos audiovisuales, parten de la confianza de las plataformas para garantizar que sus animes serán vistos. Esta visibilidad en sus animes genera el ciclo de negocio entre las plataformas y las casas de animación. Por tanto, contextualizando la situación a Latinoamérica donde tenemos países que hacen parte de los mayores consumidores de anime mundiales, se apunta a generar recomendaciones pertinentes donde los usuarios decidan su permanencia en una plataforma de streaming gracias a su interacción con esta.

- **Business objectives:**
  . Clasificar usuarios en categorías de anime, con el fin de generar métricas para recomendación de anime.
- La clasificación se basará en la puntuación del anime dada por los usuarios.
- La clasificación de anime, se basará en categorías mundialmente conocidas y nombradas por convención.
- Al clasificar los usuarios se tendrán parámetros con los cuales determinar el comportamiento de los usuarios.
  . Recomendar anime al usuario en función de las métricas de clasificación.
- Las recomendaciones permitirán a los usuarios percibir unicidad en sus apartados y ver lo que les interesa.
- Las recomendaciones permitirán a plataformas de streaming ampliar la base de usuarios, gracias a la implementación de recomendaciones personalizadas.

- **Business Success Criteria:**
- Basados en el historial de actividad de los usuarios, se generará un porcentaje, el cual refleja la cantidad de recomendaciones arrojadas que estos usuarios decidieron ver y calificar, en función del total promedio de animes que estos vieron y calificaron semanalmente, así validando que las recomendaciones están siendo afines a los usuarios. El porcentaje nace de:
  ((Recomendados Vistos y calificados) / (promedio de calificación del usuario)) *100

Tarea 2: Assess Situation

- **Inventory of resources:** Como recursos tenemos:

- **Personal:** Contamos con 4 estudiantes de ingeniería en sistemas, con habilidades en python y aproximaciones al uso de algoritmos de inteligencia artificial. Estudiantes con total disposición al proyecto además, una oportunidad de ampliación en conocimiento gracias al apoyo de profesores guía con conocimiento previo en algoritmos de inteligencia artificial.
- **Datos:** Base de datos en kaggle de las preferencias de 73.516 usuarios para 12.294 animes.
- **Recursos computacionales y de software:** Nuestro equipo de trabajo cuenta con 4 computadores capaces de ejecutar software competente, e implementar proyectos en plataformas como Jupyter notebook para la elaboración de archivos ipynb o py.

- **Requirements, assumptions, and constraints:**

  Requerimientos: El proyecto debe estar en la capacidad de:
  - Procesar información desde una base de datos.
  - Analizar la información extraída con la finalidad de clasificar en grupos característicamente homogéneos a los usuarios.
  - Generar métricas de recomendación en función de la clasificación de los usuarios. Entendiendo que las métricas se basarán en lo afín que puede ser un anime a un usuario.
  - Presentar una recomendación en la plataforma del usuario en función de las métricas de recomendación.
  - Generar un porcentaje de aceptación, basado en la interacción del usuario con las recomendaciones.
  - Depurar las métricas de recomendación para que el porcentaje de aceptación de estas se acerque lo más posible al 100%.
  - Validar cómo se comportan los usuarios frente a las recomendaciones gracias al porcentaje de aceptación y así determinar si estas afectan directamente a la base de usuarios en plataformas de streaming.

  Suposiciones:
  - Las métricas de recomendación se harán en función de los géneros de los animes y su cantidad de episodios.
  - Al generar las clasificaciones de los usuarios, se denotará que están basados en la preferencia de estos reflejada en su historial.
  - De la lista de animes que la persona verá en la plataforma la gran mayoría, un 70%, son aquellos generados por el modelo de recomendación.
  - La clasificación de ovas (Original Video Animation) se basará en su género y cantidad de episodios.
  - Entender que los usuarios son entes diversos, los cuales tendrán gustos relacionados a su humor diario, será un factor crítico a la hora de mejorar las métricas de recomendación.
  - A la hora de presentar el modelo, el resultado de este debe ser una plataforma interactiva en la cual se generan los listados de anime correspondiente, en función de la información suministrada por el usuario.

Restricciones:

- La falta de información estadística latinoamericana respecto al consumo del anime, debido a que las grandes compañías como Crunchyroll no se han dado la tarea de brindar información detallada, gracias a la no muy antigua popularidad exponencial de este contenido.
- Para mantener la integridad de los usuarios se debe validar que en las plataformas a implementar el sistema cuente con una autenticación de usuario, así permitiendo la unicidad del resultado.
- Se debe verificar detalladamente qué categorías de anime son lo suficientemente similares con otras para ser reducidas y fusionadas, así evitando hacer esfuerzos de recomendación innecesarios.
- Se debe validar el uso correcto de algoritmos eficientes para evitar la demora en la generación de las recomendaciones.
- Reforzar en los conocimientos sobre inteligencia artificial ha de ser necesario para la correcta resolución del proyecto.

- **Risk and contingencies:**

  - **Riesgos:**
    1. La información contenida en nuestra base de datos puede no ser valiosa o puede estar anticuada.
    2. La base de usuarios puede llegar a no ser representativa gracias a la cantidad inmensa de animes que se pueden recomendar.
    3. Las métricas de recomendación tienen que evitar ser arbitrarias y llegar a una convención basada en las respuestas brindadas por el análisis de los datos.
    4. El daño o manipulación errónea de la base de datos depurada será vital para las entregas de avances correctos en el proyecto.
    5. Generar un ambiente de trabajo idóneo para los miembros y así evitar el abandono del proyecto.

  - **Contingencia:**
    1. La búsqueda de nuevas bases de datos puede facilitarse puesto que el tema del proyecto es relativamente popular.
    2. Usando la base de datos que tenemos y las clasificaciones por género respectivas podemos evitar la baja representatividad de los datos y hacerlos relevantes en el contexto.
    3. Se debe buscar convenciones basadas en el contexto cultural y económico del anime, entendiendo como funciona y así corroborar esto clasificándolas en función de los parámetros arrojados por el modelo.
    4. La base de datos será alojada en línea priorizando su persistencia e integridad, además de ser manejada en versiones con cada alteración.

5. Los miembros del equipo tienen cargos respectivos dentro de este con la finalidad de llegar a una correcta realización del proyecto.

- **Terminology:**

  Datos discretos: Datos que solo pueden contener un dominio de valores determinado previamente
  Algoritmo supervisado: Técnica para deducir una función a partir de datos de entrenamiento.
  Segmentación: Separación o división de información con la finalidad de analizarla.
  Mapeo: Trazado de un mapa de elementos de datos entre dos modelos de datos distintos.
  Muestreo: Selección de un conjunto de personas o cosas que se consideran representativas del grupo al que pertenecen, con la finalidad de estudiar o determinar las características del grupo.
  Anime: Género de animación de origen japonés que se caracteriza por un grafismo crudo y argumentos que frecuentemente tratan temas fantásticos o futuristas.

- **Costs and benefits:**
  El costo principal del proyecto es la inversión en tiempo que realizaremos para él. Uno de los beneficios será el aporte al conocimiento grupal, del cual obtendremos nuevas habilidades tanto prácticas como conceptuales sobre el desarrollo de proyectos de software y su aplicación en contextos socio-económicos.

Tarea 3: Determine Data Mining Goals

- **Data mining goals**
  El modelo al momento de implementar la minería de datos debe:
  - Segmentar correctamente a los usuarios entre las diversas categorías
  - Clasificar adecuadamente un nuevo usuario en una categoría
  - Asociar un usuario con otro con base en el historial de animes visto
  - Poder implementarse de acuerdo a los objetivos de negocio planteados
  - Asignar una categoría a un usuario con un 80% de efectividad
  - Usar información histórica acerca de animés vistos para retroalimentar el modelo
  - Predecir qué categorías específicas y animes son las más afines con ciertos usuarios
  - Arrojar una lista de recomendaciones las cuales encajan con el interés del usuario

- **Data mining success criteria:**
  El criterio de éxito de nuestro modelo estaría basado en la cantidad de recomendaciones exitosas a los usuarios, siendo esta mayor al 80%. De modo que este porcentaje puede ir enlazado también a la calificación que dan los usuarios al contenido que están consumiendo. Es decir, el éxito del modelo se evaluará con un 80% de buenas calificaciones, esto a través del sistema de calificación de la plataforma que, como ya se cuenta con un dataset, se evaluará si las series que el sistema predice han sido calificadas con más de 6.5 por el usuario, junto con la afinidad que hay entre los diversos usuarios con gustos similares por el mismo género del usuario en cuestión. Por último, otra métrica que puede ayudar a determinar el éxito del modelo es la retención de usuarios a través del tiempo, considerando la estancia de un usuario por más de tres (3) meses como buen indicador del modelo.

Tarea 4: Produce Project Plan

- **Project plan:** el proyecto contará con las fases de:

    - **Descarga y entendimiento de la base de datos**: dado que es una base de datos hecha por un tercero, es necesario destinar tiempo (una semana) al completo entendimiento del funcionamiento de la misma. El recurso es la base de datos que se descargará de kaggle. No hay ninguna salida.

    - **Clasificación de usuarios**: para hacer recomendaciones de forma más eficiente necesitaremos poder clasificar a los usuarios según los géneros que más vean o mejor califiquen. Esta etapa tomará una semana aproximadamente, y solo requiere la base de datos que se descargará pues esta posee la información sobre los usuarios y las producciones que han calificado. La salida sería una lista con las categorías tentativas a las que puede pertenecer un usuario.

    - **Implementación del algoritmo de clasificación**: una vez se hayan definido las categorías, procederemos a implementar el algoritmo que clasifique a estos usuarios en la lista que hemos construido en la fase anterior. Esta fase tomará aproximadamente 4 semanas, y necesita la lista de categorías a las que puede pertenecer un usuario, así como la base de datos para poder entrenar el modelo. La salida será un algoritmo que clasifique a los usuarios

    - **Implementación del algoritmo que conecte las clasificaciones de usuarios con el género de las producciones**: esta etapa probablemente tome cuatro semanas. Necesitará la base de datos y las categorías a las que han sido asignados los usuarios. La salida será un algoritmo que retorne una lista de producciones recomendadas para cada usuario según su categoría, y calificación a los géneros de las producciones.

- **Initial assessment of tools and techniques:**
  El proyecto contará con una serie de herramientas y técnicas dependiendo de la fase del proyecto. En esencia, se determinó que el lenguaje que se usará para la evaluación, análisis y clasificación de los datos será Python a través de una serie de librerías matemáticas, analíticas y gráficas para documentar los resultados obtenidos junto con técnicas de segmentación como lo son el algoritmo KNN (k-nearest neighbors). También, el ambiente de desarrollo y documentación a utilizar serán los notebooks de Jupyter, puesto que estos permiten implementar, como se ha dicho anteriormente, la documentación paralelamente al desarrollo del modelo. Sin embargo, cabe aclarar que para mayor documentación explícitamente del código utilizado se acogerá así mismo la herramienta Sphinx. Por último, la plataforma donde residirá nuestro proyecto será Github, la cual nos permite hacer un versionamiento del mismo y así llevar un control de cambios. Finalmente, el documento en texto y formal para presentar se llevará a cabo a través del gestor de documentos Microsoft Word por convención.