Clasificador de Especies

Seminario de Analítica y Ciencia de Datos

Resumen Descriptivo del Proyecto

El proyecto surgió en una reunión con el grupo de investigación de bioinformática de la Universidad de Antioquia, los cuales tienen la necesidad de entrenar, parametrizar y evaluar modelos de redes neuronales para clasificar secuencias de ADN, con el fin de publicar un artículo científico sobre los resultados obtenidos. En términos de despliegue e infraestructura, se entrenaran los modelos en clusters locales, los modelos serán de uso libre y serán compartidos en plataformas de código abierto como github.

1. Problema.

a. ¿Cuál es mi problema alrededor de mi proyecto?

Identificar la mejor arquitectura de redes neuronales para la clasificación de secuencias de ADN

b. ¿Por qué se manifiesta el problema?

El grupo de investigación de bioinformática no tiene una arquitectura definida donde se entrenen, parametrice, monitoree y despliegue los modelos

c. ¿De quién es el problema?

Grupo de investigación de bioinformática

2. Datos.

a. Describa la fuente de datos (Sea más específico que en documento pasado)

Bases de datos de secuencias de adn comentadas de uso público como: DDBJ, GSA o NCBI GenBank

b. ¿Cuál es la calidad que posee dicha fuente?

La información está estandarizada por la comunidad internacional para asegurar su calidad, además con análisis preliminares hechos: conteo de vacíos, pertinencia de la data en su campo y descripciones estadísticas se encontró una buena calidad de datos.

c. ¿Es accesible y disponible?

Si, las bases de datos son de uso público y los servidores de las entidades son altamente disponibles

3. Hipótesis.

a. ¿Qué es lo que pretende probar?

existe una arquitectura de red neuronal ideal que permite obtener resultados satisfactorios a la hora de aplicarse en procesos de clasificación de cadenas de ADN

b. ¿Cuáles pueden ser las respuestas esperadas a lo que pretende probar?

* no existe una arquitectura inherentemente mejor para problemas de clasificación de adn
* una arquitectura muestra mejores resultados que las demás

c. ¿Qué acciones debo tomar para cada respuesta anterior?

* Proponer metodologías externas, como cambiar la decodificación de las entradas para entrenar los modelos
* Recopilar las características críticas del modelo y definir cuales son los parámetros ajustables para que la arquitectura se ajuste mejor a los problemas de clasificación de las cadenas de ADN

4. Solución.

a. Tipo. (¿Qué tipo de algoritmo de ML es? ¿Qué técnica es?

Red neuronal de multiples capas

b. ¿Cómo debe ser implementado?

se debe implementar en un contexto de fácil uso para profesionales bioinformáticos

c. ¿Qué resultados se esperan?

Una red neuronal base que permita tener resultados favorables a la hora de su implementación para clasificación de secuencias de adn

5. KPI’s

a. ¿Cómo espera evaluar el modelo?

Con múltiples fuentes de información que permitan desarrollar la misma actividad (clasificación de cadenas de adn)

b. ¿Qué métricas debe usar?

Matriz de confusión con las métricas especificidad, precisión y sensibilidad

6. Actores.

a. ¿Quién es su cliente?

Grupo de investigación de bioinformática de la UdeA

b. ¿Quiénes son los interesados (Stakeholders)?

Grupo de investigación de bioinformática de la UdeA

c. ¿Quién usará la solución?

Cualquier científico que planee desarrollar una solución de clasificación de secuencias de adn

d. ¿A quién impactará?

A aquellos bioinformáticos que busquen clasificar cadenas de adn de manera transparente, sin entrenar sus propios modelos.

Rúbricas

R01: Completitud y consistencia: el estudiante realiza una entrega del momento evaluativo con todos los elementos desarrollados, manteniendo una relación coherente entre los elementos solicitados. El estudiante usa una consistencia argumentativa que incluye el uso de referencias cuando hay mérito de autoría entre otros.

R02: Pertinencia: el proyecto del estudiante está dentro del ámbito de la analítica y la ciencia de datos y es evidente el desarrollo de cada uno de los elementos del Data Product Canvas

R03: Solución: el estudiante es capaz de responder a las preguntas del Data Product Canvas con la consistencia necesaria para relacionarla con el proyecto que desarrolla. Hay una relación evidente entre ellos. Las herramientas de MAchine Learning son una solución viable dentro de la descripción de la necesidad y objetivos del proyecto.

R04: Datos: los datos sugeridos son pertinentes al proyecto y al alcance delimitado en el momento evaluativo anterior.