1.1 Objetivos

Los objetivos principales del proyecto son los siguientes:

- Investigar y entender cuáles son los algoritmos de aprendizaje automático usados actualmente para analizar textos.
- Determinar cuáles de estas técnicas son las más adecuadas para solucionar el problema planteado en el trabajo (es decir, que tipos y que algoritmos de *machine learning* utilizaremos).
- Desarrollar el sistema utilizando las técnicas elegida que nos permita identificar los mensajes sospechosos de contener *phishing* o *spam*.
- Aplicar el sistema desarrollado a un caso real de correos en línea, como pueda ser *Gmail* o *Outlook*.

1.2 *Phishing*

El *phishing* es el delito de engañar a las personas para que compartan información confidencial y frágil, como puedan ser contraseñas o el número de la tarjeta de crédito. Hay una técnica muy común, que es enviar a la víctima un correo electrónico o un mensaje de texto que imita (o, dicho de otro modo, suplanta la identidad) a una persona u organización de confianza, como un compañero de trabajo, un banco, o una oficina gubernamental.

Si el usuario pica el anzuelo y hace clic en el enlace, se le envía a un sitio web, siendo este una imitación de la página web legitima. Es entonces cuando se le pide al usuario que se registre con sus credenciales de nombre de usuario y de la contraseña, y el atacante le puede robar la identidad a la víctima, saquear su cuenta bancaria, o vender esta información en el mercado negro.

1.3 *Spam*

El *spam* nunca es solicitado. Es molesto, normalmente es promocional, se envía a muchísimas personas y llega, aunque no se haya pedido.

El *spamming* es el acto de enviar estos mensajes, mientras que a la persona que participa en esta práctica se la denomina *spammer*. La mayoría de las veces, el *spam* es de naturaleza comercial y, aunque es preocupante, no es necesariamente malicioso o fraudulento.

El *ham* se define y se entiende como "correo electrónico que generalmente se desea y no se considera spam".

1.4 Algoritmos de *machine learning*

- **Árboles de decisión**: En este algoritmo basado en árboles, los datos de entrada son continuamente separados y clasificados según unos parámetros que se deben indicar. Funciona, como bien dice el nombre, como un árbol, con ramas de decisión como las líneas que dividen los datos; y con hojas, como las decisiones tomadas por el algoritmo para clasificar los datos (es decir, los datos de salida de la red de entrenamiento).
- **K-vecinos más próximos**: Este modelo basado en distancia, clasifica cada dato nuevo en el grupo que le corresponda según tenga *k* vecinos más cerca de un grupo o de otro, calculando la distancia del nuevo dato a cada uno de los existentes, y ordenando dichas distancias de menor a mayor para selecciona el grupo al que pertenece.
- **Bayesiano ingenuo**: Este modelo probabilístico se basa en el teorema de Bayes, el cual encuentra la probabilidad de que ocurra un evento *A* ocurra sabiendo que otro evento *B* ya ha ocurrido. Su representación matemática es:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

donde *P(X)* es la probabilidad de que suceda el evento X y *P(X|Y)* es la probabilidad de que suceda el evento *Y* habiendo observado el evento *X*.

- **Bosques aleatorios**: Como bien indica su nombre, este modelo se basa en el ensamblado de árboles de decisión individuales que operan como conjuntos. Cada árbol individual en el bosque aleatorio suelta una clase predictora, y la clase con más votos se convierte en nuestro modelo de predicción.

1.5 Herramientas usadas

1.5.1 Entorno de trabajo

Anaconda es una distribución de los lenguajes de programación *Python* y *R* para computación científica, es decir, se usa para *data science*, *machine learning*, procesamiento de datos de gran escala, analíticas predictivas, etc…, que sirven para simplificar la gestión y despliegue de paquetes.

JupyterLab es un entorno extensible para la computación interactiva y reproducible, basada en *Jupyter Notebook* y su arquitectura interna.

1.5.2 Librerías usadas

- ***pandas***: Librería de *Python* para el manejo y análisis de estructuras de datos.
- ***nltk***: *Natural Language ToolKit* (*NLTK*) es una biblioteca para el procesamiento del lenguaje natural.
- ***seaborn***: Librería de visualización de datos construida en base a *matplotlib*, que ayuda en la exploración y comprensión de los datos.
- ***wordcloud***: Representación visual de datos textuales. Obtiene las palabras individuales de un texto, y la importancia de cada palabra es representada en una figura con un color y tamaño de fuente dependiente de la importancia de dicha palabra.
- ***scikit-learn***: Esta librería va a ser el eje central en el que se van a implementar los algoritmos de *machine learning*.
- ***Imaplib/ email***: Esta librería permite leer y manejar los correos de una cuenta de *Gmail* personal, incluidas las partes *MIME* y otros documentos de mensajes basados en *RFC-2822*.
- ***pickle***: El módulo *pickle* de *Python* se usa principalmente para serializar y de serializar objetos estructurados en *Python*.

1.6 Análisis de resultados obtenidos

¿Cómo es posible que los bosques aleatorios tengan peores resultados que los arboles de decisión? Al fin y al cabo, los bosques aleatorios son algoritmos de ensamblado, que están compuestos por un conjunto determinado de árboles de decisión.

La respuesta a esta pregunta puede ser compleja y no concisa. Por ejemplo, los arboles de decisión son más eficientes cuando se quiere crear un modelo simple, cuando el conjunto de *datasets* y características pueden ser usados por completo, cuando se tiene un poder computacional limitado, o cuando no hay mucha preocupación acerca de la precisión en *datasets* futuros.

Aun así, en general debería darse el caso normalmente de que los bosques aleatorios tengan una mejor precisión que los arboles de decisión, por lo que este caso que se ha dado en mi proyecto es bastante excepcional.

Las principales razonesde que haya ganado este algoritmo son que Bayes ingenuo tiene un muy buen rendimiento cuando trabaja con múltiples clases, y en clasificación de textos. Además, algunas de las ventajas de este modelo de Bayes ingenuo que han podido beneficiar a que tenga mejores rendimientos que el resto de modelos son:

- Es un algoritmo simple, y si el supuesto de independencia condicional se cumple, este modelo de Bayes ingenuo convergerá más rápido en sus clasificaciones que otros modelos, por lo que necesitará menos datos de entrenamiento para obtener mejores resultados.
- El modelo de Bayes ingenuo requiere menos tiempo de entrenamiento del modelo para los *datasets* que se le proporcione.

1.7 Conclusiones

No se han podido completar todas las tareas como se había planeado al principio. Por ejemplo, la revisión de la bibliografía ha durado más tiempo del pensado inicialmente puesto que había mucha información que analizar para el proyecto y mis conocimientos previos de *machine learning* eran bastante limitados.

Por otro lado, la elaboración del corpus no me ha llevado tanto tiempo como el que pensaba puesto que pude encontrar *datasets* ya fabricados en Internet, y usarlos como datos de entrenamiento y de prueba.

He tenido un gran contratiempo, y es que al principio este proyecto lo había planteado para identificar solo correos de tipo *phishing*, pero al no haber podido encontrar un *dataset* completo de corpus específicos de correos de tipo *phishing*, he tenido que usar varios *dataset* que pertenecen a *emails spam*, *emails ham*, y *urls* de tipo *phishing* y de tipo legítimo.

Finalmente, la implementación del código se ha retrasado y acortado un poco debido a que el tema de documentación ha sido más denso y largo, y el código, gracias a las librerías que estoy usando de *sckit-learn*, es bastante sencillo de hacer.

Por último, me gustaría señalar algunos aspectos que han quedado fuera del trabajo realizado debido a las restricciones de tiempo y que servirían para poder tener más modelos entre los que comparar y poder estar más seguros de elegir el mejor:

- No utilizar métodos de hiper-parámetros y *cross-validation*, para poder sacar un mejor rendimiento de los distintos algoritmos, y obtener resultados aún mejores de los que he obtenido aquí.
- Crear modelos con otros algoritmos más potentes de aprendizaje automático, como redes neuronales, *SVM* (*Support Vector Machines*), o regresión lineal.
- Crear nuevos modelos cambiando la forma en que se manejan los textos, considerando, por ejemplo, no solo la frecuencia de las palabras, sino también información morfosintáctica.

1.8 Análisis de Impacto

1.8.1 Impacto personal

3.b. *Reforzar la capacidad de todos los países, en particular los países en desarrollo, en materia de alerta temprana, reducción de riesgos y gestión de los riesgos para la salud nacional y mundial*.

4.a. *Construir y adecuar instalaciones educativas que tengan en cuenta las necesidades de los niños y las personas con discapacidad y las diferencias de género, y que ofrezcan entornos de aprendizaje seguros, no violentos, inclusivos y eficaces para todos*

4.c. *De aquí a 2030, aumentar considerablemente la oferta de docentes calificados, incluso mediante la cooperación internacional para la formación de docentes en los países en desarrollo, especialmente los países menos adelantados y los pequeños Estados insulares en desarrollo*.

1.8.2 Impacto empresarial

8.8. *Proteger los derechos laborales y promover un entorno de trabajo seguro y sin riesgos para todos los trabajadores, incluidos los trabajadores migrantes, en particular las mujeres migrantes y las personas con empleos precarios*.

1.8.3 Impacto social

11.5. *De aquí a 2030, reducir significativamente el número de muertes causadas por los desastres, incluidos los relacionados con el agua, y de personas afectadas por ellos, y reducir considerablemente las pérdidas económicas directas provocadas por los desastres en comparación con el producto interno bruto mundial, haciendo especial hincapié en la protección de los pobres y las personas en situaciones de vulnerabilidad*.

11.b. *De aquí a 2022, aumentar considerablemente el número de ciudades y asentamientos humanos que adoptan e implementan políticas y planes integrados para promover la inclusión, el uso eficiente de los recursos, la mitigación del cambio climático y la adaptación a él y la resiliencia ante los desastres, y desarrollar y poner en práctica, en consonancia con el Marco de Sendai para la Reducción del Riesgo de Desastres 2015-2030, la gestión integral de los riesgos de desastre a todos los niveles*.

1.8.4 Impacto económico

9.4. *De aquí a 2030, modernizar la infraestructura y reconvertir las industrias para que sean sostenibles, utilizando los recursos con mayor eficacia y promoviendo la adopción de tecnologías y procesos industriales limpios y ambientalmente racionales, y logrando que todos los países tomen medidas de acuerdo con sus capacidades respectivas*

9.c. *Aumentar significativamente el acceso a la tecnología de la información y las comunicaciones y esforzarse por proporcionar acceso universal y asequible a Internet en los países menos adelantados de aquí a 2022*.

10.5. *Mejorar la reglamentación y vigilancia de las instituciones y los mercados financieros mundiales y fortalecer la aplicación de esos reglamentos*.

1.8.5 Impacto cultural

16.4. *De aquí a 2030, reducir significativamente las corrientes financieras y de armas ilícitas, fortalecer la recuperación y devolución de los activos robados y luchar contra todas las formas de delincuencia*.

16.5. *Reducir considerablemente la corrupción y el soborno en todas sus formas*.

16.a. *Fortalecer las instituciones nacionales pertinentes, incluso mediante la cooperación internacional, para crear a todos los niveles, particularmente en los países en desarrollo, la capacidad de prevenir la violencia y combatir el terrorismo y la delincuencia*.

16.10. *Garantizar el acceso público a la información y proteger las libertades fundamentales, de conformidad con las leyes nacionales y los acuerdos internacionales*.