Propuesta de Proyecto

Seminario de Analítica y Ciencia de Datos

# Presentación del Proyecto

# Analítica y modelo de machine learning en reportes financieros para la detección de anomalías en información financiera, con el fin de identificar transacciones inusuales que puedan afectar la consolidación y presentación de los estados de resultado en la empresa. Actualmente la compañía tiene la necesidad de identificar las transacciones inusuales que puedan afectar los estados financieros, ya que pueden ser un error o una intención de fraude la presentación de informes. Pero el volumen de datos es muy alto lo que no ha permitido que las personas puedan etiquetar los datos en usuales o inusuales. Para esto se buscar crear un modelo de Autoencoders LSTM de aprendizaje automático no supervisado en Python usando Keras.

# Objetivo del Proyecto

# Desarrollar un modelo de aprendizaje automático no supervisado, que permita identificar las etiquetas para reconocer transacciones e información financiera inusual en el libro diario, que pueda afectar la consolidación y presentación de los estados de resultados.

# Contexto del Problema

Los primeros escándalos de fraude financiero que han movido el mundo surgieron en estados únicos, los escándalos más grandes fueron protagonizados por compañías como Enron, Worldcom, Volkswagen, Madoff. Todas estas compañías tienen algo en común mentían a sus inversionistas, presentando cifras ficticias que no correspondían a la realidad financiera. (CNNMoney, 2015)

.

El problema ha impactado tanto la economía de los países, que estados unidos la ley SOX en 2002, cuyo propósito es que las compañías que quieran cotizar en la bolsa de Nueva york deben garantizar una seguridad razonable sobre la confiabilidad de la información financiera brindada a los accionistas y demás grupos de interés. la SEC en estados unidos, es la entidad encargada de velar por el cumplimiento de la ley SOX, en las compañías que cotizan en n la bolsa de Nueva York. (COMMISSION, 2016)

Debido a lo anterior, las compañías deben buscar cada vez más herramientas para poder identificar a tiempo inconsistencias en la información financiera, y dar garantía de la veracidad de sus datos. Las cuentas contables de las empresas deben ser monitoreadas con el fin de identificar a tiempo cifras inusuales y solicitar las causas y justificaciones correspondientes, antes de presentar información errada o ficticia a los inversionistas y demás áreas de interés. Cada empresa debe contar con un gobierno corporativo que le permita vigilar las operaciones y asegurar la transparencia de los estados financieros. (Delloitte)

La compañía objeto de este trabajo, es una empresa que esta en búsqueda de cumplir con los más altos estándares de presentación de estados financieros y la implementación de COSO III. Para logara su objetivo es necesario crear mecanismos y herramientas que le permitan identificar información inusual ene los estados financieros, pero el libro diario cuenta con gran cantidad de información que es difícil etiquetar la información e identificar cuales transacciones son inusuales y cuales no. Es la razón por la que se busca recurrir a mecanismos de inteligencia artificial que permitan desarrollar métodos más eficientes y confiables.

# Propuesta de Solución y Alcance

La propuesta para solucionar el inconveniente anterior e scon Python utilizando Keras, implementar un autoencoder para datos de secuencia, utilizando una arquitectura Codificador- decodificador LSTM. Se escoge este modelo ya que los autoencoders son especialmente buenos trabajando con series de tiempo, dada la capacidad para retener información temporal. Se busca entrenar el modelo con los datos del libro diario con cuentas, fecha y valor de las transacciones y una vez entrenado establecer el umbral para decidir cuando una reconstrucción s suficientemente diferente de los de entrada para considerarse ‘Anormal’. Para revisar y ajustar las etiquetas se piensa usar herramientas de visualización de datos y de revisión en pares, e incluso se piensa evaluar la necesidad de herramientas como Podigy. Para combinar l aprendizaje automático con el etiquetado manual.

Además, se buscará combinar el autoencoder con otras s herramientas de análisis estadístico como la prueba Grubbs, para la identificación de datos atípicos. También tomar unas cuentas específicas, hacer análisis exploratorio y usar modelos como vectores de soporte (SVM) o bosques aleatorios para detección de anomalías. También explorar el uso de LOF (Local Outlier Factor) para visualizar puntos de datos que están lejos de otro grupo de datos ‘normal’.

# Definición del Alcance

El proyecto pretender entregar las etiquetas para identificar cuales transacciones pueden identificarse como usuales o inusuales, para que a partir de ese resultado se pueda entrenar un modelo supervisado. Se realizará por Júpiter o en su defecto se desplegará en Azure ML.

# Riesgos e Impacto del Negocio

1. **Sobreajuste (Overfitting):** Como cualquier otro modelo de aprendizaje, es propenso a sobreajustes, y esto dependeré de la cantidad de información o datos.
2. **Complejidad computacional:** Tiempos de entrenamiento largos ypuede requerir más recursos de hardware a los que disponemos.
3. **Dificultad para interpretar el modelo:** Los autoencoders pueden identificar datos que parecen anomalías, pero no explican por que es anómalo. Por lo tanto, los resultados son difícil de identificar y entender.
4. **Elección de arquitectura del modelo:** La selección el número de capas ocultas y las unidades en cada capa, es desafiante. Además de la sensibilidad a los hiper parámetros.
5. **Grado de error en la reconstrucción**
6. **Falta de experiencia:** Esto puede afectar la interpretación del resultado del modelo, por un conocimiento menos fuerte en el funcionamiento del aprendizaje y redes neuronales. Dificultad para la elección de parámetros y el formato especifico necesario para el modelo.

# 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 en su descripción, objetivos y contexto

R03: Solución: el estudiante es capaz de relacionar herramientas tecnologico-metodológicas con el contexto de un problema de analítica y ciencia de datos.

R04: Herramientas: Las herramientas seleccionadas son pertinentes a la solución descrita. Las herramientas descritas evidencian una relación de flujo de proceso de los datos involucrados en la solución.

R05: Delimitación: dado un contexto de solución, el estudiante define claramente los límites de la solución que propone y el impacto, consecuencias y riesgos asociados a su implementación.