Juan Luis Acebal Rico

PEC 1 IDS

Yo, Juan Luis Acebal Rico, he realizado este trabajo.

Ejercicio 1:

1) Responded:

a) Son la programación-informática (Computer Science) y las matemáticas y estadística [1]

b) El dominio de la informática y del sector donde se trabaja [1]

c) Por sacar conclusiones de un sector desconocido para el investigador puede ser contraproducente además de que puede no tener sentido. [1]

d) 7) Entender las relaciones entre los datos. Es decir, entender cómo se relacionan entre sí en todos los sentidos, si tienen algún tipo de correlación, si un dato o campo depende de otro, etc. [4]

e) Exploración de los datos. Es el proceso inicial de entender el contexto de los datos, hacerse las preguntas adecuadas para ponerse en situación y preparar el posterior análisis completo de los datos [2]

f) Seguridad [5]. Si podra saberse antes del robo quien lo va a hacer, quizás aun no pero en algún momento podrá pasar ya que los modelos de predicción podrían llegar a dar una probabilidad total a que una persona fuera la responsable futura de algo que no ha ocurrido.

g) El ajuste de un modelo haciendo hipótesis, regresiones, navaja de occam, clasificación, etc. Es decir, habla de la complejidad de encontrar el equilibrio en modelos de machine learning que no resultan validos después, buscando el humor.

h) Rol unicornio. Controla las 3 partes de la ciencia de datos, cosa inusual en tu periodo formativo, ya que normalmente controlas dos y la tercera con los años de experiencia laboral podría llegar lo llama en modo jocoso asi ya que es algo no tan habitual[1].

i) Además, científico de datos en un tuit de Josh Wills [3].

2)

a) Es el adecuado en general, aunque quizás un científico de datos tal y como el ejemplo podría primero preparar los soportes físicos y lógicos para su uso después.
b) También es el adecuado, aunque ingeniero de datos y analista de datos les falta explorar los datos. Si todos damos por sentado que los datos están en perfecto estado, al final sea por datos en mal estado o por corrupción de datos, se puede llegar a que sea muy contraproducente por no haberlos revisado. Por ultimo cambiaria el contenido de los roles de ingeniero en machine learning y científico de datos.

c) Hay un gran error a mi juicio de exploración, limpieza de los datos y coordinación. El peor quizás de todos ellos es la falta de trabajo en equipo.

d) Los dos primeros roles no hacen adecuadamente su trabajo. Ya que no puedes mostrar gráficos o hacer modelos sobre datos sin revisar.

3) "En un proyecto de análisis de datos en salud, cada rol trabaja de manera **coordinada**, demostrando la confianza en la capacidad de cada profesional para llevar a cabo sus tareas eficientemente en equipo.

Para empezar, la Data Scientist valora la oportunidad de mejorar la calidad de los datos y modelos al regresar al principio del proceso y preparar toda la infraestructura para el almacenamiento y el flujo de los datos, demostrando su capacidad para asegurar la adecuada preparación de los datos y de los sistemas antes de su análisis.
La Data Analyst inicia el proceso haciendo al final las visualizaciones y haciendo una revisión exhaustiva de la calidad de los datos aun existiendo una alta eficacia de los sistemas de recopilación de datos de hospitales y laboratorios.
El Data Engineer comienza a crear modelos después de realizar una limpieza previa de los datos recopilados de diversas fuentes lo que le permite concentrarse directamente en la construcción de modelos predictivos correctos para predecir diagnósticos.
El Machine Learning Engineer, a pesar de las posibles limitaciones en los datos provenientes de registros médicos y pruebas de laboratorio, lleva a cabo un análisis exploratorio sólido, considerando la situación como una oportunidad para identificar patrones complejos y desafiantes en la información de salud. Este

enfoque, en coordinación y limpieza inicial de los datos, explota y resalta las habilidades y capacidades del equipo en el análisis de datos en salud. Además, aumenta la velocidad del procesamiento de los datos y entrenamiento de los modelos ahorrando tiempo en volver a pasos precedentes por no haber revisado bien los datos."

Ejercicio 2:

1. Medicina predictiva (según el punto de vista quizás también/o medicina preventiva). Detección precoz de cáncer en estados iniciales.
    i. Si existiera la información organizada y criterios de actuación, ¿podríamos diagnosticar un gran porcentaje de canceres en estados iniciales quedando casi siempre en un buen pronóstico?

2. Utilizaría resonancias, mamografías, análisis de sangre con marcadores cancerígenos, y donde se disponga, datos genéticos con predisposiciones al cáncer, historiales médicos, aunque empezaría claramente con pruebas de imagen y de sangre ya que es algo que estará para todos los pacientes.

3. Se podrían guardar en la nube, protegidos o en un sistema local. Lo importante seria que fueran anónimos ya que el modelo que me gustaría construir pretende predecir cáncer en base a datos que se le entrena con otros pacientes que han tenido cáncer, entonces no nos sirve sus datos personales, sino que haya suficientes pruebas de el para que participe en el modelo predictivo y saber si ha tenido o no cáncer ya.

4. Si usamos marcadores cancerígenos en la sangre podría usarse estadística como las correlaciones y si tenemos acceso a imágenes podríamos usar una clasificación. Si quisiera encontrar a grupos de riesgo quizás haría un agrupamiento. La regresión aquí quizás tendría utilidad en crear un modelo para dar una probabilidad al desarrollo de cáncer en ciertas personas sanas, y en base a ello que tengan más seguimiento.
5.
    i. Crear protocolos de actuación para que se use el sistema después. Es decir, una vez que está claro que puede hacer el modelo, hay que implementarlo de una manera que se utilice por médicos de atención primaria integrado directamente en el software de historias clínicas.

ii. Lo presentaría como una parte de un sistema de historia clínica que prometa avisar al médico cuando tiene que mandar más pruebas al paciente para detectar precozmente el cáncer.
iii. Los destinatarios serían los médicos de familia o atención primaria, los oncólogos, radiólogos y los pacientes con riesgo medio-alto.

Ejercicio 3:

Caso de estudio elegido: Gestión energética de edificios.

1. Contexto.

Dada la creciente importancia de la eficiencia energética y la sostenibilidad, la ciencia de datos se ha convertido en una herramienta crucial para la optimización de la gestión de la energía en edificios. A pesar de esta novedad, los sistemas de gestión de la energía en los edificios tradicionalmente no han operado de manera energéticamente eficiente debido a la complejidad de los edificios grandes y las diversas necesidades de sus habitantes. Sin embargo, la aparición de tecnologías inteligentes ha abierto numerosas oportunidades para mejorar estos procesos, dada la disponibilidad de grandes cantidades de datos. Utilizando inteligencia artificial y análisis de big data, los desarrolladores de la empresa Grid Edge han creado sinergias entre los distintos volúmenes de datos para que interactúen entre sí.

2. Objetivos.

El sistema “Flex2X” desarrollado por la empresa Grid Edge en el Reino Unido se ha convertido en un buen ejemplo de cómo la ciencia de datos se puede utilizar en la gestión energética de los edificios. Este sistema utiliza algoritmos de inteligencia artificial para analizar los datos en tiempo real de los sistemas de gestión de energía del edificio (BMS en inglés), así como otras fuentes, como el clima. Al ser enseñados y ajustados por los datos, los algoritmos de IA pueden predecir el uso de energía del edificio un día de antelación. Dicho de otra manera, esta medida permite una gestión más precisa y eficiente de la energía consumida.

3. Impacto.

El impacto de dicho sistema para la sociedad y la empresa puede subdividirse en:

- Para los ocupantes del edificio: Es una forma muy útil de que el confort sea optimo y los equipos o sistemas de climatización/calefacción estén

disponibles cuando se necesiten y al mismo tiempo puede llegar a reducir hasta un 40% los costes energéticos al casi eliminar el mal uso junto los datos en tiempo real como la cantidad de carbono utilizada ayudan a concienciar al usuario.

- Para el propietario: reduce costos, mejora su huella de carbono, maximiza el confort mediante el uso de los sistemas cuando realmente se necesita. Además, se recuperará la inversión más rápidamente (más del 10% de ahorro). Quizás esto puede ayudar a que a largo plazo haya una gran cantidad de edificios donde sea una tendencia la gestión por IA.
- Para los operadores de la red eléctrica: Si sistemas como este fueran utilizados por el operador de electricidad, se podría predecir el uso de electricidad a gran escala, algo muy útil en momentos de alta demanda, donde ciertas fuentes de energía renovable son intermitentes como la energía solar.

Conclusión

Flex2X muestra lo que la ciencia de datos puede hacer por revolucionar muchos campos de nuestra sociedad como por ejemplo el de edificios. Al integrar los algoritmos de IA que están siempre aprendiendo y en constante evolución, aunque claramente la mayor diferencia es que con sistemas como este, se consigue un nivel de eficacia y eficiencia nunca vistos antes, y que eso es muy positivo para la sociedad al poder hacer sistemas casi perfectos, dando una ventaja competitiva muy grande a las sociedades que hagan uso de ella. Con todo esto podemos llegar a consumir, crear, fabricar lo necesario ya que los modelos predictivos se podrían acabar usando un día para saber las necesidades exactas de algo que hay que fabricar en una ciudad cada día, acabando con el desperdicio y siendo mucho más sostenible.

Ejercicio 4:

1) Es una gran fuente de datos de la ciudad de Barcelona que está disponible para su utilización publica

2) Son datos que están disponibles para su uso para cualquier persona que quiera descargarlos.

i) Sus principales características son su uso sin restricciones y compartirlos nombrando al creador (es como Creative commons BY, es decir, una licencia abierta donde nombras al creador)

3) Hay 562
   i) Descripción: Tenemos 8 secciones con diferentes informaciones, además de las que nombro a continuación, faltarían por nombrar, sobre este sitio, actualidad, desarrolladores, agenda 2030-ODS y proyectos de difusión. La página web es muy interesante para estudiantes y trabajadores de data science ya que nos puede proveer de gran cantidad de datos para utilizarlos.
   ii) Catálogo de datasets: Es donde podemos encontrar todos los datos abiertos del ayuntamiento de Barcelona.
   iii) Visualizaciones y aplicaciones: Son proyectos que usan estos datos y mostrando las visualizaciones más interesantes.
   iv) Estadísticas: (del sitio) donde podemos ver que tienen 562 datasets, usan 36 formatos, tienen en total 3601 series históricas
4) El grado de apertura de los datasets en el portal se evalúa según el esquema de las cinco estrellas de Tim Berners-Lee. Los recursos que alcanzan un nivel más alto en este esquema son aquellos disponibles en formatos abiertos y enlazables, y que permiten su integración con otros datos en la web. A modo de ejemplo, un archivo pdf o una foto será nivel 1, un archivo propietario de datos estructurados como por ejemplo una hoja de cálculos (pero xls) será nivel 2, CSV será de nivel 3, y nivel 4 y 5(si lo he entendido bien) son datos enlazados, la diferencia entre 4 y 5 es que 4 no siempre enlazas, es decir, a veces copias el valor, y eso hace que no cumplas el criterio de nivel 5 de forma entera, es decir, un nivel 4 se usan estándares abiertos de W3C tales como RDF y SPARQL, identificas las partes para poder ser enlazado y señalado por los demás pero te faltaría enlazar hacia otros sitios web [(6)]. Cualquier nivel sería una licencia abierta para optar al menos 1 estrella.
5)
   i) CSV: Es la forma más común de recibir datos. Es una tabla de una hoja de cálculo con columnas y filas. Seria de 3 estrellas.
   ii) XLS: Es el formato propietario de Microsoft Excel para hojas de cálculo. Seria dos estrellas.
   iii) JSON: Se usa para JS y es una forma de manipular datos estilo XML en donde tiene clave-valor. Seria tres estrellas.
   iv) Por ejemplo, una API seria la forma más interesante de manipular datos ya que permite automatizar cosas tales como actualizar datos de forma automática en tiempo real (como el tiempo o el precio de kw/h). Normalmente seria 4 o 5 estrellas.

Bibliografía

(1) Marçal Mora Cantallops. Los roles, ámbitos y nombres de la ciencia de datos. Paginas 15 y 16

(2) Marçal Mora Cantallops. Los roles, ámbitos y nombres de la ciencia de datos. Pagina 38

(3) Marçal Mora Cantallops. Los roles, ámbitos y nombres de la ciencia de datos. Pagina 14

(4) Marçal Mora Cantallops. Los roles, ámbitos y nombres de la ciencia de datos. Pagina 19

(5) Marçal Mora Cantallops. Los roles, ámbitos y nombres de la ciencia de datos. Pagina 31 y 32

(6) 5stardata.info/es