Micro curso 2

**Lenguajes de Data Science**

Sem 1.1: Introducción al curso

¡Bienvenido al curso! Has comenzado una de las más completas revisiones de las herramientas de la ciencia de los datos que encontrarás en Internet actualmente. Esto no quiere decir que cubramos todas y cada una de las herramientas, pero más adelante en el curso introduciremos una lista comprensiva de las tares que un científico de datos necesita realizar y te daremos las dos o tres principales herramientas comerciales y de código abierto disponibles para completarlas. De igual forma explicamos cómo las herramientas se entrelazan en funcionalidad, cuáles son sus pros y contras y cómo estas herramientas pueden abordar todas las tuberías de datos de la ciencia de los datos. Vamos a comenzar con los datos. Los datos son obviamente importantes para los científicos de datos. En este curso, te mostraremos cómo gestionar, extraer, transformar, analizar y visualizar los datos. Ahora, podrías ser capaz de sobrevivir a la ciencia de los datos son ningún conocimiento en programación si usas el conjunto de herramientas adecuadas. Sin embargo, recomendamos que te familiarices con la programación y los marcos de programación relacionados con la ciencia de los datos. Para ayudarte, te presentaremos los lenguajes y marcos de programación más usados en la ciencia de los datos. Bien, dicho esto, hay demasiadas opciones de automatización disponibles en las ultimas herramientas que un científico de datos pueda usar. En este curso, explicaremos cómo hacer uso de dichas herramientas para ahorrar tiempo y descubrir la inspiración. La programación visual está disponible en muchas herramientas. En este curso, aprenderás como la programación visual puede ser usada para acelerar el tiempo de desarrollo y ayuda a los no programadores a entrar en el campo de la ciencia de los datos. El software de código abierto es el líder en el campo de la ciencia de los datos, pero el Costo Total de Propiedad, o “TCO”; a veces puede ser más altos para los software de código abierto. Debido a los costos de configuración, personalización y mantenimiento. Como resultado, el software comercial también tiene su lugar, especialmente desde la nueva generación de software comerciales de la ciencia de los datos que aprovecha el software de código abierto y los estándares abiertos. Lo cual facilita la migración entre las herramientas y puede reducir el TCO. En este curso, te presentaremos tanto los software de código abierto así como los comerciales y señalaremos sus fortalezas y debilidades para la ciencia de los datos. De igual forma te mostraremos las formas de aprovechar sus fortalezas. Finalmente, te mostraremos cómo la computación en la nube puede ser usada para acelerar y facilitar el trabajo de los científicos de datos. Te presentaremos las herramientas basadas en la nube más nuevas y comúnmente usadas para la ciencia de los datos. Además de las lecciones, este curso tiene numerosos laboratorios para que te familiarices con el material y también obtengas experiencia en las prácticas. También hay múltiples pruebas para evaluar tu aprendizaje. ¡No queda más nada por decir, así que empecemos! Estamos muy contentos de tenerte con nosotros para comenzar tu viaje en la ciencia de los datos. Si tienes algún problema con el material del curso, no dudes en contactarnos en el foro de discusión. ¡Así que vamos a comenzar!

Sem 1.2: Lenguajes de la Ciencia de los Datos

Los Lenguajes de la Ciencia de los Datos. Para cualquiera que este comenzando su viaje en la ciencia de los datos, la variedad de opciones técnicas puede ser abrumadoras. Hay una gran cantidad impresionante de opciones cuando se trata de lenguajes de programación. Cada uno tiene sus propias fortalezas y debilidades y no hay una respuesta correcta a la pregunta de cual deberías aprender primero. La respuesta a esa pregunta depende en gran medida a tus propias necesidades, los problemas que intentas resolver o para quien los estas resolviendo. Python, R, y SQL son los lenguajes que te recomendamos que consideres como primeras opciones. Pero hay muchos otros que tienen sus propias fortalezas y características interesantes. Scala, Java, C++ y Julia son algunos de los más populares. Javascript, PHP, Go, Ruby y Visual Basic todos estos tienen también sus propios casos de uso. El lenguaje que escojas para aprender dependerá de las cosas que necesites lograr y de los problemas que necesites resolver. Esto también dependerá de la compañía en la que trabajes el rol que tengas y la edad de tu aplicación actual. Exploraremos las respuestas a esta pregunta a medida que profundicemos en los lenguajes populares en la industria de la ciencia de los datos. Hay muchos roles disponibles para las personas que están interesadas en involucrase en la ciencia de los datos. Analista de Negocios Ingeniero de Base de Datos, Analista de Datos, Ingeniero de Datos, Científico de Datos, Científico de Investigación, Ingeniero de Software, Gestor de Producto Estadísticos, Gestor de Proyecto y muchos más. Profundicemos un poco en lo que aprenderemos en la Lección 1. Nos enfocaremos en los tres principales lenguajes de la ciencia de los datos: Python, R y SQL donde cada uno de los cuales tienen sus propias lecciones. Después pasaremos a resaltar otros lenguajes notables y lo que los hace especiales. ¡Luego, terminaremos con una pequeña prueba para que demuestres tus conocimientos! Aquí está lo que discutiremos: Python R SQL, otros lenguajes dignos de tener en cuenta en la Ciencia de los Datos y pruebas prácticas.

Sem 1.3: Introducción a Python

En este video, revisaremos las características del alto nivel del lenguaje de programación de Python. Python es un lenguaje muy poderoso. Es el lenguaje de programación más popular para la ciencia de los datos. De acuerdo a la Encuesta Kaggle de la Ciencia de los Datos y Aprendizaje Automático, el 75% de los 10,000 encuestados alrededor del mundo reportaron que usan Python con regularidad. Glassdoor informó que en el año 2019 más del 75% de los puestos de la ciencia de los datos listados incluían Python en sus descripciones de cargo. Cuando se les preguntó cuál lenguaje debería aprender primero un aspirante a científico de datos, la mayoría de los científicos de datos contestaron Python. Probablemente estés pensando, ¿Por qué Python es tan popular? Bien, comencemos con las personas que usan Python. Si ya sabes cómo programar, entonces Python es excelente para ti porque usa una sintaxis, clara y legible. Puedes hacer muchas de las cosas que estás acostumbrado a hacer en otros lenguajes de programación pero con Python puedes hacerlos usando menos códigos. Si quieres aprender a programar, este es un gran lenguaje para comenzar debido a la enorme comunidad global y a la gran cantidad de documentación. De hecho, muchas encuestas realizadas en el año2019, revelaron que más del 80% de los profesionales de datos alrededor del mundo usan Python. Python es útil para muchas situaciones, incluyendo la ciencia de los datos, inteligencia artificial, aprendizaje automático, desarrollo web y dispositivos IoT como Raspberry Pi. Las grandes organizaciones que usan Python incluyen IBM, Wikipedia, Google, Yahoo!, CERN, NASA, Facebook, Amazon, Instagram, Spotify, y Reddit. Python es un poderoso lenguaje de programación de propósito general con el que se pueden hacer demasiadas cosas. Es ampliamente soportado por una comunidad global y dirigido por la Python Software Foundation.

Python es un lenguaje de programación de propósito general de muy alto nivel que puede ser aplicado a diferentes clases de problemas.

Este tiene una gran librería estándar la cual proporciona las herramientas adecuadas para tareas diferentes, incluyendo pero no limitado a las bases de datos, automatización, web scraping, procesamiento de texto, procesamiento de imágenes, aprendizaje automático y análisis de datos.

Para la ciencia de los datos, puedes usar las librerías de computación científica de Python tales como Pandas, NumPy, SciPy, y Matplotlib.

Para la inteligencia artificial, tiene TensorFlow, PyTorch, Keras, y Scikit-learn.

Python también puede ser usado para el Procesamiento del Lenguaje Natural o por sus siglas en inglés (NLP) utilizando Natural Language Toolkit (NLTK). Otro punto de venta atractivo es la comunidad Python, la cual tiene una historia muy bien documentada de pavimentar el camino para la diversidad y los esfuerzos de inclusión en la industria de la tecnología como un conjunto. El lenguaje de Python tiene un código de conducta ejecutado por la Python Software Foundation la cual busca garantizar la seguridad y la inclusión de todos, tanto en las comunidades Python en línea como en persona. De igual forma hay comunidades como lo es PyLadies la cual busca la forma de crear espacios para que las personas interesadas en Python aprendan en entornos seguros e inclusivos. PyLadies es un grupo internacional de mentores que se enfoca en ayudar a más mujeres a convertirse en participantes activas y líderes en la comunidad de código abierto de Python.

Sem 1.4: Introducción al Lenguaje R

En este video, discutiremos una breve visión general acerca del lenguaje de programación R. Después de nuestro último video acerca de Python, donde discutimos su amplia adopción, puede que te estés preguntando por qué deberías considerar el aprendizaje de cualquier otro lenguaje. Bien, según los resultados de la Encuesta Kaggle de la Ciencia de los Datos del año 2019, la cual tuvo más de 10k encuestados alrededor del mundo, ¡el aprender hasta 3 lenguajes puede aumentar tu salario! Y el lenguaje R tiene mucho que ofrecerte. Así como Python, R es gratuito, pero es un Proyecto GNU – en vez de ser de código abierto, es en realidad software libre. Por lo que si Python es de código abierto R es software libre, ¿cuál es la diferencia? Bueno, tanto el código abierto como el software libre generalmente se refieren al mismo conjunto de licencias. Muchos proyectos de código abierto usan la Licencia Pública General GNU, por ejemplo. Tanto el código abierto como el software libre soportan la colaboración. En la mayoría de los casos (pero no en todos), estos términos pueden ser usados intercambiablemente. La Open Source Initiative (OSI) apoya el código abierto mientras que la Free Software Foundation (FSF) defiende el software libre. El código abierto se enfoca más en las empresas, mientras que el software libre se enfoca más en un conjunto de valores. Regresemos a por qué deberías aprender Lenguaje R. Porque este es un proyecto de software libre, puedes usar el lenguaje de la misma forma que contribuyes al código abierto y permite la colaboración pública al igual que el uso privado y comercial. Además, el lenguaje R es otro lenguaje soportado por una amplia comunidad mundial de personas apasionadas por hacer posible el uso del lenguaje para poder resolver grandes problemas. ¿Para quién es el lenguaje R? Es muy usado por estadísticos, matemáticos y mineros de datos para poder desarrollar los software estadísticos, gráficos y análisis de datos. La sintaxis orientada a los arrays del lenguaje facilita la traducción de las matemáticas al código, especialmente para alguien sin o con tan solo un poco de conocimiento de programación. De acuerdo a la Encuesta Kaggle de la Ciencia de los Datos y Aprendizaje Automático, la mayoría de las personas aprende el lenguaje R aunque ya tengan unos cuantos años en su carrera de la ciencia de los datos, pero sigue siendo un lenguaje muy cómodo para aquellos que no tiene experiencia con los software de programación. El lenguaje R es muy popular en el mundo académico, pero entre las empresas que usan R están IBM, Google, Facebook, Microsoft, Bank of America, Ford, TechCrunch, Uber, y Trulia. ● R se ha convertido en el mayor repositorio de conocimiento estadístico al nivel mundial. ● A partir del año 2018, R tiene más de 15,000 paquetes públicamente lanzados, haciendo posible realizar complejos análisis exploratorios de datos. ● R se integra muy bien con otros lenguajes de computación, como lo son C++, Java, C, .Net y Python. ● Operaciones matemáticas convencionales como lo son la multiplicación de matrices la cuales funcionan desde la base. ● R posee más facilidades de programación orientadas a los objetos en comparación de la mayoría de los lenguajes de computación estadística. Existen muchas formas de conectarse con otros usuarios de R a nivel mundial. Las comunidades como user!, WhyR?, SatRdays y R-Ladies son fabulosas para poder conectarse. Y también puedes echar un vistazo a los sitios webs de R project para saber más acerca de las conferencias y eventos R.

Sem 1.5: Otros Lenguajes

Hasta ahora, hemos discutido acerca de Python, R y SQL. En este video, discutiremos algunos otros lenguajes que tienen casos de uso convincentes para la ciencia de los datos. Bueno, entonces indiscutiblemente Python, R y SQL son los tres lenguajes más populares que utilizan los científicos de datos. Pero, hay muchos otros lenguajes que también valen la pena que te tomes un tiempo para considerar cual lenguaje usar para resolver un problema particular de la ciencia de los datos. Scala, Java, C++ y Julia son probablemente los lenguajes más tradicionales de la ciencia de los datos en esta diapositiva. ¡Pero JavaScript, PHP, Go, Ruby, Visual Basic y otros también han encontrado su lugar en la comunidad de la ciencia de los datos! No profundizaré tanto en cada uno de estos lenguajes, pero mencionaré algunos aspectos resaltantes. Java es un lenguaje de programación orientado a objetos de propósito general probado y verdadero. Este ha sido ampliamente adoptado en el espacio empresarial y está diseñado para ser rápido y escalable. Las aplicaciones Java son compiladas en código de bytes y se ejecutan en la Máquina Virtual Java o “JVM” Algunas herramientas notables de la ciencia de datos construidas con Java incluyen Weka para la minería de datos, Java-ML, el cual es una librería de aprendizaje automático, Apache MLlib el cual hace que el aprendizaje automático sea escalable y Deeplearning4j para el aprendizaje profundo. Apache Hadoop es otra aplicación basada en Java. Esta gestiona el procesamiento y almacenamiento de datos para las grandes aplicaciones de datos que son ejecutadas en sistemas clúster. Scala es un lenguaje de programación de uso general que proporciona el soporte a la programación funcional y a un sistema de tipo estático fuerte. Muchas de las decisiones de diseño en la construcción del lenguaje Scala fueron hechas para responder a las criticas de Java. Scala igualmente es interoperable con Java, ya que se ejecuta en JVM. El nombre “Scala” es una combinación de “scalable” y “language”. Este lenguaje está diseñado para crecer junto con las demandas de los usuarios. En la ciencia de los datos, el programa con mayor popularidad construido utilizando Scala es Apache Spark. Spark es un sistema de computación en clúster rápido y de propósito general. Este proporciona APIs que hacen más fácil la escritura de los trabajos paralelos y un motor optimizado que soporta los gráficos de computación general. Spark incluye Shark el cual es un motor de consulta, MLlib, para el aprendizaje automático, GraphX, para el procesamiento de gráficos y Spark Streaming. Apache Spark fue diseñado para ser más rápido que Hadoop. C++ es lenguaje de programación de propósito general. Es una extensión del lenguaje de programación C o “C con Clases”. C++ mejora la velocidad de procesamiento, permite la programación de Sistema y también permite un mejor control sobre la aplicación de software. Hay muchas organizaciones que usan Python u otros lenguajes de alto nivel para el análisis de datos y las tareas de exploración aún siguen confiando en C++ para desarrollar programas que alimentan esos datos a los clientes en tiempo real. Para la ciencia de los datos, una librería popular de aprendizaje profundo para el flujo de datos llamada TensorFlow fue construida con C++. Pero mientras que C++ es la base de TensorFlow, es ejecutada en una interfaz de Python, por lo cual no necesitas saber C++ para poder usarlo. MongoDB, una base de datos NoSQL para la gestión de grandes volúmenes de datos, fue construida con C++. Caffe es un repositorio de algoritmos de aprendizaje profundo construido con C++, Python y MATLAB bindings. Una tecnología central para la World Wide Web, JavaScript es un lenguaje de propósito general que se ha extendido más allá del navegador con la creación de Node.js y otros alcances del server side. Javascript NO está relacionado con el lenguaje Java. Para la ciencia de los datos, la implementación más popular es sin duda alguna TensorFlow.js. TensorFlow.js hace posible el aprendizaje automático y el aprendizaje profundo bien sea en Node.js o en el navegador. TensorFlow.js igualmente fue adoptado por otras librerías de código abierto incluyendo brain.js y machinelearn.js. El Proyecto R-js es otra gran implementación de JavaScript para la ciencia de los datos. R-js ha reescrito las especificaciones de algebra lineal del lenguaje R en Typescript. Esta reescritura permitirá una base para que otros proyectos implementen marcos de trabajo de base matemáticas más poderosos como lo son Numpy y SciPy de Python. Typescript es un super conjunto de JavaScript. Julia fue diseñado en el MIT para el análisis numérico de alto desempeño y la ciencia computacional. Permite un desarrollo rápido como Python o R, mientras produce programas que se ejecutan tan rápido como los programas de C o Fortran. Julia esta compilado lo que quiere decir que el código es ejecutado directamente en el procesador como código ejecutable y llama a las librerías de C, Go, Java, MATLAB, R, Fortran y Python e igualmente ha refinado el paralelismo. El lenguaje Julia es relativamente nuevo, ya que fue escrito en el año 2012, pero es muy prometedor para el futuro impacto en las industrias de la ciencia de los datos. JuliaDB es particularmente una aplicación útil de Julia para la ciencia de los datos. Es un paquete para trabajar con conjuntos de datos grandes y persistentes. Hasta ahí profundizaremos con respecto a los muchos lenguajes que son usados para resolver los problemas de la ciencia de los datos. Si tienes experiencia con algún lenguaje en particular, te recomiendo que hagas una búsqueda en la web para ver que podría ser posible en términos de usarlo en la ciencia de los datos. ¡Podrías estar sorprendido de las posibilidades!

**Herramientas de Data Science**

Sem 1.6: Categorias de las Herramientas de la Ciencia de los Datos

Las herramientas de código abierto están disponibles para diversas tareas de la ciencia de los datos. En este video, veremos las diferentes tareas de la ciencia de los datos. En los siguientes videos veremos las herramientas de código abierto más usadas para esas tareas. Las herramientas más importantes se cubren a lo largo de este curso. La Gestión de Datos es el proceso de persistencia y recuperación de datos. La Integración y Transformación de Datos, comúnmente conocida como Extracción, Transformación y Carga o por sus siglas en inglés “ETL”, es el proceso de recuperación de datos de sistemas de gestión de datos remotos. La transformación y carga de los datos en un Sistema local de gestión de datos también es parte de la Integración y Transformación de Datos. La Visualización de Datos es parte de un proceso inicial de exploración de datos, así como parte de un producto final. La Construcción de Modelo es el proceso de creación de un modelo de aprendizaje automático o aprendizaje profundo usando un algoritmo apropiado con muchos datos. El despliegue del modelo permite poner a disposición a las aplicaciones de terceros un modelo de aprendizaje automático o de aprendizaje profundo. El seguimiento y evaluación de los modelos, asegura un control continuo de la calidad del rendimiento de los modelos desplegados. Estas verificaciones son para la exactitud, equidad y la robustez del adversario. La gestión de activos de código usa el control de versiones y otras características de colaboración para facilitar el trabajo en equipo. La gestión de activos de datos implica los mismos componentes de versiones y de colaboración en los datos. La gestión de activos de datos a su vez apoya la replicación, copia de seguridad y la gestión de los derechos de acceso. Los entornos de desarrollo, comúnmente conocidos como Entornos de Desarrollo Integrado o “IDEs”, son herramientas que ayudan al científico de datos a implementar, ejecutar, probar y desplegar su trabajo. Los entornos de ejecución son herramientas en las que se realiza el pre procesamientos de datos, la capacitación de modelos y el despliegue. Finalmente, hay disponible una herramienta visual completamente integrada que cubre todos los componentes de la herramienta previa, bien sea parcial o por completo. Con esto concluye este video. En el siguiente video empezaremos a ver las herramientas de código abierto para tareas de la ciencia de los datos.

Sem 1.7: Herramientas de Código Abierto para la Ciencia de los Datos - Parte 1

En la parte uno de esta serie de dos partes, cubriremos la gestión de datos, integración de datos de código abierto, transformación y las herramientas de visualización. Las herramientas de gestión de datos de código abierto más usadas son las bases de datos relacionales como MySQL y PostgreSQL; NoSQL, MongoDB Apache CouchDB y Apache Cassandra y las herramientas basadas en archivos como Hadoop File System o los sistemas de archivos en la nube como Ceph. Por último, Elasticsearch es principalmente usado para almacenar datos de texto y crear un índice de búsqueda para una rápida recuperación de documentos. La tarea de integración de datos en el mundo clásico de almacenamiento de datos es llamado ETL que según sus siglas quiere decir “extraer, transformar y cargar”. Hoy en día los científicos de datos generalmente proponen el término “ELT” – Extraer, Cargar y Transformar “ELT” resalta el hecho de que los datos son volcados en algún lugar y el ingeniero de datos o el mismo científico de datos es el responsable de los datos. Hay otro termino para este proceso que ha surgido actualmente: “refinamiento y limpieza de datos”. Estas son las herramientas de integración y transformación de datos de código abierto más usadas: Apache AirFlow, originalmente creada por AirBNB; KubeFlow la cual permite ejecutar las tuberías de datos de la ciencia de los datos en Kubernetes; Apache Kafka la cual se originó de LinkedIn, Apache Nifi la cual ofrece un editor visual amigable, Apache SparkSQL (el cual te permite usar ANSI SQL y escalar para computar clústeres de 1000s de nodos) y NodeRED el cual también proporciona un editor visual. NodeRED consume tan pocos recursos que incluso funciona en pequeños dispositivos como Raspberry Pi. Bueno, ahora presentaremos las herramientas de visualización de datos de código abierto más usadas. Debemos distinguir entre las librerías de programación en las que debes usar el código y las herramientas que contienen una interfaz de usuario. Las librerías más populares son cubiertas en los próximos videos. Pixie Dust es una librería, pero viene con una interfaz de usuario la cual ayuda enormemente a cualquiera quien necesite graficar en Python. Hue tiene un enfoque similar el cual puede crear visualizaciones a partir de consultas en SQL. Kibana, es una aplicación web de exploración y visualización de datos está limitada a Elasticsearch (el proveedor de datos). Finalmente, Apache Superset es una aplicación web de exploración y visualización de datos. El despliegue del modelo es extremadamente importante. Una vez que has creado un modelo de aprendizaje automático capaz de predecir algunos aspectos claves del futuro, deberías hacer que ese modelo sea consumible por otros desarrolladores luego convertirlo a una API. Apache PredictionIO actualmente solo soporta modelos Apache Spark ML para el despliegue pero que sea soportada por cualquier tipo de librería está en camino. Seldon es un producto interesante ya que soporta todos los marcos de trabajo concluyendo TensorFlow, Apache SparkML, R, y scikit-learn. Seldon puede funcionar en Kubernetes y Redhat OpenShift. Otra manera de desplegar los modelos de SparkML es utilizando MLeap. Finalmente, TensorFlow puede servir con cualquiera de sus modelos utilizando el servicio de TensorFlow. Puedes desplegarlo en un dispositivo integrado como Raspberry Pi o un teléfono inteligente usando TensorFlow Lite e incluso desplegarlo en un navegador web usando TensorFlow dot JS. El monitoreo del modelo es otro paso crucial. Una vez que has desplegado un modelo de aprendizaje automático, es necesario hacer un seguimiento de su desempeño de predicción a medida que los nuevos datos llegan con el fin de mantener los modelos desactualizados. A continuación, encontrarás algunos ejemplos de herramientas de seguimiento de modelos: ModelDB es una base de metadatos de modelos de máquinas donde la información sobre los modelos puede ser almacenada y consultada. Nativamente soporta Apache Spark ML Pipelines y scikit-learn. Una herramienta genérica multi propósitos llamada Prometheus es también usada ampliamente para el seguimiento del modelo de aprendizaje automático, aunque no está hecha específicamente para este propósito. El desempeño del modelo no es medido exclusivamente con respecto a la precisión. Los modelos de tendencias contra grupos protegidos como el género o la raza también son importantes. El kit de herramientas de código abierto de IBM AI Fairness 360 hace precisamente esto. Detecta y mitiga las tendencias en los modelos de aprendizaje automático. Los modelos de aprendizaje automático, especialmente los modelos de aprendizaje profundo basados en las redes neuronales que pueden ser objeto de ataques adversos donde un atacante trata de engañar al modelo con datos manipulados o manipulando el propio modelo. El IBM Adversarial Robustness 360 Toolbox puede ser usado para detectar la vulnerabilidad de los ataques adversarios y ayuda a hacer mucho más robusto el modelo. Los modelos de aprendizaje automático frecuentemente son considerados como una caja negra que aplica algo de “magia” misteriosa. El kit de herramientas IBM AI Explainability 360 hace el proceso de aprendizaje automático mas entendible al encontrar ejemplos similares dentro del conjunto de datos que puede ser presentado a un usuario para su comparación de forma manual. El kit de herramientas IBM AI Explainability 360 también puede ilustrar la capacitación de un modelo de aprendizaje automático más simple al explicar cómo las diferentes variables de entrada afectan la decisión final del modelo. Las opciones de las herramientas de gestión de activos de código han sido ampliamente simplificadas: Para la gestión de activos de código – también conocida como gestión de versiones o control de versiones – Git es hoy en día el estándar. Han surgido múltiples servicios para soportar a Git, siendo claramente el más destacado GitHub, el cual proporciona alojamiento para la gestión de versiones de desarrollo de software. El segundo puesto es definitivamente para GitLab el cual ha tenido la ventaja de ser una plataforma completamente de código abierto que puedes alojar y administrar tú mismo. Otra opción interesante es Bitbucket. La gestión de activos de datos también conocida como gobernanza de datos o linaje de datos es otra parte crucial de la ciencia de los datos a nivel empresarial. Los datos tienen que ser versionados y anotados con metadatos. Apache Atlas es una herramienta que soporta esta tarea. Otro proyecto interesante es ODPi Egeria el cual es gestionado a través de Linux Foundation y es un entorno abierto. Este ofrece un conjunto de APIs abiertas, tipos y protocolos de intercambio que los repositorios de metadatos usan para compartir e intercambiar los datos. Y por último, Kylo que es una plataforma de software de código abierto para la gestión de Data Lake que provee un amplio soporte para la gran variedad de tareas de gestión de activos de datos. Con esto concluimos la primera parte de esta serie de dos partes. Ahora continuemos con la segunda parte.

Sem 1.8: Herramientas de Código Abierto para la Ciencia de los Datos - Parte 2

Bienvenido a la segunda parte de esta serie. En esta sección, vamos a cubrir el entorno de desarrollo, la integración de datos de código abierto, transformación y herramientas de visualización. Uno de los entornos de desarrollo actualmente más populares que los científicos de datos están utilizando es “Jupyter”. Jupyter surgió primeramente como una herramienta para la programación interactiva en Python, ahora soporta más de cien lenguajes de programación por medio de “kernels”. Kernels no deberían ser confundidos con los kernels del sistema operativo. Los Jupyter kernels están encapsulando los distintos interpretes interactivos para los diferentes lenguajes de programación. Una propiedad clave de Jupyter Notebooks es la habilidad de unificar la documentación, código, salida del código, los interpretes de comandos y las visualizaciones en un solo documento. JupyterLab es la siguiente generación de Jupyter Notebooks y a largo plazo reemplazará a Jupyter Notebooks. Los cambios a nivel de arquitectura que se están introduciendo en JupyterLab hacen que Jupyter sea más moderno y modular. Desde la perspectiva del usuario, la principal diferencia introducida por JupyterLab es la habilidad de abrir distintos tipos de archivos, incluyendo Jupyter Notebooks, datos y terminales. Entonces puedes arreglar estos archivos en el canvas. Aunque Apache Zeppelin ha sido completamente reimplantado, este está inspirado por Jupyter Notebooks y proporciona una experiencia similar. Un elemento clave es la capacidad de graficar integrado. En Jupyter Notebooks, es necesario el uso de librerías externas en Apache Zeppelin y el graficar no requiere de un código. También puedes extender estas capacidades mediante el uso de librerías adicionales. RStudio es uno de los entornos de desarrollo más antiguos para la estadística y la ciencia de los datos, siendo introducida en el año 2011. Es exclusivamente ejecutada en R y todas las librerías asociadas a R. Sin embargo, el desarrollo en Python es posible y R está estrechamente integrado en esta herramienta para proporcionar una experiencia de usuario óptima. RStudio unifica la programación, ejecución, depuración, acceso remoto a los datos, exploración de datos y visualización en una sola herramienta. Spyder intenta imitar el comportamiento de RStudio para llevar su funcionalidad al mundo de Python. Sin embargo, Spyder no tiene el mismo nivel de funcionalidad que RStudio, los científicos de datos lo consideran como una alternativa. Pero en el mundo de Python, Jupyter es usado con más frecuencia. Este diagrama muestra cómo Spyder integra el código, la documentación, las visualizaciones y otros componentes en un simple canvas. A veces tus datos no entran en la capacidad de almacenamiento o memoria principal de una sola computadora. Es allí donde entran los entornos de ejecución de clúster. El bien conocido marco de trabajo de computación en clúster Apache Spark es uno de los proyectos Apache más activos y es usado a lo largo de todas las industrias, incluyendo muchas compañías en la lista de Fortune 500. La propiedad clave de Apache Spark es la escalabilidad lineal. Lo que quiere decir que, si duplicas el número de servidores en un clúster, también duplicarás su desempeño. Después de que Apache Spark comenzó a ganar cuota de mercado y Apache Flink fue creado. La diferencia clave entre Apache Spark y Apache Flink es que Apache Spark es un motor de procesamiento de datos por lotes, capaz de procesar grandes cantidades de archivos de datos por archivo. Apache Flink, por otra parte, es una imagen de procesamiento de flujo de datos con su enfoque principal en el procesamiento de flujos de datos en tiempo real. Aunque el motor soporta ambos paradigmas de procesamientos de datos, Apache Spark es generalmente la opción en la mayoría de los casos de uso. Uno de los últimos desarrollado en los entornos de ejecución de la ciencia de los datos es llamado “Ray”, el cual se enfoca claramente en la formación de modelos de aprendizaje profundo a gran escala. Echemos un vistazo a las herramientas de código abierto para los científicos de datos que están completamente integradas y son visuales. Con estas herramientas, no necesitas conocimientos de programación. Las tareas más importantes son soportadas por estas herramientas, estas tareas incluyen la integración de datos, la transformación, la visualización de datos y la construcción de modelos. KNIME desarrollado en la Universidad de Constanza en el año 2004. Como puedes ver KNIME tiene una interfaz de usuario visual con la habilidad de arrastrar y soltar. También tiene capacidades de visualización ya incorporadas. Knime puede ser extendida para programar con R y Python y tiene conectores con Apache Spark. Otro ejemplo de este grupo de herramientas es Orange. Es menos flexible que KNIME pero mucho más fácil de usar. En esta sección, has aprendido acerca de las tareas más comunes en la ciencia de los datos y cuales herramientas de código abierto son relevantes para estas tareas. En el próximo video, discutiremos algunas herramientas comerciales establecidas que encontrarás en tu experiencia con la ciencia de datos. Vamos al siguiente video para obtener más detalles.

Sem 1.9: Herramientas Comerciales para la Ciencia de los Datos

Previamente cubrimos las herramientas de código abierto para la ciencia de los datos. Ahora, veamos las opciones comerciales que encontrarás en muchos proyectos empresariales. Vamos a repasar nuestra visión general de las distintas categorías de herramientas. En la gestión de datos, la mayoría de los datos relevantes de una empresa se almacenan en una base de datos como lo son Oracle Database, Microsoft SQL Server, o IBM Db2. Aunque las bases de datos de código abierto están ganando popularidad, esos tres productos de gestión de datos se siguen considerando como los estándares de la industria. Lo que quiere decir que no desaparecerán en un futuro cercano. No se trata solamente de funcionalidad. Los datos son el corazón de cualquier organización y la disponibilidad de soportes comerciales juegan un papel muy importante. El soporte comercial es ofrecido directamente por los proveedores de software, los socios influyentes y las redes de soporte. Cuando nos enfocamos en las herramientas comerciales de integración de datos, hablamos de herramientas de “extracción, transformación y carga” o “ETL”. De acuerdo a Gartner Magic Quadrant (Cuadrante Mágico de Gartner), Informatica Powercenter e IBM InfoSphere DataStage son los líderes, seguidos por los productos de SAP, Oracle, SAS, Talend y Microsoft. Estas herramientas apoyan el diseño y despliegue de las tuberías de procesamiento de datos de ETL a través de una interfaz gráfica. También proporcionan conectores a la mayoría de los sistemas de información de destino comercial al igual que para los de código abierto. Finalmente, Watson Studio Desktop incluye un componente llamado Data Refinery, el cual permite La definición y ejecución de procesos de integración de datos en un estilo de hoja de cálculo. En el ámbito comercial, las visualizaciones de datos utilizan herramientas de inteligencia empresarial o “BI”. Su enfoque principal es crear informes visualmente atractivos y fáciles de comprender y tableros de control en vivo. Los ejemplos comerciales más importantes son: Tableau, Microsoft Power BI e IBM Cognos Analytics. Otro tipo de visualización se destina a los científicos de los datos en lugar de a los usuarios habituales. Un problema de muestra podría ser “¿Cómo pueden relacionarse las distintas columnas de una tabla?”. Este tipo de funcionalidad está incluida en Watson Studio Desktop. Si quieres construir un modelo de aprendizaje automático usando una herramienta comercial, deberías considerar el uso de un producto de minería de datos. Los más importantes de estos tipos de productos son: SPSS Modeler y SAS Enterprise Miner. Además, una versión de SPSS Modeler está igualmente disponible en Watson Studio Desktop, basada en la versión de la nube la de herramienta. Discutiremos más sobre las herramientas basadas en la nube en el próximo video. En el software comercial, el despliegue de modelos está estrechamente integrado en el proceso de construcción de modelos. Este diagrama muestra un ejemplo de SPSS Collaboration and Deployment Services en el cual son usados para desplegar cualquier tipo de activo creado por la suite de herramientas de Software SPSS. Otros proveedores usan el mismo tipo de proceso. Los software comerciales igualmente pueden exportar modelos en formato abierto. Ejemplo, SPSS Modeler permite la exportación de modelos como Predictive Model Markup Language, o PMML, el cual puede ser leído por numerosos paquetes de software comerciales y abiertos. Model monitoring es una nueva disciplina y actualmente no existen herramientas comerciales relevantes disponibles. Por lo tanto, el código abierto es la primera opción. Esto mismo pasa con la gestión de activos de código. El código abierto con Git y GitHub es el estándar efectivo. La gestión de activos de código, con frecuencia llamado gobernanza de datos o linaje de datos, es una parte fundamental de la ciencia de datos a nivel empresarial. Los datos deben ser versionados y anotados utilizando metadatos. Los proveedores, incluyendo Informatica Enterprise Data Governance e IBM, proporcionan herramientas para estas tareas específicas. IBM InfoSphere Information Governance Catalog cubre las funciones como el diccionario de datos, lo cual facilita el descubrimiento de los activos de datos. Cada activo de datos se asigna a un administrador de datos - - el propietario de los datos. El propietario de los datos es responsable de ese activo de datos y puede ser contactado. El linaje de datos también es cubierto; esto permite al usuario hacer un seguimiento de los pasos de transformación seguidos en la creación de los activos de datos. El linaje de datos también incluye una referencia a los daos de la fuente real. Se pueden agregar reglas y políticas para reflejar los complejos requisitos normativos y comerciales de privacidad y retención de datos. Watson Studio es un entorno de desarrollo completamente integrado para los científicos de datos. Es normalmente consumido a través de la nube, y cubriremos más sobre este en una siguiente lección. También hay una versión de escritorio disponible. Watson Studio Desktop combina Jupyter Notebooks con herramientas gráficas para maximizar el rendimiento de los científicos de datos. Watson Studio, conjuntamente con Watson Open Scale, es una herramienta completamente integrada que cubre el ciclo de vida completa de la ciencia de los datos y todas las tareas que hemos discutido previamente. Hablaremos más de ambos en la siguiente lección. Tan solo ten en cuenta que pueden ser desplegados en un centro de datos local que pueden ser Kubernetes o RedHat OpenShift. Otro ejemplo de una herramienta comercial totalmente integrada es H2O Driverless AI, la cual cubre el ciclo de vida completo de la ciencia de los datos. En esta lección, has aprendido cómo las tareas más comunes de la ciencia de los datos son soportadas por las herramientas comerciales. En el siguiente video, descubriremos herramientas de la ciencia de los datos que están disponibles exclusivamente en la nube.

Sem 1.10: Herramientas Comerciales para la Ciencia de los Datos

Previamente cubrimos las herramientas de código abierto para la ciencia de los datos. Ahora, veamos las opciones comerciales que encontrarás en muchos proyectos empresariales. Vamos a repasar nuestra visión general de las distintas categorías de herramientas. En la gestión de datos, la mayoría de los datos relevantes de una empresa se almacenan en una base de datos como lo son Oracle Database, Microsoft SQL Server, o IBM Db2. Aunque las bases de datos de código abierto están ganando popularidad, esos tres productos de gestión de datos se siguen considerando como los estándares de la industria. Lo que quiere decir que no desaparecerán en un futuro cercano. No se trata solamente de funcionalidad. Los datos son el corazón de cualquier organización y la disponibilidad de soportes comerciales juegan un papel muy importante. El soporte comercial es ofrecido directamente por los proveedores de software, los socios influyentes y las redes de soporte. Cuando nos enfocamos en las herramientas comerciales de integración de datos, hablamos de herramientas de “extracción, transformación y carga” o “ETL”. De acuerdo a Gartner Magic Quadrant (Cuadrante Mágico de Gartner), Informatica Powercenter e IBM InfoSphere DataStage son los líderes, seguidos por los productos de SAP, Oracle, SAS, Talend y Microsoft. Estas herramientas apoyan el diseño y despliegue de las tuberías de procesamiento de datos de ETL a través de una interfaz gráfica. También proporcionan conectores a la mayoría de los sistemas de información de destino comercial al igual que para los de código abierto. Finalmente, Watson Studio Desktop incluye un componente llamado Data Refinery, el cual permite La definición y ejecución de procesos de integración de datos en un estilo de hoja de cálculo. En el ámbito comercial, las visualizaciones de datos utilizan herramientas de inteligencia empresarial o “BI”. Su enfoque principal es crear informes visualmente atractivos y fáciles de comprender y tableros de control en vivo. Los ejemplos comerciales más importantes son: Tableau, Microsoft Power BI e IBM Cognos Analytics. Otro tipo de visualización se destina a los científicos de los datos en lugar de a los usuarios habituales. Un problema de muestra podría ser “¿Cómo pueden relacionarse las distintas columnas de una tabla?”. Este tipo de funcionalidad está incluida en Watson Studio Desktop. Si quieres construir un modelo de aprendizaje automático usando una herramienta comercial, deberías considerar el uso de un producto de minería de datos. Los más importantes de estos tipos de productos son: SPSS Modeler y SAS Enterprise Miner. Además, una versión de SPSS Modeler está igualmente disponible en Watson Studio Desktop, basada en la versión de la nube la de herramienta. Discutiremos más sobre las herramientas basadas en la nube en el próximo video. En el software comercial, el despliegue de modelos está estrechamente integrado en el proceso de construcción de modelos. Este diagrama muestra un ejemplo de SPSS Collaboration and Deployment Services en el cual son usados para desplegar cualquier tipo de activo creado por la suite de herramientas de Software SPSS. Otros proveedores usan el mismo tipo de proceso. Los software comerciales igualmente pueden exportar modelos en formato abierto. Ejemplo, SPSS Modeler permite la exportación de modelos como Predictive Model Markup Language, o PMML, el cual puede ser leído por numerosos paquetes de software comerciales y abiertos. Model monitoring es una nueva disciplina y actualmente no existen herramientas comerciales relevantes disponibles. Por lo tanto, el código abierto es la primera opción. Esto mismo pasa con la gestión de activos de código. El código abierto con Git y GitHub es el estándar efectivo. La gestión de activos de código, con frecuencia llamado gobernanza de datos o linaje de datos, es una parte fundamental de la ciencia de datos a nivel empresarial. Los datos deben ser versionados y anotados utilizando metadatos. Los proveedores, incluyendo Informatica Enterprise Data Governance e IBM, proporcionan herramientas para estas tareas específicas. IBM InfoSphere Information Governance Catalog cubre las funciones como el diccionario de datos, lo cual facilita el descubrimiento de los activos de datos. Cada activo de datos se asigna a un administrador de datos - - el propietario de los datos. El propietario de los datos es responsable de ese activo de datos y puede ser contactado. El linaje de datos también es cubierto; esto permite al usuario hacer un seguimiento de los pasos de transformación seguidos en la creación de los activos de datos. El linaje de datos también incluye una referencia a los daos de la fuente real. Se pueden agregar reglas y políticas para reflejar los complejos requisitos normativos y comerciales de privacidad y retención de datos. Watson Studio es un entorno de desarrollo completamente integrado para los científicos de datos. Es normalmente consumido a través de la nube, y cubriremos más sobre este en una siguiente lección. También hay una versión de escritorio disponible. Watson Studio Desktop combina Jupyter Notebooks con herramientas gráficas para maximizar el rendimiento de los científicos de datos. Watson Studio, conjuntamente con Watson Open Scale, es una herramienta completamente integrada que cubre el ciclo de vida completa de la ciencia de los datos y todas las tareas que hemos discutido previamente. Hablaremos más de ambos en la siguiente lección. Tan solo ten en cuenta que pueden ser desplegados en un centro de datos local que pueden ser Kubernetes o RedHat OpenShift. Otro ejemplo de una herramienta comercial totalmente integrada es H2O Driverless AI, la cual cubre el ciclo de vida completo de la ciencia de los datos. En esta lección, has aprendido cómo las tareas más comunes de la ciencia de los datos son soportadas por las herramientas comerciales. En el siguiente video, descubriremos herramientas de la ciencia de los datos que están disponibles exclusivamente en la nube.

Sem 1.11: Herramientas basadas en la nube para la ciencia de los datos

Dado que anteriormente cubrimos las herramientas de código abierto para la ciencia de datos, veamos las opciones comerciales que encontrará en muchos proyectos empresariales. Eche otro vistazo a la descripción general de las diferentes categorías de herramientas. Dado que los productos en la nube son una especie más nueva, siguen la tendencia de tener múltiples tareas integradas en las herramientas. Esto es especialmente cierto para las tareas marcadas en verde en el diagrama. Comencemos con la categoría de herramientas visuales totalmente integradas. Dado que estas herramientas introducen un componente donde la ejecución a gran escala de flujos de trabajo de ciencia de datos ocurre en clústeres de cómputo, aquí cambiamos el título y agregamos la palabra "Plataforma". Estos clústeres están compuestos por varias máquinas servidor, de forma transparente para el usuario, en segundo plano. Watson Studio, junto con Watson OpenScale, cubre el ciclo de vida de desarrollo completo para todas las tareas de ciencia de datos, aprendizaje automático e IA. Otro ejemplo es Microsoft Azure Machine Learning. Esta es también una oferta totalmente alojada en la nube que respalda el ciclo de vida de desarrollo completo de todas las tareas de ciencia de datos, aprendizaje automático e IA. Y finalmente, otro ejemplo es H2O Driverless AI, que ya presentamos en el último video. Aunque es un producto que descarga e instala, la implementación con un solo clic está disponible para los proveedores de servicios en la nube comunes. Dado que las operaciones y el mantenimiento no los realiza el proveedor de la nube, como es el caso de Watson Studio, Open Scale y Azure Machine Learning, este modelo de entrega no debe confundirse con la plataforma o el software como servicio: PaaS o SaaS. En la gestión de datos, con algunas excepciones, existen versiones SaaS de herramientas comerciales y de código abierto existentes. Recuerde, SaaS significa "software como servicio". Significa que el proveedor de la nube opera la herramienta por usted en la nube. Como ejemplo, el proveedor de la nube opera el producto haciendo una copia de seguridad de sus datos y configuración e instalando actualizaciones. Como se mencionó, existen herramientas propietarias, que solo están disponibles como un producto en la nube. A veces, solo está disponible a través de un único proveedor de nube. Un ejemplo de dicho servicio es Amazon Web Services DynamoDB, una base de datos NoSQL que permite el almacenamiento y la recuperación de datos en un formato de almacenamiento de documentos o clave-valor. La estructura de datos de documentos más destacada es JSON (pronunciado "jay-sun"). Otro tipo de servicio de este tipo es Cloudant, que es una oferta de base de datos como servicio. Pero, bajo el capó, se basa en Apache CouchDB de código abierto. Tiene una ventaja: aunque el proveedor de la nube realiza tareas operativas complejas como la actualización, la copia de seguridad, la restauración y el escalado, en el fondo esta oferta es compatible con CouchDB. Por lo tanto, la aplicación se puede migrar a otro servidor CouchDB sin cambiar la aplicación. E IBM también ofrece Db2 como servicio. Este es un ejemplo de una base de datos comercial disponible como una oferta de software como servicio en la nube, que quita las tareas operativas al usuario. Cuando se trata de herramientas comerciales de integración de datos, no solo hablamos de herramientas de "extracción, transformación y carga" o "ETL", sino también de herramientas de "extracción, carga y transformación" o "ELT". Esto significa que los pasos de transformación no los realiza un equipo de integración de datos, sino que se llevan al dominio del científico de datos o el ingeniero de datos. Dos herramientas comerciales de integración de datos ampliamente utilizadas son Informatica Cloud Data Integration y IBM's Data Refinery. Data Refinery permite la transformación de grandes cantidades de datos sin procesar en información consumible y de calidad en una interfaz de usuario similar a una hoja de cálculo. Data Refinery es parte de IBM Watson Studio. El mercado de las herramientas de visualización de datos en la nube es enorme y todos los principales proveedores de la nube tienen una. Un ejemplo de la herramienta de visualización de datos basada en la nube de una empresa más pequeña es DataMeer. IBM también ofrece su famosa suite de inteligencia empresarial Cognos como solución en la nube. IBM Data Refinery también ofrece funciones de exploración y visualización de datos en Watson Studio. Nuevamente, estos son solo algunos ejemplos de un ecosistema comercial que cambia y crece rápidamente entre una gran cantidad de proveedores establecidos y emergentes. En Watson Studio, se puede utilizar una gran cantidad de visualizaciones diferentes para comprender mejor los datos. Por ejemplo, este gráfico de barras 3D le permite visualizar un valor objetivo en la dimensión vertical, que depende de otros dos valores en las dimensiones horizontales. Colorear te permite visualizar una tercera dimensión. La agrupación perimetral jerárquica le permite visualizar correlaciones y afiliaciones entre entidades. Si es suficiente, un gráfico de barras clásico también puede hacer el trabajo, mientras que un gráfico de dispersión 2D con un mapa de calor muestra dos campos de datos dependientes, uno en el eje y y otro como intensidad de color. Un mapa de árbol muestra la distribución de subconjuntos dentro de un conjunto, el famoso gráfico circular hace lo mismo pero de una manera no jerárquica y, finalmente, una nube de palabras muestra términos significativos en un corpus de documentos. La creación de modelos se puede realizar mediante un servicio como Watson Machine Learning. Watson Machine Learning puede entrenar y construir modelos utilizando varias bibliotecas de código abierto. Google tiene un servicio similar en su nube llamado AI Platform Training. Casi todos los proveedores de la nube tienen una solución para esta tarea. La implementación de modelos en software comercial suele estar estrechamente integrada con el proceso de creación de modelos. Este es un ejemplo de SPSS Collaboration and Deployment Services, que se puede utilizar para implementar cualquier tipo de activo creado por el paquete de herramientas de software de SPSS. Lo mismo vale para otros vendedores. Además, el software comercial puede exportar modelos en un formato abierto. Por ejemplo, SPSS Modeler admite la exportación de modelos como Predictive Model Markup Language, o "PMML", que pueden leer muchos otros paquetes de software abiertos y comerciales. Watson Machine Learning también se puede utilizar para implementar un modelo y ponerlo a disposición de los consumidores mediante una interfaz REST. Amazon SageMaker Model Monitor es un ejemplo de una herramienta en la nube que monitorea continuamente los modelos de aprendizaje automático y aprendizaje profundo implementados. Una vez más, todos los principales proveedores de la nube tienen herramientas similares. Este es también el caso de Watson OpenScale. OpenScale y Watson Studio… …unifican el paisaje. Todo lo marcado en verde se puede hacer con Watson Studio y Watson OpenScale. Cubriremos Open Scale en un video posterior. Aprendió cómo las tareas más comunes en la ciencia de datos son compatibles con herramientas comerciales en la nube. La integración nos proporciona la capacidad de utilizar las mismas herramientas para múltiples tareas. En los próximos videos, veremos paquetes, API, conjuntos de datos y modelos para la ciencia de datos.

**Paquetes, APIs, conjuntos de datos y modelos**

Sem 1.12: Librerías de la Ciencia de los Datos

En este video, revisaremos diferentes librerías de la ciencia de los datos. Las librerías son una colección de funciones y métodos los cuales te permiten realizar una gran variedad de acciones sin tener que escribir el código tú mismo. Nos enfocaremos en la librería de Python: Librerías de Computación Científica en Python Librerías de Visualización en Python, Librerías de Aprendizaje Automático de Alto Nivel y Librerías de Aprendizaje Profundo – “Alto Nivel” sencillamente quiere decir que no tienes que preocuparte por los detalles, aunque esto dificulta el estudio o la mejora de las Librerías de Aprendizaje Profundo en Python Las librerías en otros lenguajes generalmente las librerías contienen módulos incorporados que proporcionan distintas funcionalidades que puedes usar directamente; algunas veces son llamados “frameworks o marcos de trabajo”. También hay muchas librerías, que ofrecen una amplia variedad de posibilidades. Pandas ofrece estructuras de datos y herramientas para la depuración, manipulación y análisis de datos efectivos. Esto proporciona herramientas para trabajar con los distintos tipos de datos. El instrumento principal de Pandas es una tabla de dos dimensiones que consiste en columnas y filas. Esta tabla es llamada “DataFrame” y está diseñada para proporcionar una fácil indexación para que puedas trabajar con tus datos. Las librerías NumPy están basadas en matrices, lo que te permite aplicar funciones matemáticas a estas matrices. Pandas está realmente construido basado en NumPy Los métodos de Visualización de Datos son una grandiosa forma de comunicarse con los demás y mostrar los resultados significativos del análisis. Estas librerías te permiten crear gráficos, tablas y mapas. El paquete Matplotlib es la librería más conocida para la visualización de datos y es excelente para hacer gráficos y diagramas. Igualmente, los gráficos son altamente personalizables. Otra librería de visualización de alto nivel, es Seaborn, el cual está basado en matplotlib. Seaborn hace fácil la generación de gráficos como mapas de calor, series temporales y gráficos de violín. Para el aprendizaje automático, la librería Scikit-learn contiene herramientas para el modelado estadístico, incluyendo regresión, clasificación, clustering o agrupación entre otros. Está construida basada en NumPy, SciPy y matplotlib y es prácticamente para comenzar. Para este enfoque de alto nivel, tu defines el modelo y especificas los tipos de parámetros que te gustaría usar. Para el aprendizaje profundo, Keras te permite construir el modelo estándar de aprendizaje profundo. Así como Scikit-learn, la interfaz de alto nivel te permite construir modelos de forma rápida y simple. Esto puede funcionar usando las unidades de procesamiento de gráficos (GPU), pero para muchos casos de aprendizaje profundo un entorno de nivel menor. TensorFlow es un marco de nivel bajo usado en la producción a gran escala de modelos de aprendizaje profundo. Está diseñado para la producción, pero puede ser difícil de manejar para la experimentación. Pytorch es usado para la experimentación, haciéndolo que sea simple para los investigadores probar sus ideas. Apache Spark es un marco de computación de clúster de propósito general el cual permite el procesamiento de datos usando clústeres de computación. Esto quiere decir que procesas los datos de forma paralela, utilizando múltiples computadoras simultáneamente. La librería Spark una funcionalidad muy similar a la de Pandas Numpy Scikit-learn Apache Spark data processing jobs puede usar Python R Scala o SQL Hay muchas librerías para Scala, la cual se usa predominantemente en la ingeniería de datos pero de igual forma se usa a veces en la ciencia de los datos. Vamos a discutir algunas de las librerías que son complementarias de Spark Vegas es una librería Scala para la visualización de datos estadísticos. Con Vegas, puedes trabajar con los archivos de datos, así como Spark DataFrames. Para el aprendizaje profundo, puedes usar BigDL. R tiene una funcionalidad incorporada para el aprendizaje automático y la visualización de datos, pero también existen diferentes librerías complementarias: ggplot2 es una librería popular para la visualización de datos en R. Igualmente puedes utilizar librerías que te permiten hacer la interfaz con Keras y TensorFlow. R ha sido el estándar de-facto para la ciencia de datos de código abierto, pero ahora está siendo sustituido por Python.

Sem 1.13: Interfaces de Programación de Aplicaciones

En este vídeo discutiremos acerca de las Interfaces de Programación de Aplicaciones o también conocidas como APIs. Específicamente, discutiremos: Qué es una API Librerías API – API REST, incluyendo: Solicitud y Respuesta Una API permite que dos elementos de un software se hablen entre sí. Por ejemplo, tienes tu programa, tienes tus datos, tienes otros componentes de software. Usas la API para comunicarte con otros componentes de software. Usas la API para comunicarte con la API usando las entradas y salidas. No tienes que saber cómo funciona la API, solamente necesitas saber sus entradas y salidas. Recuerda, que la API sólo se refiere a la interfaz, o a la parte de la librería que tú ves. La “Librería” se refiere a todo el tema. Considera la librería pandas. Pandas es en realidad un conjunto de componentes de software, muchos de los cuales ni siquiera están escritos en Python. Tienes algunos datos. Tienes un conjunto de componentes de software. Nosotros usamos la Pandas API para procesar los datos comunicándonos con los otros componentes del software. Puede haber tan solo un componente de software en el back end, pero puede estar separado La API para diferentes lenguajes. Considera TensorFlow, escrito en C++. Hay APIs separadas en Python JavaScript C++ Java Go. La API es sencillamente la interfaz. También hay múltiples APIs desarrolladas por voluntarios para TensorFlow; como por ejemplo Julia MATLAB R Scala y muchas más. Las APIs REST son otro tipo de API muy popular. Estas te permiten comunicarte usando la Internet, aprovechando el almacenamiento y mayor acceso a los datos, algoritmos de inteligencia artificial y muchos otros recursos. RE significa “Representación” la S significa “Estado” y la T Significa “Transferencia”. En la API REST, tu programa es llamado el “cliente”. La API se comunica con un servicio web el cual llamas a través de Internet. Hay un conjunto de reglas que rigen (clic 4)Comunicación, (clic 5)Entrada o Solicitud, y (clic 6 ) Salida O Respuesta. Aquí hay algunos términos comunes relacionados con la API. Tú o tú código pueden ser vistos como un cliente. El servicio web se refiere a un recurso. El cliente encuentra el servicio a través de un punto final. El cliente envía la solicitud al recurso y la respuesta al cliente. Los métodos HTTP son una forma de transmitir datos a través de Internet le decimos a las APIs REST qué hacer enviando una solicitud. La solicitud es usualmente comunicada a través de un mensaje HTTP. Generalmente el mensaje HTTP contiene un archivo JSON, que contiene las instrucciones para la operación que nos gustaría que el servicio realizara. Esta operación se transmite al servicio web a través de Internet. El servicio realiza la operación. Similarmente, el servicio web devuelve una respuesta a través de un mensaje HTTP, donde la información se devuelve generalmente usando un archivo JSON. Esta información es transmitida de vuelta al cliente. La API Watson Text to Speech es un ejemplo de una API REST. Esta API convierte la voz a texto. En las API de llamadas, se envía una copia del archivo de audio a la API; este proceso es llamado solicitud de envío. La API envía la transcripción de texto de lo que el individuo está diciendo. La API está haciendo una solicitud de obtención. La API Watson Language-Translator provee otro ejemplo. Envías el texto que quieres traducir a la API, y ésta traduce el texto y te envía la traducción. En este caso traducimos de inglés a español. En este video, hemos discutido lo que es una API Librerías API – API REST, incluyendo Solicitud y Respuesta Gracias por ver este video.

Sem 1.17: Conjuntos de datos - Mejorando la ciencia de los datos

En este video discutiremos los conjuntos de datos: qué son, por qué son tan importantes en la ciencia de los datos, y donde encontrarlos. Primero definamos brevemente lo que es un conjunto de datos. Un conjunto de datos es una colección estructurada de datos. Los datos incluyen información que podría ser representada como texto, números o archivos multimedia como imágenes, audio o video. Un conjunto de datos que es estructurado como datos tabulares comprende una colección de filas, que a su vez Comprenden columnas que almacenan la información. Un formato popular de datos tabulados es el de “valores separados por comas” o CSV. Un archivo de tipo CSV es un archivo de texto delimitado en el que cada línea representa una fila y los valores de los datos están separados por una coma. Por ejemplo, imagina un conjunto de datos de observaciones de una estación meteorológica. Cada fila representa una observación en un momento determinado, mientras que cada columna contiene la información sobre esa observación en particular, como la temperatura, humedad y cualquier otra condición meteorológica. Las estructuras de datos jerárquicas o de red son normalmente usadas para representar las relaciones entre los datos. Los datos jerárquicos son organizados en una estructura de árbol, mientras que los datos de red pueden ser almacenados de forma gráfica. Por ejemplo, las conexiones entre personas en un sitio web de redes sociales normalmente son representadas en forma de gráfico. Un conjunto de datos también puede incluir archivos de datos sin procesar, como lo son las imágenes o audios. El conjunto de datos MNIST es muy popular para la ciencia de los datos. Contiene imagines de dígitos escritos a mano y es comúnmente usado para entrenar los sistemas de procesamiento de imágenes. Tradicionalmente, la mayoría de los conjuntos de datos eran considerados privados porque contenían información patentada o confidencial, como por ejemplo datos de clientes, datos de precios o cualquier otra información comercialmente sensible. Estos conjuntos de datos no son normalmente compartidos públicamente. Con el tiempo, más y más entidades públicas y privadas como instituciones científicas, gobiernos, organizaciones e incluso empresas han comenzado a poner a disposición del público conjuntos de datos como “datos abiertos”, que permiten obtener una gran cantidad de información de manera gratuita. Por ejemplo, las Naciones Unidas al igual que gobiernos federales y municipales alrededor del mundo han publicado muchos conjuntos de datos en sus sitios web, que incluyen economía, sociedad, asistencia media, transporte, medio ambiente y mucho más. El acceso a estos y otros conjuntos de datos abiertos permite a los científicos de datos, investigadores, analistas y otras personas descubrir conocimientos anteriormente desconocidos, pero potencialmente útiles. Ellos pueden crear nuevas aplicaciones bien sea para fines comerciales como para el bien público. De igual forma pueden llevar a cabos nuevas investigaciones. Los datos abiertos han jugado un papel muy importante en el crecimiento de la ciencia de los datos, el aprendizaje automático y la inteligencia artificial y han proporcionado a los profesionales una forma de perfeccionar sus habilidades sobre una amplia variedad de conjuntos de datos. Existen muchas fuentes de datos abiertos en Internet. Puedes encontrar una lista extensa de portales de datos abiertos en todo el mundo en el sitio web de la Fundación de Conocimiento Abierto (FCA) datacatalogs.org. Las Naciones Unidas, la Unión Europea y muchas otras organizaciones gubernamentales e intergubernamentales mantienen repositorios de datos los cuales dan acceso a una amplia variedad de información. En Kaggle, la cual es una popular comunidad en línea de ciencia de los datos, podrás encontrar y contribuir con conjuntos de datos que podrían ser de interés general. Por último, pero no menos importante, Google proporciona un motor de búsqueda de conjuntos de datos que podrían ayudarte a encontrar los que tengan un valor particular para ti. Es importante reconocer que la distribución y el uso de datos abiertos podrían estar restringidos, según lo definan las condiciones de su licencia. En caso de la falta de una licencia para la distribución de datos abiertos, muchos conjuntos de datos eran compartidos en el pasado bajo licencias de software de código abierto. Estas licencias no fueron diseñadas para cubrir las consideraciones especificas relacionadas con la distribución y uso de los conjuntos de datos. Para tratar el tema, Linux Foundation creó el Community Data License Agreement, o CDLA. Dos licencias fueron inicialmente creadas para compartir datos: CDLA-Sharing y CDLA-Permissive. La licencia CDLA-Sharing te otorga permiso para utilizar y modificar los datos. La licencia estipula que si publicas tu versión modificada de los datos debes hacerlo bajo los mismos términos de la licencia de los datos originales. La licencia CDLA-Permissive igualmente te otorga permiso para utilizar y modificar los datos. Sin embargo, no estás obligado a compartir los cambios que le hagas a los datos. Debes tener en cuenta que ninguna de las dos licencias impone restricciones en los resultados que podría conseguir al usar los datos, lo cual es importante en la ciencia de los datos. Digamos, por ejemplo, que estas construyendo un modelo el cual realiza una predicción. Si estás capacitando a el modelo usando conjuntos de datos bajo la licencia de CDLA, no estás obligado a compartir el modelo, o a compartirlo bajo una licencia especifica si decides compartirlo. En este video has aprendido acerca de los conjuntos de datos abiertos, su rol en la ciencia de los datos y donde puedes encontrarlos. También hemos visto algo de Community Data License Agreement, el cual facilita el intercambio y como compartir los datos abiertos. Un aspecto importante que no hemos cubierto en este video es la calidad y precisión de los datos, los cuales podrían variar mucho dependiendo de quien haya recolectado y contribuido al conjunto de datos. Mientras que algunos conjuntos de datos abiertos pueden ser lo suficientemente buenos para uso personal, es posible que no cumplan los requisitos de la empresa debido al impacto que pueden tener en el negocio. En el próximo modulo, aprenderás acerca de Data Asset eXchange, el cual es un repositorio de datos abierto y curado.

Sem 1.18: Intercambio de activos de datos

A pesar del crecimiento de los conjuntos de datos abiertos que están disponibles al público, aun asípuede ser difícil descubrir conjuntos de datos que sean a la vez de alta calidad y que tengantérminos de licencia y uso claramente definidos.Para ayudar a resolver este reto, IBM creó Data Asset eXchange, o “DAX”, el cualque mostraremos en este vídeo.DAX proporciona una fuente de confianza para encontrar conjuntos de datos abiertos que están listos para ser usados enaplicaciones empresariales.Estos conjuntos de datos los cuales cubren una amplia variedad de dominios, incluyendo imagines, video, textoy audio.Debido a que DAX proporciona un alto nivel de conservación de la calidad de los conjuntos de datos, así como las licenciasy condiciones de uso, los conjuntos de datos DAX usualmente son fáciles de adoptar, bien sea en proyectos de investigacióno comerciales.Siempre que sea posible, DAX apunta a poner a disposición los conjuntos de datos en el ámbito de una de las variantesde CDLA Community Data License Agreement, con el fin de fomentar el intercambio de datos y la colaboración.DAX también provee un lugar único para acceder a los conjuntos de datos únicos, en particular de los proyectos deInvestigación de IBM.Para facilitar a los desarrolladores el inicio en el uso de los conjuntos de datos, DAX también proveetutoriales notebooks que explican los fundamentos de la limpieza de datos, pre-procesamientoy el análisis exploratorio.Para algunos conjuntos de datos, también hay notebooks que ilustran cómo realizar análisis más complejos,como la creación de gráficos, análisis estadísticos, análisis de series cronológico, capacitación en los modelos de aprendizaje automáticoy la integración de aprendizaje profundo a través del uso de Model Asset eXchange, (un proyectomuy relacionado con DAX y también está disponible en el sitio web de IBM Developer).De esta forma, DAX ayuda a los desarrolladores a crear flujos de trabajo analíticos y de aprendizaje automático de principio a finy a consumir datos y modelos abiertos con seguridad bajo los términos de licencia claramente definidos.Digamos que has encontrado un conjunto de datos que podrían ser de tu interés.En la página de conjunto de datos podrás descargar el archivo comprimido de los conjuntos de los datos de almacenamiento en la nube,explore el conjunto de los datos usando Jupyter Notebooks, revisar los metadatos del conjunto de datos, así como el formato,términos de licencia y el tamaño y una vista previa de algunas partes de los conjuntos de datos.La mayoría de los conjuntos de datos en DAX son complementados con uno o más Jupyter Notebooks que puedesusar para realizar la limpieza de datos, pre-procesamiento y el análisis exploratorio.Estos notebooks funcionan “tal cual” como en Watson Studio, la plataforma de ciencia Data Sciencedata de IBM.Jupyter Notebooks y Watson Studio serán discutidos más adelante durante este curso.En este video, has aprendido sobre el repositorio de datos abiertos de IBM y acerca de Data Asset eXchange.En el laboratorio practico tendrás la oportunidad de explorar el repositorio.

Sem 1.19: Modelos de Aprendizaje Automático

En este video, te presentaremos los modelos de aprendizaje automático al igual que aprendizaje profundo.Los datos contienen una enorme cantidad de información que puede ser usada para resolver ciertos tipos de problemas.El alcance tradicional del análisis de datos, como una persona que inspecciona manualmente los datos oun programa computarizado especializado que automatice el análisis de los humanos, los cuales llegan rápidamente a sus limitesdebido a la cantidad de datos para ser analizados o a la complejidad del problema.El aprendizaje automático usa algoritmos – también conocidos como “modelos” – para identificar los patronesen los datos.El proceso por el cual el modelo aprende estos patrones a partir de los datos es llamado “capacitación del modelo”.Una vez que un modelo es capacitado, entonces puede ser usado para hacer predicciones.Una vez que se le presentan los nuevos datos al modelo, intenta hacer predicciones o decisionesbasadas en los patrones que ha aprendido de los datos anteriores.Los modelos de aprendizaje automático pueden ser divididos en tres clases básicas: aprendizaje supervisado,aprendizaje no supervisado y aprendizaje reforzado.El aprendizaje supervisado es uno de los modelos de aprendizaje automático mas comúnmente usados.En el aprendizaje supervisado, un humano provee los datos de entrada y las salidas correctas.El modelo intenta identificar las relaciones y dependencias entre los datos de entrada yla salida correcta.En términos generales, el aprendizaje supervisado es usado para resolver los problemas de regresióny clasificación.Ahora vemos un ejemplo para cada tipo de problema:Los modelos de regresión son usados para predecir un valor numérico “real”.Por ejemplo, dada la información acerca de las ventas anteriores de casas, como lo son la ubicación geográfica, tamaño,número de cuartos y precio de venta, puedes capacitar un modelo para predecir el precio de ventaestimado para otras casas con características muy similares.Los modelos de clasificación son usados para predecir si algo pertenece a una categoría o“clase”Por ejemplo, dado un conjunto de correos electrónicos con una designación de si son consideradoso no como spam, un algoritmo puede ser capacitado para identificar los correos electrónicos no solicitados.En el aprendizaje no supervisado, los datos no son etiquetados por un humano.Los modelos deben analizar los datos e intentar identificar los patrones y la estructura dentro de losdatos basándose solamente en las características de los propios datos.La agrupación y detección de anomalías son dos ejemplos claros de este tipo de aprendizaje.Los modelos de agrupación son usados para dividir cada registro de un conjunto de datos en uno de un pequeño numerode grupos similares.Un ejemplo de un modelo de agrupación podría ser proporcionar recomendaciones de compra para unatienda e-commerce basada en el comportamiento de las compras anteriores y el contenido de la cesta de compra.La detección de anomalías identifica los valores atípicos en un conjunto de datos, tales como, transacciones fraudulentas con tarjetas de créditoo intentos sospechosos de inicios de sesión en línea.El tercer tipo de aprendizaje, el cual es aprendizaje reforzado, esta vagamente basado en la forma en que los sereshumanos y otros organismos aprenden.Piensa en un ratón en un laberinto.Si el ratón llega al final del laberinto, él obtiene un trozo de queso.Esta es la “recompensa” por haber terminado la tarea.El ratón aprende – por ensayo y error – cómo recorrer el laberinto para conseguirtodo el queso que pueda.De forma similar, un modelo de aprendizaje reforzado aprende el mejor conjunto de acciones que debe tomar,dado su entorno actual, para obtener la mayor recompensa a lo largo del tiempo.Este tipo de aprendizaje ha sido muy exitoso recientemente al vencer a los mejores jugadores humanosen juegos como go, ajedrez y los videojuegos de estrategia más populares.El aprendizaje profundo es un tipo especializado de aprendizaje automático.Lo que se refiere a un conjunto general de modelos y técnicas que tratan vagamente de emular la manera en que elcerebro humano resuelve una amplia variedad de problemas.Es comúnmente usado para analizar el lenguaje natural, bien sea hablado y texto, así como las imágenes, audioy video para pronosticar los datos de series temporales y mucho más.El aprendizaje profundo ha tenido demasiado éxito recientemente en estas y otras áreas, por lo tantose está convirtiendo en una herramienta cada vez más popular e importante para la ciencia de los datos.El aprendizaje profundo normalmente suele requerir conjuntos de datos muy grandes etiquetados para capacitar a un modelo,es intensivo en computación y usualmente requiere un hardware de propósito especial para lograr tiempos decapacitación aceptables.Puedes construir un modelo personalizado de aprendizaje profundo desde cero o también usar modelos pre-capacitadosde los repositorios de modelos públicos.Los modelos de aprendizaje profundo son implementados usando marcos de trabajos populares como lo son TensorFlow, PyTorch,y Keras.Los marcos de trabajo de aprendizaje profundo generalmente proporcionan una API de Python API y la mayoría son compatibles con otros lenguajes deprogramación como C++ y JavaScript.Puedes descargar modelos de última generación pre-capacitados de los repositorios que son comúnmentereferidos como modelos “zoos”Los modelos zoos más populares incluyen los proveídos por TensorFlow, PyTorch, Keras y ONNX.Los modelos igualmente son publicados por grupos de investigación académica y comercial.Mientras está fuera del alcance de este video explicar en detalle cómo podríasconstruir un modelo, vamos a resumir brevemente las tareas de alto nivel usando un ejemplo.Vamos a suponer que quieres habilitar una aplicación para identificar objetos en imágenes capacitando unmodelo de aprendizaje profundo.Primero, recoges y preparas los datos que serán usados para capacitar el modelo.La preparación de los datos puede ser un proceso que consume mucho tiempo y trabajo.Para capacitar un modelo para detectar objetos en imágenes, necesitas etiquetar los datos brutos de capacitación,ejemplo, dibujando cajas limitadoras alrededor de los objetos y etiquetándolos.Después, construirás un modelo desde cero o seleccionarás un modelo existente que podría ser adecuadopara la tarea bien sea desde un recurso público o privado.Entonces entrenas el modelo con los datos que has preparado.Durante la capacitación, tu modelo aprende a partir de los datos etiquetados, cómo identificar los objetos queson representados en una imagen.Una vez que has comenzado la capacitación, analizas los resultados y repites el proceso hastaque el desempeño del modelo entrenado cumpla con tus requisitos.Cuando el modelo entrenado se comporta como se desea, lo despliegas para ponerlo a disposición de tusaplicaciones.En este video, has aprendido acerca del aprendizaje automático al igual que el aprendizaje profundo, para que sonusados y dónde encontrar los modelos de código abierto.En el siguiente video, te presentaremos el Model Asset eXchange, una colección curadade modelos de aprendizaje profundo listos para ser usados, así como también personalizables.

Sem 1.20: El Model Asset Exchange

En este video, te presentaremos el Model Asset eXchange en IBM Developer, un recursogratuito de código abierto para modelos de aprendizaje profundo.Durante este video nos referiremos a Model Asset eXchange como “MAX”En el video previo, describimos resumidamente las tareas de alto nivel que hay que completarpara capacitar a un modelo desde cero.Debido al volumen de datos, trabajo, tiempo y recursosrequeridos para completar las tareas, el tiempo de valoración puede ser bastante largo.Para reducir el tiempo de valoración, considera la posibilidad de tomar ventaja de los modelos pre-capacitados para ciertos tipos deproblemas.Estos modelos pre-capacitados pueden estar listos para ser utilizados inmediatamente o pueden tomar menos tiempo paraser capacitados.El Model Asset eXchange es un repositorio gratuito de código abiertopara microservicios de aprendizaje profundo listos para ser usados y personalizables.Estos microservicios están configurados para usar losmodelos de aprendizaje profundo de última generación, pre-capacitados o personalizables, para resolver los problemas empresariales más comunes.Estos modelos han sido revisados, probados y pueden ser rápidamente desplegados en entornos locales yen la nube.Todos los modelos en MAX están disponibles bajo las licencias permisivas de código abierto, haciendo más fácilel uso para propósito personales, empresariales y reduce el riesgo de responsabilidades legales.En MAX, puedes encontrar los modelos para una variedad de dominios, incluyendo imagen, audio, videoy análisis del lenguaje natural.Esta lista incluye una pequeña selección.En el laboratorio para este módulo, tendrás la oportunidad de explorar esos modelos.Echemos un vistazo a los componentes de un típico microservicio de modelos.Cada microservicio incluye los siguientes componentes:Un modelo de aprendizaje profundo pre-capacitado.Código que pre-procesa las entradas antes de ser analizada por el modelo y código que post-procesala salida del modelo.Una API publica estandarizada que permite a las aplicaciones acceder a la funcionalidad de los servicios.Los microservicios de modelos de MAX son construidos y distribuidos como imágenes de Docker de código abierto.Docker es una plataforma de contenedores que hace mas fácil la construcción de aplicaciones y su despliegueen un entorno de desarrollo, prueba o producción.La fuente de imágenes Docker es publicada en GitHub y puede ser descargada, personalizada según sea necesarioy utilizada en entornos personales o comerciales.Puedes desplegar y ejecutar estas imágenes en un entorno de prueba o producción usando Kubernetes,un sistema de código abierto para automatizar el despliegue, escalado y la gestión de aplicaciones en contenedores ennubes privadas, hibridas o públicas.Una plataforma popular Kubernetes a nivel empresarial es Red Hat OpenShift la cual está disponibleen IBM Cloud, Google Cloud Platform, Amazon Web Services y Microsoft Azure.Los microservicios de modelos exponen una API REST que los desarrolladores pueden usar para incorporarel aprendizaje profundo en sus aplicaciones y servicios.Debido a que los API REST pueden ser consumidos usando cualquier lenguaje de programación, fácilmente puedes integrarestos servicios en un ecosistema existente.La API expone un punto final o “endpoint” de predicción y unos o más puntos finales de metadatos.Este ejemplo muestra los puntos finales del Microservicio de Detección de Objetos.El punto final del modelo/predicción tiene una imagen como entrada y regresa como respuesta una listade objetos que fueron detectados en la imagen, junto con las coordenadas del cuadro delimitador que identificadonde el objeto detectado está ubicado.Algunos puntos finales de predicción igualmente pueden aceptar los parámetros de entrada adicionales que impactan en losresultados producidos, tales como los filtros.Este microservicio expone dos puntos finales de metadatos, /modelo/etiquetas y /modelo/metadatos.Estos puntos finales proporcionan información como los objetos que pueden ser detectados y el modelo de aprendizajeprofundo usado para derivar las respuestas dada la entrada.En la parte del laboratorio para este módulo, tendrás la oportunidad de explorar y probar estos puntos finalesutilizando un navegador web.Cada punto final acepta entradas de amigable aplicación, como una imagen en formato JPG, PNG o GIFen vez de una estructura de datos especifica del modelo.Cada punto final también genera salidas de amigable aplicación, como el JSON estandarizado, el cuales un formato poco pesado de intercambio de datos.Vamos a echar un vistazo más de cerca a lo que pasa cuando una aplicación invoca el punto finalde la predicción.En este ejemplo, un usuario ha seleccionado una imagen en una aplicación web, el punto final de la predicciónes invocado y la imagen es cargada.El microservicio prepara la imagen de entrada para el procesamiento, ejecuta el modelo de aprendizaje profundoque identifica los objetos de la imagen, genera una respuesta usando los resultados de predicción yregresa el resultado a la aplicación.La aplicación muestra los resultados dibujando cajas y etiquetas delimitadoras.En este video, hemos presentado el Model Asset eXchange, que es un repositoriogratuito y de código abierto para microservicios que pone a disposición de las aplicaciones yservicios en entornos locales, la nube y la funcionalidad de aprendizaje profundo.En el laboratorio, tendrás la oportunidad de probar un modelo de microservicio, explorar suAPI y aprender más acerca de cómo puedes aprovecharlo desde una aplicación web y una aplicación de Internetde las Cosas (IoT).

**Lectura: Laboratorio: Explorar los Conjuntos de Datos y Modelos (1 hora)**

LAB: Empezando con el Model Asset Exchange y el Data Asset Exchange

En este laboratorio explorarás el Model Asset Exchange (MAX) y el Data Asset Exchange (DAX), que son dos recursos de código abierto de Data Science resources en IBM Developer.

Al completar la parte 1 de este laboratorio (“Explorar los modelos de aprendizaje profundo”) podrás:

● Encontrar modelos de aprendizaje profundo listos para usar en el Model Asset Exchange

● Localizar recursos que te guíen en el despliegue en un entorno local o de nube

● Explorar la API de microservicio de modelo de aprendizaje profundo usando tu navegador web

● Articular cómo los desarrolladores pueden consumir esos microservicios.

Al completar la parte 2 (“Explorar los conjuntos de datos de aprendizaje profundo”) conocerás:

● Dónde encontrar los conjuntos de datos abiertos en IBM Developer

● Cómo explorar esos conjuntos de datos

Te tomará aproximadamente 30 minutos completar el laboratorio. Sólo se requiere un navegador web para completar las tareas.

Parte 1: Explorar los modelos de aprendizaje profundo

El Model Asset Exchange es un repositorio curado de los modelos de aprendizaje profundo de código abierto para una variedad de dominios, como el procesamiento de texto, imagen, audio y vídeo.

1. Abre https://developer.ibm.com/ en tu navegador web.

2. En el menú principal selecciona “Open Source at IBM” > “Model Asset eXchange”. Se muestra la página principal de MAX.

En este ejercicio de laboratorio introductorio, nos vamos a centrar en algunas características clave de MAX. Se puede obtener información más detallada en la Learning Path o Ruta de Aprendizaje, que abarca los escenarios comunes de despliegue y consumo.

3. Abre la Learning Path o Ruta de Aprendizaje (“Try the tutorial”) en una nueva ventana (o pestaña) del navegador y revisa brevemente los objetivos y módulos

4. Cierra la ventana (o la pestaña) del explorador de la Learning Path o Ruta de Aprendizaje.

5. La página principal muestra una pequeña selección de modelos. Desplázate hacia abajo y haz clic en “View all models” para revisar la lista completa.

Puedes filtrar los modelos por tipo y dominio usando el filtro de la izquierda.

6. Selecciona “Object Detection in Images” de la lista de Model Asset Technologies en el lado izquierdo.

Se deben mostrar cuatro modelos: Nucleus Segmenter, Human Pose Estimator, Image Segmenter y Object Detector.

7. Selecciona el modelo Object Detector.

En la página del modelo puedes aprender sobre la tecnología subyacente, leer sobre la investigación relacionada y explorar las opciones de despliegue y consumo.

8. La fuente del modelo está publicada en GitHub y puede ser descargada y modificada si se desea. Haz clic en “Get this model” para abrir el repositorio en un nueva pestaña (o ventana) del navegador.

El repositorio contiene todo lo que un desarrollador necesita para construir manualmente una versión personalizada del modelo de microservicio.

9. Cierra la pestaña (o ventana) del navegador GitHub

10. En la página Object Detector (https://developer.ibm.com/exchanges/models/all/max-object-detector/) navega a la sección “Deployment”. Todos los modelos MAX se distribuyen como aplicaciones en contenedores, que pueden desplegarse como un microservicio en un entorno local, en un entorno de nube híbrido o en un entorno de nube utilizando Docker o Kubernetes.

Está fuera del alcance de este laboratorio discutir los pasos del despliegue. Sin embargo, puedes explorar el modelo de microservicio sin tener que desplegar nada, lo cual discutiremos en la siguiente sección.

Explora un microservicio de modelos MAX

Para explorar un microservicio se puede acceder a una instancia de microservicio de evaluación que está alojada en IBM Cloud.

11. Desplázate a la parte superior de la página y haz clic en “Try the API”

12. La página de especificación de la API del microservicio se abre en una nueva pestaña (o ventana) del navegador.

13. Amplía la sección “model”. Se muestran los puntos finales del microservicio REST.

El microservicio de Object Detector expone tres puntos finales a los que las aplicaciones pueden acceder: un punto final de labels, un punto final de metadata, y un punto final de predict. Exploremos el punto final de labels.

14. Haz clic en “GET /model/labels”. Este punto final devuelve una lista de objetos que este modelo puede detectar.

15. Haz clic en “ Pruébalo ” y “ Ejecutar ”.

La respuesta indica que este modelo fue entrenado para identificar 80 tipos diferentes de objetos, como personas (person), bicicletas (bicycles), y autos (cars).

Puedes capacitar algunos modelos, como el Object Detector, usando tus propios datos con la ayuda del servicio Watson Machine Learning en IBM Cloud. (Aunque no lo harás en este laboratorio.)

16. Haz clic en “POST /models/predict” y “Try it out” para probar el punto final del servicio de modelos. Este punto final es lo que tus aplicaciones usarían para analizar el contenido de una imagen.

17. Busca un archivo JPG o PNG en tu máquina local, acepta el parámetro threshold por defecto y haz clic en “Execute” para invocar el punto final. El microservicio analiza la imagen y devuelve una lista de objetos identificados.

18. Revisa la respuesta. Para cada objeto detectado se devuelve una entrada en la lista de predictions,identificando la clase de objeto (por ejemplo, person), la probabilidad de que la clase de objeto identificada sea correcta (por ejemplo. 0.9879012107849121) y las coordenadas normalizadas en las que en la imagen se detectó el objeto. La probabilidad es un valor numérico entre 0 y 1. Cuanto más alto es el valor, más confianza tiene el modelo en la exactitud de su predicción.

El valor de entrada del threshold se utilizó como filtro, eliminando los objetos del resultado en los que el modelo no tiene suficiente confianza.

19. Reduzca (o incremente) el valor del threshold a partir de su valor predeterminado de 0,7 e invoca nuevamente el punto final. Los resultados devueltos pueden cambiar dependiendo del valor que hayas seleccionado.

20. Cierra la pestaña (o ventana) del navegador.

En la siguiente sección revisaremos brevemente cómo las aplicaciones pueden consumir el punto final de la predicción para analizar la entrada.

Consumir un microservicio de modelos MAX

Dado que el microservicio expone una API REST, los desarrolladores pueden implementar aplicaciones y servicios que consumen el punto final de la predicción en cualquier lenguaje de programación.

1. En la página del Object Detector (https://developer.ibm.com/exchanges/models/all/max-object-detector/) navega a la sección “Example Usage”.

2. Para facilitar a los desarrolladores el comienzo, cada modelo viene con un conjunto de ejemplos. Para el modelo del Object Detector estas muestras incluyen:

● Un ejemplo de línea de comando que utiliza curl (https://curl.haxx.se/), una popular herramienta de código abierto para descargar y subir archivos de la web.

Algunos modelos, como el Object Detector, también vienen con una aplicación de muestra que puedes probar sin tener que instalar nada.

3. En la página del Object Detector (https://developer.ibm.com/exchanges/models/all/max-object-detector/) desplázate hasta la parte superior.

4. Haz clic en “Try the web app” para ejecutar la aplicación de demostración en tu navegador web.

La aplicación web de demostración sube una imagen (o toma una foto usando la cámara web), envía una solicitud a un microservicio del Object Detector y visualiza la respuesta dibujando cajas delimitadoras alrededor de los objetos detectados y pegando una etiqueta.

Parte 2: Explore conjuntos de datos de aprendizaje profundo

Data Asset Exchange es una colección seleccionada de conjuntos de datos abiertos de IBM Research y de terceros que puede utilizar para entrenar modelos.

1. Abra https://developer.ibm.com/ en su navegador web.

5. Sube una imagen (o toma una foto con la cámara web) e inspecciona los resultados visualizados.

6. Cambia las condiciones del filtro (por ejemplo, disminuye el threshold de probabilidad) y observa cómo afectan a los resultados visualizados. Observa que la aplicación web almacena en el caché la respuesta y aplica el filtro sobre los datos almacenados en el caché. Esto se hizo por dos razones:

● Elimina la necesidad de reprocesar la imagen usando el microservicio de aprendizaje profundo (lo que reduce significativamente el tiempo de respuesta de la aplicación)

● Ilustra cómo se puede aplicar el filtrado del lado del cliente a los resultados si el punto final de la predicción no admite el filtrado según los criterios deseados.

Si estás interesado en saber más sobre cómo se implementó la aplicación, echa un vistazo el patrón de código:https://developer.ibm.com/patterns/create-a-web-app-to-interact-with-objects-detected-using-machine-learning/

Esto concluye la parte 1 de este laboratorio, donde discutimos el Model Asset Exchange.

Parte 2: Explorar los conjuntos de Aprendizaje Profundo

El Data Asset Exchange es una colección curada de conjuntos de datos abiertos de IBM Research y terceros que se puede utilizar para capacitar a los modelos.

1. Abre https://developer.ibm.com/ en tu navegador.

2. En el menú principal selecciona “Open Source at IBM” > “Data Asset eXchange”. Se muestra la página de inicio del DAX.

. En el menú principal, seleccione Código abierto en IBM > Intercambio de activos de datos . Se muestra la página de inicio de DAX.

La colección incluye los conjuntos de datos del Debater project (https://www.research.ibm.com/artificial-intelligence/project-debater/), conjuntos de datos que pueden utilizarse para capacitar a los modelos en el análisis de la disposición de los documentos, el procesamiento del lenguaje natural, el análisis de serie temporales y más.

3. Abre el conjunto de datos NOAA Weather Data (https://developer.ibm.com/exchanges/data/all/jfk-weather-data/), que contiene los datos de una estación meteorológica en el Aeropuerto John F. Kennedy en New York durante 8 años. Este conjunto de datos se utilizó para entrenar el modelo de predicción meteorológica en MAX (https://developer.ibm.com/exchanges/models/all/max-weather-forecaster/).

Puedes descargar el conjunto de datos usando el enlace “Get this dataset”. Los conjuntos de datos se almacenan en forma de archivos comprimidos, que pueden extraerse utilizando cualquier utilidad que soporte el formato tar.gz. Si no estás familiarizado con este formato de archivo, echa un vistazo a este breve tutorial de código abierto https://opensource.com/article/17/7/how-unzip-targz-file.

4. Inspecciona los metadatos del conjunto de datos.

● Este conjunto de datos se almacena como datos tabulares y formateado como un archivo de valores separados por comas (CSV), que es un formato básico de intercambio de datos muy popular.

● El conjunto de datos se publicó bajo la amigable ciencia de los datos de acuerdo a la licencia CDLA-Sharing (https://cdla.io/).

● El conjunto de datos contiene series temporales y puede utilizarse para predecir las tendencias meteorológicas.

5. La mayoría de los conjuntos de datos se complementan con Python notebooks que se pueden utilizar para explorar, pre-procesar y analizar los datos. Puedes acceder al notebook (o a los notebooks) haciendo clic en el enlace “Explore in Watson Studio” o “Try the notebook”:

Los notebooks están alojados en Watson Studio, la plataforma de la ciencia de los datos de IBM. Más adelante en este curso aprenderás más sobre Watson Studio, los notebooks y cómo manejarlos.

6. [Opcional] Si ya estás familiarizado con los notebooks y Watson Studio siéntete libre de abrir el enlace e importar el proyecto o el notebook. En el ejemplo que figura a continuación se muestran los activos del proyecto de conjuntos de datos meteorológicos, que incluyen el archivo de datos en bruto y dos notebooks.

Esto incluye la parte 2 de este laboratorio, que introduce a Data Asset Exchange.

**GitHub**

Sem 2.1: Visión General de Git/GitHub

En este video, tendrás una visión general de Git y GitHub, cuales son los entornos populares entre los desarrolladores y los científicos de datos para realizar el control de versiones de los archivos de código fuente y proyectos y la colaboración con otros. No puedes hablar de Git y GitHub sin el entendimiento básico de lo que es el control de versiones.

Un sistema de control de versiones te permite hacer un seguimiento de los cambios a tus documentos. Esto hace mucho mas fácil la recuperación de versiones antiguas de tu documento en caso de cometer un error y hace que la colaboración con otros sea mucho más fácil. Aquí tenemos un ejemplo para ilustrar cómo funciona el control de versiones. Vamos a decir que tienes una lista de compras y quieres que tus compañeros de habitación confirmen las cosas que necesitas y agregar otros artículos adicionales. Sin el control de versiones, tienes un gran desastre que limpiar antes de ir a realizar las compras. Con el control de versiones, tu sabes EXACTAMENTE lo que necesitas después de que cada uno haya contribuido con sus ideas.

Git es un software libre y de código abierto distribuido bajo la GNU General Public License. Git es un sistema de control de versiones distribuido, lo que quiere decir que los usuarios de cualquier parte del mundo pueden tener una copia de tu proyecto en su propia computadora, cuando ellos hayan hecho cambios, pueden sincronizar su versión con un servidor remoto para compartirlo contigo. Git no es el único sistema de control de versiones que existe allá fuera, pero el aspecto distribuido es una de las principales razones por las que se ha convertido en uno de los sistemas de control de versiones más común disponible. Los sistemas de control de versiones son ampliamente utilizados para cosas que implican código, pero igualmente puedes controlar imágenes, documentos y cualquier tipo de archivo. Puedes utilizar Git sin ninguna interfaz web usando tu interfaz de línea de comandos, pero GitHub es uno de los servicios de alojamiento web más populares para los repositorios de Git. Otros incluyen GitLab, BitBucket y Beanstalk. Existen algunos términos que necesitarás conocer antes de poder comenzar. El protocolo SSH es un método para acceso remoto seguro de una computadora a otra. Un repositorio contiene las carpetas de tu proyecto que están configuradas para el control de versiones. Una bifurcación es una copia de un repositorio. Una pull request es la manera en que solicitas que alguien revise y apruebe tus cambios antes de que se conviertan en la versión final. Un directorio de trabajo contiene los archivos y subdirectorios de tu computadora los cuales están asociados con un repositorio Git. Hay algunos comandos básicos de Git los cuales siempre usarás. Cuando comiences con un nuevo repositorio, solamente necesitarás crearlo una vez: bien sea localmente y luego hacer push a GitHub o clonando un repositorio ya existente utilizando el comando “git init”. “git add” mueve los cambios desde el directorio de trabajo a la staging area. “git status”, permite ver el estado de tu directorio de trabajo y la snapshot del stage de tus cambios. “git commit”, toma una captura instantánea del stage de los cambios y los confirma en el proyecto. “git reset”, deshace los cambios que hayas hecho en los archivos de tu directorio de trabajo. “git log”, permite navegar por los cambios previos de un proyecto. “git branch”, permite crear un entorno aislado dentro de tu repositorio para hacer los cambios. “git checkout”, permite ver y cambiar las branches existentes. “git merge” permite volver a unir todo nuevamente. Para aprender a usar Git efectivamente y comenzar a colaborar con los científicos de datos alrededor del mundo, necesitarás aprender los comandos esenciales. Por suerte para nosotros, GitHub tiene increíbles recursos disponibles para ayudarte a comenzar. Visita try.github.io para descargar las hojas con los distintos trucos y repasar los tutoriales. En los siguientes módulos, te daremos un curso intensivo acerca de cómo configurar tu entorno local y como comenzar un proyecto.

Lectura 2:

Taller de GitHub Parte - 1

Taller práctico: Empezando en GitHub

Esfuerzo: 20 min

En este taller, comenzará con GitHub, creando una cuenta y un proyecto GitHub y agregándole un archivo a través de su interfaz Web.

Objetivos

Después de completar este taller, podrá:

Describir GitHub

Crear una cuenta GitHub

Agregar un Proyecto / Repo

Editar / Crear un archivo

Subir un archivo y presentar

Visión general de GitHub

Primero, presentemos a GitHub. GitHub en palabras simples es una colección de carpetas y archivos. Es un servicio de alojamiento de repositorios Git, pero añade muchas de sus propias características. Mientras que Git es una herramienta de línea de comandos, un servidor necesita ser alojado y mantenido también a través de la línea de comandos, GitHub proporciona este servidor Git para ti y una interfaz gráfica basada en la web. También proporciona control de acceso y varias características de colaboración, como wikis y herramientas básicas de gestión de tareas para cada proyecto. GitHub proporciona almacenamiento en la nube para el código fuente, soporta todos los lenguajes de programación populares y agiliza el proceso de iteración. GitHub incluye un plan gratuito para desarrolladores individuales y para alojar proyectos de código abierto.

Ejercicio 1: Crear una cuenta GitHub

Por favor, siga los pasos que se indican a continuación para crear una cuenta GitHub:

Paso 1: Crear una cuenta: https://github.com/join

NOTA: Si ya tienes una cuenta en GitHub, puedes saltarte este paso y simplemente entrar en tu cuenta.

Paso 2: Indicar los detalles necesarios para crear una cuenta, como se muestra a continuación:

Y haga clic en “Create account”.

Paso 3: Hacer clic en “Verify” para verificar la cuenta y haga clic en “Done”

Paso 4: Después de la verificación, haga clic en “Join a Free Plan”

Paso 5: Seleccionar los detalles, como se muestra a continuación y haga clic en “Complete Setup”

Paso 6: Vaya a su correo electrónico, busque el correo de verificación de GitHub, y haga clic en el enlace o botón de ese correo electrónico para proceder con la verificación.

NOTA: Si no recibe el correo electrónico de verificación, haga clic en “Resend verification email.”

El correo electrónico ha sido verificado

Ejercicio 2: Añadir un Proyecto / Repo

Paso 1: Haga clic en el símbolo +, y luego clic en “New repository”.

Paso 2: Indique un nombre para el repositorio e inícielo con el archivo README.md vacío.

Y haga clic en “Create repository”.

Ahora, será redirigido al repositorio que ha creado.

Empecemos a editar el repositorio.

Ejercicio 3: Crear / Editar un archivo

Ejercicio 3a: Editar un archivo

Paso 1: Una vez creado el repositorio, la carpeta raíz aparece listada por defecto y tiene un solo archivo ReadMe.md. Haga clic en el icono del lápiz para editar el archivo.

Paso 2: Agregue texto al archivo.

Paso 3: Desplácese hacia abajo de la página después de agregar el texto y haga clic en “Commit Changes”.

Ahora, compruebe que su archivo ha sido editado con el nuevo texto.

Ejercicio 3b: Crear un nuevo archivo

Paso 1: Haga clic en el nombre del repositorio para volver a la rama master como en este testrepo.

Paso 2: Haga clic en “Add file” y seleccione “Create New file” para crear un archivo en el repositorio.

Paso 3: Proporcione el nombre del archivo y la extensión del mismo. Por ejemplo, firstpython.py y añadir las líneas.

Paso 4: Desplácese hacia abajo de la página después de agregar el texto. Agregue la descripción del archivo (opcional) y haga clic en “Commit new file”.

Paso 5: Su archivo ahora fue añadido a su repositorio y en el listado del repositorio muestra cuando fue añadido/cambiado.

Ejercicio 4: Subir un archivo y Comprometerse

Paso 1: Haga clic en “Add file” y seleccione “Upload files” para subir un archivo (Subir cualquier archivo .txt, .ipynb, .png) en el repositorio de su computadora local.

Paso 2: Haga clic en “choose your files” y elija cualquier archivo de su ordenador.

Paso 3: Una vez que el archivo termine de ser subido, haga clic en “Commit Changes”

Paso 4: Ahora, su archivo se agregó al repositorio.

Resumen

En este documento, ha aprendido a crear un nuevo repositorio, añadir un nuevo archivo, editar y subirlos a un repositorio, así como confirmar los cambios.

Autor(es)

Romeo KienzlerMalika Singla

Otro(s) Colaborador(es)

Rav Ahuja

Registro de modificación

Fecha

Versión

Modificado por

Descripción de Modificación

2020-16-07

0.5

Malika Singla

Comprobación de ortografía, número de paso añadido

2020-14-07

0.4

Rav Ahuja

Cambio de logo, se actualizó el esfuerzo, el título, la introducción, los objetivos, añadió los autores y el registro de modificación

2020-13-07

0.3

Malika Singla

Se añadió a GitLab y se hicieron algunos cambios de formato, se añadieron objetivos, etc

2020-03-07

0.2

Malika Singla

Reducido y añadido el registro de la cuenta GitHub, nuevas capturas de pantalla, etc.

2020-30-06

0.1

Romeo Kienzler

Versión inicial redactada

Â© IBM Corporation 2020. Todos los derechos reservados.

Sem 2.2: GitHub en Mac/Linux - Parte 1

¡Bienvenido a una pequeña introducción a Git y Github! Es muy sencilla por lo que si no tienes absolutamente ni idea de lo que es esto, de lo contrario, podría ser aburrido. Sólo te digo como lo estoy utilizando y soy capaz de sobrevivir en la mayorías de los casos con ese conocimiento. Algunas veces tengo que leer la documentación para cosas más avanzadas, pero en el ochenta por ciento de los casos lo que te diré ahora es suficiente. Así que comencemos a usar Github el cual hace las cosas un poco más fáciles, también hay competencia con Github. Esto es GitLab, GitLab es particularmente interesante porque todo el sistema detrás de GitLab es de código abierto, por lo que es la versión hosted, pero se puede tener todo el sistema en prem al igual que pagar por el soporte de GitLab. Estas personas están particularmente en Berlín, por lo que son muy agradables. En fin, comencemos con Github, así que lo que haré ahora es crear un repositorio. Lo cual es algo como, la ubicación donde van tus cosas. Vamos a llamar un demo, y podemos seleccionar en caso de querer crear un repositorio público o privado. Así que ese repositorio es remoto. Está alojado en Github, y está inicializado con un README y por ahora vamos a saltar .gitignore, el cual te mostraré más tarde. Así que ahora vamos a crear un repositorio y una vez que este creado es muy fácil tener una copia de este repositorio localmente. Existen dos opciones, una es clonándolo a través de HTTPS, o mediante SSH. Yo siempre uso SSH, SSH hace las cosas un poco más fáciles y lo único que necesitas hacer, es agregar tu identidad publica a Github. Puedo mostrarte eso más tarde. Ahora lo único que tenemos que hacer es ir a una carpeta vacía y luego digamos que git clone este enlace así que git lo hace. Lo que pasa ahora es que el contenido de esto es copiado a mi máquina. Aquí debo especificar mi contraseña para mi SSH identity, así que esto es algo que necesito explicarte más tarde. Así que ahora ve al demo, aquí verás el archivo README y verás que el contenido del archivo README. Bien ahora vamos a agregar el archivo, así que versa que solamente está el README y la carpeta .git. Por favor nunca jamás lo toques. Vamos a crear otro archivo vi test.txt. Y hola, esto es un test y ahora le vamos a agregar el reop. Así que primero todos nosotros, digamos git agrega test.txt, y ahora le dices git status. Entonces, podemos observar que este archivo es ahora llamado área de pruebas lo que es algo entre tu carpeta loca y tu repositorio. Aquí tienes una carpeta local con los contenidos, tienes un repositorio local y también tienes un repositorio remoto. Y personalmente trabajas con tu repositorio local. Y todas las demás personas, que trabajan contigo, lo hacen con sus repositorios locales y puedes sincronizar todos tus contenidos con el repositorio remoto. Ahora queremos ir a commit, entonces digamos, o hagamos algo más. Digamos que creamos un archivo test2 y decimos git agrega test2.txt. Ahora interesantemente, si decimos git status ambos estarán en el área de prueba. Y ahora si decimos git commit y decimos message initial commit, eso es ahora nuestro repositorio local en esta máquina. Y la idea es que mientras tengas el área de prueba, tú puedas controlar cuales archivos van a commit porque es una entidad atómica en git y es commit por lo que todo esta alrededor de commit. Por lo que técnicamente puedes controlar cuales archivos van a single commit y ahora lo que quieres hacer es push el contenido de tu repositorio local a un repositorio remoto. Porque si ahora vas a esto aquí a el proyecto git hub, entonces verás que no hay ningún archivo adicional. Así que digamos ahora git push y lo que pasa ahora es que tengo que proporcionar nuevamente una contraseña. Y no tengo el gesto de contraseñas o el agente de contraseñas funcionando, por lo que ahora ves aquí que hemos confirmado este cambio local desde el repositorio local al repositorio remoto. Ahora puedes ver que tienes estos dos archivos aquí y si haces clic en ese archivo primeramente verás el contenido y también puedes ver cuales tienes en commit el cual fue el responsable de agregar este archivo. Otra cosa interesante es, si estas comenzados desde cero también puedes hacer lo siguiente. Vamos a borrar este proyecto lo que es un poco difícil solamente para asegurarnos de que no lo haces accidentalmente, así que borra este repositorio y digamos que vamos a demo. Oh no, creo que tienes que escribir todo completamente, sí, así que este proyecto ha sido borrado. Bueno, vamos a hacer algo más. Vamos a crear un nuevo repositorio, pero vamos a iniciar el repositorio desde mi máquina. Vamos a llamarlo nuevamente demo, no lo iniciamos con un archivo README. Decimos crear un repositorio y si tenemos suerte github nos mostrará cómo hacerlo. Bien, vamos a deshacernos de la carpeta demo y creémosla nuevamente. Ahora vamos a hacer lo que nos dicen, entonces creamos el archivo README y decimos git init. Ahora que lo hemos iniciado tenemos la carpeta dot git la cual no deberías tocar nunca más. Ahora le decimos a git que agregue el archivo README, pero también le podemos decir a git que agregue dot. Así que se agrega todo lo que estaba en nuestra carpeta, que es solamente el archivo README.md y luego decimos git commit como prueba habitual o lo que sea. Ahora no podemos decir git push, si decimos git push no sabrá donde hacerlo porque en realidad no le hemos dicho a esta carpeta local que debería conectarse al repositorio remoto. Eso lo hacemos con el comando git add remote add origin y esa es la URL de nuestro proyecto. Luego podemos decir git push - u origin master y eso es prácticamente todo. Así que tenemos que recrear, disculpen, volver a introducir la contraseña, para mi clave RSA y si tenemos suerte y todo funcionó perfectamente. Podremos ir al proyecto y deberíamos ver el archivo README, el cual es el del caso. ¡Bien, eso es todo por ahora! Esa es la primera parte del siguiente video, en el cual te mostraré cómo hacer las ramas (branches) y en realidad también a cómo crear una pull request. ¡Gracias por ver este video y hasta pronto!

Sem 2.3: GitHub en Mac/Linux - Parte 2

Bienvenido de nuevo a la segunda parte de mi pequeña serie GitHub for Dummies. Una cosa que olvidé en el último video es cómo sincronizar si alguien más ha hecho un cambio. Bueno. Consideremos este depósito aquí, no tenemos nada adentro; así que este es en realidad el repositorio remoto aquí en GitHub. Entonces, crearemos nuevamente un pequeño archivo README, agregaremos "README: y luego decimos "git add" y luego decimos "git commit" "blahblah" y luego "git push". Luego tenemos el archivo README aquí, y ahora vamos a crear un nuevo archivo en el repositorio remoto, eso es algo que sucederá también si alguien más envía algo a ese repositorio. Verá un nuevo archivo... y ahora solo está en el repositorio remoto y no en mi local (repositorio ). Verá aquí, solo está el LÉAME y no el nuevo archivo. Para obtener este nuevo archivo, simplemente diga "git pull" y luego verá que el nuevo archivo también está aquí. Y sí, eso es básicamente todo. Y luego curiosamente, supongamos que está, por ejemplo, trabajando en un gran subproyecto en el repositorio y no quiere interferir con el trabajo principal que se está realizando. El trabajo principal se llama básicamente una rama, que es "maestro". crea tu propia rama Entonces puedes decir, "Git branch mybranch" por ejemplo. So, then you say "git status" we still see that we are here on the master branch. Now we say "git checkout mybranch" and now, see, we are switched to my branch. and now we say "git status" and you also see here we are on mybranch. Now what I do is... here we have those two files, and we say "# stuff on branch" here and we say "stuffonbranch.txt" so "git add stuffonbranch.txt" then "git commit" and "bla" and "git push" Ah yeah, we have to also create this branch remotely. So, and now we push it to a remote branch. Now we can actually see that mybranch has been created here and you see here the file stuff on branch which is not on the master. So, what we do now here is we switch back to the main branch. So that's "git checkout master" and then we say "git merge mybranch" So this has been automatically merged so you see here we have stuff on branch also in my local folder here. But we see here it is still not in the master, ok? So we say "git push" and if you now refresh you see stuff on branch. So what happened here is we have merged from the local branch repository into the local master repository and now we push the contents of the local master repository to the remote master repository, and that way we have merged branch back into the master branch. So, that is all for now here. In the next video, I will show you how to actually create pull requests and contribute to foreign repositories.

Sem 2.4: GitHub en Mac/Linux - Parte 3 (Opcional)

Ok, este es el último video ahora, en la serie 'Git for Dummies'.Entonces, solo quiero mostrarles cómo bifurcar un repositorio en github, cómo comprometerse con un bifurcadorepositorio y cómo crear una solicitud de extracción.Entonces, consideremos que desea contribuir al mejor y más importante repositorio de githuben el planeta.Este es el repositorio de Coursera en la organización de IBM que estoy manteniendo, ysupongamos que ha encontrado un error en uno de esos componentes de software.Quiero decir que es imposible que haya un error, pero considera que hay un error y quieresarreglarlo.Entonces, lo que realmente podemos hacer es pedirme que te otorgue derechos de confirmación para ese repositorio.pero es mejor si solo crea una bifurcación de este repositorio y luego se compromete con su repositorio,a tu bifurcación y crea una solicitud de extracción.Déjame mostrarte cómo funciona, para que veas aquí 529 otras personas ya bifurcaron este repositorio.Por cierto, dame una estrella si quieres, ¡quiero tener 1000 estrellas pronto!Aquí puede hacer clic en la bifurcación y luego le preguntará dónde desea bifurcarla.Así que esto está en la organización de IBM, y luego lo introduzco en mi cuenta privada.Eso significa que tendré un repositorio de Coursera en mi cuenta aquí, y verás aquí inclusome muestra 'bifurcado de IBM/Coursera' Ahora puedo cambiar algo, puedo hacercambia directamente aquí, pero para hacerlo un poco más complicado, hagamos que elcambia localmente.Entonces lo cloné, oh, no debería olvidar mostrarte cómo agregar tu RSAclave pública para el repositorio de github, no, sistema github.Pero lo haré más tarde, así que ahora clonando este repositorio de Coursera, este repositorio remotoen mi repositorio local.Usted ve aquí nuevamente que está en la cuenta 'romeokienzler', que es básicamente una copia de launo oficial de IBM.Creo que tenemos que esperar un poco porque es bastante grande, debería haber puesto dos pequeñosrepositorios de prueba, pero de todos modos, esperemos tanto tiempo.Mientras esperamos, déjame mostrarte algo más.Eso es sobre las claves SSH, y ahora en nuestra carpeta de inicio, y en su carpeta de inicio tieneuna carpeta SSH.Ya he creado la clave, así que es una clave privada, nunca se la dé a nadie ynunca jamás mostrarlo, el contenido a nadie.Y usted crea esos y puede establecer una frase de contraseña para las claves y pondré los comandos paracrearlos en la descripción de este video.Pero esa es la clave pública que puedes dar a cualquiera de forma segura y compartir con cualquiera.Así que también puedo hacer eso aquí, así puedo mostrarles el contenido.Mira, esa es la clave pública, y lo que hago es si voy a la página de github, tengopara ir a, creo que la configuración de la cuenta, claves SSH y GPG.Usted ve aquí, aquí tiene que agregar la clave, por lo que puede decir nueva clave SSH y simplemente pegarel contenido de la clave pública.Y ahora no estoy haciendo eso porque ya lo he agregado.¡Pero eso es todo lo que necesitas hacer en github!Y a partir de ahora puedes usar SSH porque solo tú tienes la clave privada correspondiente paraesa clave pública aquí, por lo tanto, puede autenticarse de manera segura contra github.Así que también sabes esto y espero que mientras tanto, esta otra clonación esté hecha.Así que vaya a Coursera y echemos un vistazo a ese repositorio y así tenemos el archivo Léame.por ejemplo, echemos un vistazo al archivo Léame.Tal vez hay algo que pueda mejorar en el archivo Léame.Así que este es el proyecto github para lo siguienteespecialización en curso.Así que pongamos Coursera en una capital y ahorremos.Ok, como recuerdas, necesito decir 'git add' y luego, si digo 'git status', el archivo Léamese modifica el archivo.Yo digo ok, 'git commit'.Digo mensaje ahora y ahora es importante ya que es un repositorio real, simplemente no puedoescribir un mensaje de mierda.'pequeño error tipográfico corregido' y ahora digo git, presione OK.Así que ahora el archivo Léame corregido está en mi repositorio local y en el repositorio remoto.Pero solo en mi cuenta privada, si voy a este repositorio aquí, entonces en mi cuenta privada,entonces si hago clic en Léame.En primer lugar, lo que ve aquí hace 21 segundos se ha cambiado.Verá aquí el mensaje de confirmación 'pequeño error tipográfico corregido'.Y si hace clic en el archivo Léame que ve aquí ahora, la C ahora es mayúscula.Ahora quiero que este cambio vuelva al repositorio oficial IBM/Coursera.Y ven aquí que hay un colaborador, por lo que también comprometió algunas cosas a través de una solicitud de extracción.en el repositorio principal por lo que es sencillo.Puede crear una solicitud de extracción fácilmente aquí, en ese botón y luego la diferencia entremi repositorio y el repositorio oficial solo se usan y aquí ven la única diferenciaestá aquí la pequeña corrección del error tipográfico y decir crear una solicitud de extracción.Eso es algo que estoy haciendo como colaborador potencial y digo crear una solicitud de extracción.Así que ahora yo mismo, como administrador del repositorio de IBM, recibiré un correo electrónicoy ven aquí que ya estoy reenviado (al repositorio de IBM), así que consideremos queno soy Romeo, soy otra persona el administrador del repositorio y aquí veo que hayEntra una solicitud de extracción y la revisaré Así que veo que está bien que alguien haya corregido un pequeño error tipográficoen mi repositorio eso es algo que definitivamente puedo aceptar.Así que digo fusionar solicitud de extracción, confirmar fusionar y, por lo general, lo que digo se fusiona, gracias amucho, más o menos, para ser cortés con sus colaboradores.Ahora aquí estoy en el repositorio de Coursera y si ahora hace clic en Léame verá aquí,hay una mayúscula y eso es básicamente lo que parece.Ahora aquí ves que se compromete y el mensaje de confirmación.Y eso es básicamente todo, así que espero que esto te haya ayudado un poco y con ese conocimiento puedassobrevivirá alrededor del ochenta por ciento de sus problemas y trabajo de git y github.Si tiene preguntas específicas, no dude en preguntarme en el foro de discusión o en los comentarios.debajo del video y gracias por mirar y que tengas un buen día!¡Chau!¡Muchos gracias!

**Jupyter Notebook y JupyterLab**

Sem 2.5: Comenzando con Jupyter Notebooks

Entonces, si desea comenzar a usar Jupyter Notebooks, lo más fácil que puede haceres que solo usa un servicio de Jupyter Notebook de uno de los proveedores de la nube.Así que dentro de IBM se llama IBM Watson Studio.Watson Studio es más, pero tiene un servicio de Jupyter Notebook alojado.Y la otra alternativa es simplemente usar un sistema llamado Anaconda.Y eso viene con Anaconda Navigator.Y aquí puede instalar y ejecutar diferentes herramientas de ciencia de datos.Y comenzaré aquí con Jupyter Notebooks primero, pero definitivamente cambiaré de inmediatoa JupyterLab.Solo le mostraré cómo se ve Jupyter Notebooks.Ok mira, esto ahora es Jupyter Notebook.Ok, y la clave en Jupyter Notebooks es que puede mezclar y combinar código, visualizacionesy documentaciones.Así que vamos a crear un cuaderno.Y hay varios idiomas compatibles.La mayoría de los lenguajes interactivos.E iré aquí por un cuaderno de Python 3.Pero también existen los llamados núcleos para Scala, para R y cualquier otra programación interactiva.lenguaje que puedas imaginar.Entonces, lo que puedes hacer aquí es escribir código, ok.Entonces, comencemos con una expresión de Python simple y luego, lo bueno es que puede inmediatamenteevaluar una variable.Entonces no hay necesidad de imprimir esta variable.Por supuesto que también puedes hacer eso.Pero la última variable en el código siempre se evalúa.Entonces, por ejemplo, si tiene Y es igual a 2, y luego dice, Y y X.Entonces, oh, por supuesto que primero tengo que ejecutar la celda.Así que ahora ves que es ah, 1.Y si es Y, entonces es 2.¿De acuerdo?Eso es lo básico.Así que echemos un vistazo a cómo trazar cosas para que pueda, por ejemplo, crear unformación.Simplemente crearemos una matriz numpy.Entonces X es igual a una matriz de puntos numpy, y luego creamos aquí, ah ... 1, 2, 3, 3, 3, 4, 5, 6, 7, 8,9, 1, 1, 2, 3, 4, 5.Ok, eso es ahora una matriz numpy, lo tenemos aquí.Y también verifique dos veces aquí.Si dice tipo, esto es ah, de tipo nd-array.Está bien.Tan numpy es genial, pero no entraré en detalles aquí.Así que ahora importaremos una biblioteca llamada Seaborn para trazar.Para mantenerlo bien, ponemos todas las importaciones en la parte superior.Entonces podemos decir, Seaborn distplot, que es básicamente un histograma.Y luego ven aquí un histograma pintado muy bonito.Entonces usa, como puede ver aquí, matplotlib en segundo plano.Pero Seaborn en realidad actúa como una capa encima de matplotlib.Y convierte las parcelas matplotlib de aspecto muy feo en algo muy bonito,como también estamos esperando, por ejemplo, de ggplot2, que tenemos en R. Y el últimoLo que quiero mencionar aquí es que también puede agregar otra celda y convertir esto enuna celda de documentación.Así que he seleccionado aquí Markdown.Ahora puedo en realidad... así que vamos a mover esto aquí.Di, ah, esto es un título.¿De acuerdo?Entonces puedo agregar algo, esto es interesante.Y luego, esto es aún más interesante.Usted ve aquí que incrementa automáticamente el número aquí.Así que esto es genial porque puedes mezclar y combinar la sintaxis de Markdown, tu favoritolenguaje de programación, y todas las salidas.

Eso es solo una pequeña introducción y ahora pasaré directamente a explicar JupyterLab.

* ¿Verdadero o falso? Para escribir y ejecutar código en Jupyter Notebooks, primero debe escribir código en una de las celdas de código gris y luego presionar el botón Reproducir para ejecutar el código. Las alternativas a presionar el botón Reproducir incluyen métodos abreviados de teclado como Shift+Enter y Ctrl+Enter para ejecutar el código. VERDADERO

Sem 2.6: Comenzando con Jupyter Lab

Entonces, volviendo aquí al navegador anaconda, también podemos lanzar aquí, jupyter lab.Esto ahora parece un poco más complicado, pero en realidad no lo es porque lo que puedesver aquí también hay un cuaderno jupyter.Por lo tanto, siempre puede descargar un cuaderno jupyter de este tipo e importarlo a un jupyter normalcuadernosO también puede importarlo, por ejemplo, a Watson Studio o cualquier otro jupyter alojado en la nubemedioambiente.Entonces, todos son compatibles y esto también es genial y lo que puedes ver aquí es quehay algunas ventanas más.Entonces, ven aquí una pequeña sección, así que este es un explorador de archivos, y luego la otra cosacomo ve aquí, tiene varias sesiones de kernel llamadas.Entonces, una sesión del kernel en realidad está más o menos adjunta a un cuaderno jupyter.Entonces, esto significa que se está ejecutando un cuaderno jupyter en la sesión del kernel y la sesión del kernelestá ejecutando su código real.Por lo tanto, puede ver aquí que este cuaderno jupyter está, por supuesto, renderizado en mi navegador web y eso escomunicándose con ese núcleo remoto.Y el núcleo está encapsulando el entorno de tiempo de ejecución, en este caso, Python ordinariointérprete.También puede apagarlos y reiniciarlos, aquí tiene algunas preferencias y una listade las pestañas abiertas.Entonces, las cosas que puede ahora, por ejemplo, crear una nueva terminal y ver aquí hay unaapertura de la segunda pestaña para que pueda mezclar y combinar jupyter notebooks y sus terminales.Y también puedes arrastrarlos y soltarlos.Entonces puede dividir y agregar otro terminal, por ejemplo, también puede ponerlo aquí o aquí.Así que es una buena manera de obtener una visión general de su trabajo.Entonces, cree un nuevo cuaderno de python aquí, y luego pongamos esto aquí.Bien, ahora tienes dos portátiles y dos terminales abiertos al mismo tiempo y sí, eso es,muy bueno y eso es básicamente todo.Entonces, ahora cerremos todos esos.He abierto aquí un cuaderno de mi curso de IA aplicada en Coursera y puedes verlo aquípodemos ejecutar el comando de shell.Entonces, cada vez que vea un signo de exclamación, puede ejecutar directamente los comandos de shell, por lo quehagámoslo aquí.Entonces esto se llama una 'actualización de instalación de pip' y algunos paquetes aquí, por lo que algunos de ellos sonya instalados, pero algunos se están recopilando.Y eso es una especie de mejor práctica.Siempre estoy haciendo al comienzo de cada cuaderno. Tendré una instalación de pip, queasegurarme de estar siempre en las versiones correctas de los paquetes de software dependientes.Así que ves aquí la estrella, la estrella significa que la celda actual se acaba de ejecutary una vez terminada la ejecución, verás un número.Y un número es una identificación de tarea y ese es un número incremental.Según el número, sabe en qué orden ha ejecutado las celdas.También puede ver aquí que el cuaderno jupyter se está ejecutando.Ahora está hecho, así que vemos aquí uno y lo que siempre hago también es revisar elversiones correctas instaladas.Así que ven aquí, estoy buscando tensorflow uno catorce y keras 225.Y si ese no es el caso, estoy lanzando una excepción, así que no te preocupes por las advertencias.estamos oficialmente en tensorflow 2.x.Así que necesito actualizar este portátil pronto de todos modos.En caso de que tenga una tarea de ejecución prolongada y simplemente no se detenga, siempre puede reiniciarel núcleoEntonces, hagamos esto aquí, y digo que a menudo reinicie el kernel y borre todos los resultados porque,una vez que la salida desaparece, realmente veo que el núcleo se ha reiniciado.Así que ve aquí, toda su salida se ha ido y eso significa que en realidad ha cerrado elintérprete de python y creó un nuevo intérprete de python en segundo plano aquí.Otra cosa que tal vez sea interesante, en realidad, vamos a crear un nuevo cuaderno.Entonces tienes aquí, tu salida uno.Siempre puede decir "contraer todo el código" o "contraer todo el resultado".Y vuelva a abrir una celda diferente, de esa manera puede mejorar la visibilidad del contenido relevante.Eso es básicamente todo.Así que hay más cosas para explorar, pero creo que con el conocimiento se puede sobrevivir.Y creo que he cubierto alrededor del 80% de lo importante.Tal vez una última cosa, así que exporta este cuaderno a varios formatos e intentemoslas diapositivas 'reveal.js'.Ábrelo en Firefox inmediatamente y verás que ahora es una presentación deReveal.js.

Sem 2.7: Arquitectura Jupyter

Entonces, presentemos la arquitectura de Jupyter.Entonces, en el lado izquierdo, por supuesto, tiene al usuario, que interactúa con un navegador yel componente central en Jupyter es el servidor portátil.Entonces, el servidor del cuaderno está cargando y almacenando el archivo del cuaderno jupyter y envía el htmlcontenido al navegador.Una vez que se procesa este contenido html en el navegador, interactúa con el servidor del cuaderno usandoenchufes web.Entonces, es una aplicación dividida, la interfaz de usuario se ejecuta en el navegador usando javascript.Y el servidor del portátil es responsable de mantener el estado, y lo segundo es quese está comunicando con un llamado kernel.Entonces, el kernel es un contenedor, que envuelve el marco de ejecución, por ejemplo, Pythonintérprete.Entonces, un problema que podría encontrar si está ejecutando una carga pesada en jupyter o jupyterlab todo esto está instalado en una sola máquina, ya sea en su computadora portátil o en una máquina servidor.Especialmente, por ejemplo, si está utilizando Amazon Sage Maker, tiene los cuadernos Jupyter.Pero, todo está restringido a la máquina, la instancia EC2 que ha elegido antes.Y otra desventaja es, por supuesto, si no está usando la máquina, todavía pagapara ello.Y si no es suficiente, entonces solo necesita esperar mucho tiempo o es simplemente imposiblepara ejecutar su carga de trabajo.Así que ven que aquí ilustrado, aquí tenemos una sola máquina y varios usuarios paraejemplo, están ejecutando varios núcleos y, finalmente, la memoria principal se agota ysimplemente se bloquea.Así que ¡BOOM!Y la solución a esto es la puerta de enlace empresarial jupyter, que IBM creó y que IBM tambiénopen source, que ahora también forma parte de la distribución oficial de código abierto de jupyter.Entonces, la idea aquí es que tienes un proxy y detrás del proxy hay un grupo demáquinas y ese proxy decide dónde poner los núcleos individuales y los núcleos sonluego se ejecuta en esas máquinas.Para entrar un poco más en detalle, aquí estoy usando un diagrama de arquitectura, de jupyter enterprisepuerta de entrada Así que aquí tiene, los usuarios a su izquierda, como de costumbre, hay jupyterInstancias de notebook ejecutándose en el navegador.Esa es una aplicación javascript y se comunica con el servidor usando websockets,luego tiene jupyter lab y luego, la puerta de enlace empresarial de jupyter se activa.Entonces, en contraste, jupyter lab habla con un núcleo, habla con la puerta de enlace empresarial jupyter,que actúa como proxy.Entonces, básicamente actúa como un núcleo, pero en realidad reenvía todas las llamadas a un control remoto.kernel, y el kernel remoto se ejecuta dentro de un POD de Kubernetes.Entonces, básicamente es un contenedor y ese contenedor de Kubernetes se puede programar en cualquier clústermiembro.Y lo bueno aquí es que ahora puede iniciar tantos núcleos como desee, yKubernetes se encarga de distribuirlos en el clúster.Así que, por supuesto, no quiero lanzar Watson Studio aquí, pero pueden ver aquí que enel fondo, mientras este portátil se ha cargado, el kernel se ha iniciado en un control remotoMiembro del clúster de Kubernetes.Y todo lo que ejecute aquí ahora pasa por la puerta de enlace empresarial de jupyter.Solo estoy mostrando Watson Studio aquí porque no tengo una instalación de código abierto disponibleque utiliza la puerta de enlace empresarial jupyter.Y si ahora, por ejemplo, revisa esta descripción general del proyecto, entonces Watson Studio preguntala puerta de enlace empresarial jupyter sobre el estado de los diferentes contenedores, que están encapsulandolos granosY ven aquí, tengo un entorno activo.Este es este aquí, ves aquí la configuración de hardware y esto realmente se está ejecutandoen algún lugar de la nube de contenedores de IBM en Kubernetes.Eso es genial y solo una última cosa, quiero mostrarte aquí si puedes especificarentornos y puede adjuntar entornos a cuadernos.Básicamente, puede elegir cuántas CPU y cuánta RAM ejecuta un portátil en particular.Eso es particularmente útil, por ejemplo, si tiene una tarea ETL para la integración de datos queusa chispa, solo usa un cuaderno chispa.Y si es una tarea pesada puedes aumentar el número de ejecutores.O tiene, por ejemplo, un cuaderno de tensorflow que luego toma estos datos y hace algo de entrenamiento.Luego, por supuesto, también puede aumentar la cantidad de CPU y la memoria principal.Eso es básicamente todo y espero que hayas aprendido algo sobre jupyter, jupyter notebooksarquitectura y si tiene preguntas, publíquelas en la sección de comentarios oUsa el foro de discusión de Coursera.¡Muchas gracias, adios!

Laboratorio: Jupyter Notebooks – Principios

Ejercicio 1 – Crear un nuevo notebook en Python

1. Si ya tiene Skills Network Labs abierto en una pestaña, puede hacer clic en JupyterLab en la página principal.

2. Para crear un nuevo notebook, haga clic en cualquiera de los idiomas en "Notebook".

Ejercicio 2 - Escribir y ejecutar el código

En su nuevo notebook vacío (del Ejercicio 1), haga clic dentro de la celda de código gris y escriba algún código, como "1 + 1" (sin comillas).

Ejecute el código, ya sea haciendo clic en el botón Play en el menú arriba del notebook, o presionando Shift+Enter en su notebook.

Debería ver en la salida, "2".

Intente ejecutar otro código (intente operaciones matemáticas simples).

¡Grandioso! Ahora sabes cómo escribir y ejecutar código en Jupyter Notebooks

Ejercicio 3 - Crear nuevas celdas

En su Jupyter notebook, primero haga clic encualquiera de las celdas existentes para seleccionar la celda.

En el menú, haga clic en "Insert", luego "Insert Cell Above" o "Insert Cell Below".

¡Grandioso! Ahora ya sabe cómo insertar nuevas celdas en Jupyter Notebooks. Tenga en cuenta que puede usar atajos del teclado como: [a] - Insertar una Celda arriba; [b] – Insertar una celda abajo.

Ejercicio 4 - Crear celdas Markdown y añadir texto

En su notebook, haga clic en cualquier celda de código, y en el menú desplegable de la parte superior, cambie el tipo de celda de "Código" a "Markdown". Como notará, no puede crear celdas "Markdown" sin crear primero celdas de "Código" y convertirlas a "Markdown".

En la celda de Markdown, escriba un texto como "Mi título".

Para renderizar el texto de la celda markdown, asegúrese de que la celda esté seleccionada (haciendo clic dentro de ella) y pulse Play en el menú o Shift+Enter.

¡Su celda Markdown debe estar ahora renderizada!

Para editar la celda de Markdown, haga doble clic en cualquier lugar de la celda. Tenga en cuenta que puede usar el atajo de teclado: [m] - Convertir la celda en Markdown

Este curso utiliza una herramientas de terceros, Laboratorio: Jupyter Notebooks – Principios para mejorar tu experiencia de aprendizaje. La herramienta hará referencia a información básica, como tu nombre, correo electrónico e ID de Coursera.

Lectura: Jupyter Notebook en Internet

Hay miles de interesantes Jupyter Notebooks disponibles en internet para que aprendas. Una de las mejores fuentes es: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

Es importante notar que puedes descargar tales Notebooks a tu computadora local o importarlos a una herramienta de notebook basada en la nube para que puedas volver a ejecutar, modificar y seguir lo que se explica en el notebook.

Con frecuencia Jupyter Notebook ya está compartido en una vista renderizada. Esto significa que puedes verlo como si estuviera funcionando localmente en tu máquina. Pero a veces, la gente sólo comparte un enlace al archivo de Jupyter (que se puede distinguir por la extensión \*.ipynb). En este caso puedes simplemente tomar la URL de ese archivo y pegarla en el NB-Viewer => https://nbviewer.jupyter.org/

La lista anterior te da un muy buen comienzo con una enorme colección de materiales para explorar. Por lo tanto, es tal vez más útil para darte algunos consejos interesantes sobre notebooks. Como hemos cubierto algunos ejemplos con los datos de juguetes (toy data) en los laboratorios, permíteme señalar algunos trabajos que utilizan estos datos y que van más allá en el camino de la ciencia de los datos. Además, como hemos cubierto las diferentes tareas de la ciencia de los datos, también proporcionaremos un notebook de ejemplo para cada uno de ellas.

Primero se comienza con el análisis exploratorio de datos, por lo que este notebook es muy recomendable: https://nbviewer.jupyter.org/github/Tanu-N-Prabhu/Python/blob/master/Exploratory\_data\_Analysis.ipynb

Para la integración de datos/limpieza a menor escala, a menudo se utiliza la librería de Python “pandas”. Por favor, echa un vistazo a este notebook: https://towardsdatascience.com/data-cleaning-with-python-using-pandas-library-c6f4a68ea8eb

Si quieres experimentar ya lo que es la agrupación, echa un vistazo a esto:https://nbviewer.jupyter.org/github/temporaer/tutorial\_ml\_gkbionics/blob/master/2%20-%20KMeans.ipynb

Y finalmente, si quieres ir por un notebook más profundo sobre el conjunto de datos de “iris”, mira aquí: https://www.kaggle.com/lalitharajesh/iris-dataset-exploratory-data-analysis

**IDE de RStudio**

Sem 2.8: Qué es RStudio IDE?

Así que ahora echemos un vistazo a RStudio.RStudio es IDE, entorno de desarrollo integrado,y está hecho para el lenguaje de programación R. R esun lenguaje de programación estadística que ha sidoderivado del lenguaje S de código cerrado.Entonces R es de código abierto,RStudio es de código abierto,y es completamente gratis.La estructura de datos central en R es un DataFrame.Entonces tenemos aquí una ventana que contieneun editor y debajo nosotrostenemos aquí una ventana que contiene la consola.El intérprete de R es el intérprete interactivo.Entonces tiene acceso al intérprete aquí en cualquier momento,pero es útil si tiene un editor de texto aquí.Siempre puedes ejecutarel código que estás escribiendo aquí en este editor de texto.En la parte superior derecha,tienes una ventana estrechadonde se muestra su entorno,por lo que cada variable enel montón es accesible y también puede inspeccionarlo.Si haces gráficas,están acabando aquí.Entonces, comencemos con un DataFrame.Así que para saber dónde estamos,decimos, "Getwd", así que obtenga el directorio de trabajo.Entonces estamos en la carpeta "Inicio". Eso es muy bonito.Así que ahora decimos, DataFrame es igual.Es posible que haya visto este símbolo aquí, por lo que es un legado.Solo hay una ligera diferencia en los códigos de función,así que no te preocupes por esoa menos que lo esté usando en códigos de función.Entonces puede usar el operador de asignación ordinario normal,lo que hace que nuestro aspecto sea un poco menos feo.Entonces decimos ahora read.csv y ahora podemos decir,"Ir a mi carpeta de Descargas".Aquí vemos todos los archivos posibles.Entonces sé que comenzó con CU,por lo que es mensajes de los clientes.Le damos a Enter, podemos ejecutar esta línea.Entonces esta línea se copiaa la consola y se ejecuta.Al mismo tiempo, ves aquí DataFrame esbásicamente contiene los datos de ese CSV,y ahora podemos echar un vistazo ael contenido de este DataFrame.Si decimos, "Ver (df)", veráaquí ya se muestra dentro de la ventana del editor.Lo que notamos es quela primera línea se ha interpretado como encabezado.Eso es algo que no queremos.Entonces podemos decir, "Encabezado FALSO".Es muy útil que tenga autocompletado aquí comosiempre sabes qué parámetros tiene una función.Si ahora vuelve a echar un vistazo a DataFrame,ves aquí la primera línea esahora parte de los datos y no del encabezado.Así es como puedes cargar datos.Puedes, por supuesto,también cargar datos de sistemas de bases de datos remotos.Puedes usar ODBC para eso,pero eso está más allá del alcance de este tutorial.Lo que les quiero mostrar es otra forma de importar datos,puede decir aquí, "Importar desde CSV"y luego el código se crea realmente para usted.Así que abramos el archivo de nuevo.Básicamente está haciendo lo mismo,solo que tienes una idea.Tienes una vista previa de datos con el mismo problema.que tiene la primera línea interpretada como encabezado.Entonces puedes desmarcareste aquí y luego es como debe ser,y usted dice, "Importar".Notas que aquí tienes un segundo objeto.aquí en el montón que se llama mensajes de clientes.También es DataFrame y también se muestra aquí.

Eso concluye el primer video.En el siguiente vídeo,Te mostraré cómo trabajar con bibliotecas,y después de eso te mostraré cómo crear tramas.

Sem 2.9: Instalación de paquetes y carga de Librerías en RStudio IDE

Una de las principales fortalezas de R no es una sintaxis muy agradable, por supuesto, es laabundancia de paquetes.Entonces, si va a cran.r-project.org, esta es The Comprehensive R Archive Network.Tienes un enlace aquí para paquetes y ves aquí la tabla de paquetes disponibles.También puede ordenar por fecha o nombre, por lo que obtiene alrededor de 15,349 paquetes.Y eso es mucho, así que soy el tipo de audio.Así que echemos un vistazo aquí para ver el audio.Así que hay un paquete de audio.Si desea instalar un paquete de este tipo en R, es bastante sencillo.Lo que debe hacer es decir 'install.packages' y luego decir el nombre del paquete,'audio' y luego simplemente ejecutas.Me pregunta si quiero compilarlo desde la fuente, está bien.Entonces, muchos de los paquetes R están escritos en C o C ++, por lo que debe tener instalado un compilador C,y luego el paquete se compila en su máquina por usted.Así que esto está hecho, para usar el paquete necesitas cargarlo.Di biblioteca o también puedes decir require, no importa y luego dices audio.Así que ahora el paquete está cargado y ves que también aquí, que el paquete de audio ahora está cargado.Y ahora puedes usarlo.Entonces, si desea echar un vistazo al código fuente o algunos ejemplos, puede explorarabajo del archivo R.Así que voy aquí a esta URL, y ven aquí, es un pequeño ejemplo.Entonces, esto debería reproducir una onda sinusoidal, echemos un vistazo a si puedo conectarme a mi micrófonocon mi altavoz.Si eso funciona, ¿puedes oír eso?Ok, ese es un ejemplo de lo fácil que es instalar y usar paquetes en R. Y el problemano es que no encuentre un paquete para una tarea específica, el problema es más que ustedtener cinco o diez paquetes diferentes para una tarea específica.Y necesitas leer mucho y decidir cuál estás usando.Entonces, en el próximo video, mostraré cómo trazar en R studio usando ggplot.

Sem 2.10: Graficando dentro de RStudio IDE

Entonces, antes de que podamos trazar algo que solo necesita tener un poco de informacióneso es muy fácil de crear.Esa es una función rnorm que toma muestras de una distribución normal.Entonces, por defecto, tenemos media cero y desviación estándar uno, así que muestreemos cien puntos aquí.Y hagamos lo mismo con y.y de hecho aumentar un poco la desviación estándar.Y ahora creamos un marco de datos y es data.frame y dices solo (x, y).Si ahora ve este marco de datos que vemos aquíque tienes las dos columnasY ven aquí los números son un poco más altosporque la desviación estándar es mayor.Y ahora podemos trazarlo.Entonces tenemos que importar una biblioteca (ggplot2).Y luego podemos trazarlo así que decimos ggplotmás parámetros, por supuesto, el marco de datosY no entraré en detalles aquí.porque ggplot es relativamente complejopero una biblioteca muy, muy poderosa.Básicamente estamos diciendo aquí que queremos dibujar puntos.Así que ven aquí que en el eje x y el eje ysolo estamos trazando esos puntos.Bien, ahora vamos a trazar algo mejor.Entonces, hay un marco de datos integrado en R que se llama mtcars.Y tiene diferentes tipos de coches, y sus propiedades.Así que es un buen conjunto de datos para jugar.Y como dije ggplot2 es relativamente complejoasí que solo estoy copiando y pegando el códigopor trazar algo muy bien.Entonces, esto es básicamente el peso del automóvil sobre el kilometraje.Y ven aquí cuanto más pesado se pone el carro,menos es el kilometraje.Y además de eso también estás dibujandouna línea de regresión y también un intervalo de confianza o una desviación estándar.Eso es genial con solo dos líneas de código.Puedes crear parcelas que encuentresen las mejores publicaciones (científicas) porque las mejores editorialesa menudo usa R y ggplot2 para crear ese gráfico.Bien, espero que esto ayude.Si tiene alguna otra pregunta,por favor hágamelo saber en la sección de Comentarios a continuación.Y ten un buen día. ¡Chau!

**Laboratorio: Rstudio – Principios**

Este laboratorio te presenta a R y RStudio. Primero tenemos que instalarlo.

Por favor, instala R apropiado para tu sistema operativo. Puedes obtenerlo **aquí:https://cran.rstudio.com/**

Una vez que hayas instalado R, por favor descarga e instala el RStudio apropiado para tu sistema operativo. Puedes encontrarlo aquí:**https://rstudio.com/products/rstudio/download/#download**

1. Inicia RStudio. Deberías ver algo como esto:

2. Ahora haz clic en el pequeño símbolo “plus” arriba a la izquierda y selecciona “R Script” . Esto te da lo siguiente:

3. Ahora cargamos el conjunto de datos de iris que ya deberías conocer de los laboratorios anteriores. Por favor, introduce las siguientes líneas en la ventana del editor que acaba de aparecer. Luego selecciónalas todas de tal manera que se vuelvan azules. Luego haz clic en el pequeño icono de ejecución justo encima de la ventana del editor.

biblioteca (conjuntos de datos)

datos (iris)

Ver (iris)

Luego, selecciónelos todos de modo que se vuelvan azules. Luego haga clic en el pequeño icono de ejecución que se encuentra justo encima de la ventana del editor.

4. Se le dirige directamente a la pestaña de vista de datos para inspeccionar su conjunto de datos:

5. Podemos ver que hay cinco columnas en este conjunto de datos y que las primeras cuatro son de punto flotante y la última es la etiqueta de tipo cadena de datos que contiene el valor de la categoría de nuestro conjunto de datos. También vemos que tenemos 150 entradas en total de las cuales estamos viendo las primeras 19. Ahora queremos saber cuántas especies diferentes hay presentes en el conjunto de datos. Por lo tanto, por favor escribe el siguiente comando en la ventana del editor y haz clic en el icono de ejecución:

**único(iris$Especies)**

y haga clic en el icono de ejecución:

6. En la ventana de Console, en la parte inferior, verás el resultado del comando ejecutado y sabrás que sólo hay tres especies diferentes presentes en el conjunto de datos. Ahora es el momento de mirar el conjunto de datos con más detalle. Crearemos un gráfico similar como ya hemos hecho en el laboratorio de Jupyter. Para ello, primero tenemos que instalar el paquete ggplot2. Por favor, escribe y ejecuta el siguiente comando:

**install.packages(c("GGally", "ggplot2"))**

7. Ahora que has instalado las librerías necesarias para crear algunos buenos gráficos, vamos a ejecutar los siguiente comando:

biblioteca (GGally)

**ggpairs(iris, mapeo=ggplot2::aes(color = Especie))**

8. Ahora verás el siguiente gráfico en la ventana “plots”:

9. Esto nos da mucha información para una sola línea de código. Primero, vemos las distribuciones de datos por columna y especies en la diagonal. Luego vemos todos los diagramas de dispersión por pares en los recuadros a la izquierda de la diagonal, de nuevo desglosados por el color. Por ejemplo, es obvio ver que se puede dibujar una línea para separar la “setosa” de la “versicolor” y la “virginica”. En cursos posteriores, por supuesto enseñaremos cómo las especies superpuestas pueden ser separadas también. Esto se llama aprendizaje automático supervisado usando clasificadores no lineales, por cierto. Luego se ve la correlación entre las columnas individuales de los recuadros hasta la diagonal, lo que confirma nuestro pensamiento de que el “setose” es más diferente, y por lo tanto más fácil de distinguir, que el “versicolor” y la “virginica”,ya que es un valor de correlación cercano a uno significa una gran similitud, mientras que un valor cercano a cero significa una menor similitud. Los gráficos restantes de la derecha se llaman “box-plots” y las de la parte inferior se llaman “histograms” pero no entraremos en detalles aquí y lo guardaremos para un curso más avanzado de esta serie.

Esto concluye este laboratorio, ¡espero que lo hayas disfrutado!

**Laboratorio: Creación de un mapa interactivo en R**

Bienvenidos al 2o laboratorio RStudio en el que creamos un mapa interactivo usando dos librerías. “Shiny”, que es una asombrosa librería para crear prototipos de interfaces de usuarios interactivas en R y “leaflet” que nos permite usar los datos de “OpenStreetMap” para crear un mapa interactivo. “Shiny” usará “leaflet” como uno de sus componentes de UI (como cualquier otro componente de UI como botón, deslizador o campo de texto).

Instala los siguiente paquetes, puedes hacerlo simplemente pegando y ejecutando el siguiente comando en la ventana inferior izquierda de “Console”:install.packages("shiny")install.packages("leaflet")

Pasa el siguiente código en una nueva ventana de edición de R en RStudio:library(shiny)library(leaflet)

r\_colores <- rgb(t(col2rgb(colores()) / 255))

nombres(r\_colores) <- colores()

ui <- página fluida(

salida de folleto("mimapa"),

pag(),

actionButton("recalcular", "Nuevos puntos")

)

servidor <- función (entrada, salida, sesión) {

puntos <- eventReactive(input$recalc, {

puntos = cbind(rnorma(40) \* 2 + 13, rnorma(40) + 48)

devolución (puntos)

}, ignorar NULL = FALSO)

observeEvent(input$Map\_shape\_click, { # actualice la ubicación selectInput en los clics del mapa

salida$coordenadas <- renderText({

"Has seleccionado esto"

})

})

salida$mimapa <- renderHoja({

folleto() %>%

addProviderTiles(proveedores$Estambre.TonerLite,

opciones = ProviderTileOptions(noWrap = TRUE)

) %>%

addMarkers(datos = puntos())

})

}

aplicación brillante (interfaz de usuario, servidor

​ 3. Haz clic en el botón “Run App”

verás que aparece la siguiente aplicación:

4​. Si hace clic en “New points”, se generará y se mostrará un nuevo conjunto de puntos aleatorios. Por favor encuentra el código siguiente para entender la creación de esos puntos (por favor, ten en cuenta que “rnorm” es un generador de números aleatorios y “cbind” adjunta dos listas verticalmente para formar una estructura de tabla):

puntos = cbind(rnorma(40) \* 2 + 13, rnorma(40) + 48)

5. Ahora queremos cambiar el código para mostrar la lista de puntos de la interfaz de usuario. En primer lugar, tenemos que añadir este componente en la propia interfaz de usuario, lo que significa que tenemos que añadir un elemento al objeto “ui”. Por favor, hágalo añadiendo las siguientes dos líneas a la “ui”:

pag(),

textOutput("coordenadas") tal que la sección de código completa se vea así: ui <- fluidPage(

salida de folleto("mimapa"),

pag(),

actionButton("recalcular", "Nuevos puntos"),

pag(),

salida de texto("coordenadas")

​6. Entonces, es hora de cambiar también el comportamiento de esta aplicación cambiando el objeto “server”. Por lo tanto, por favor agregue las siguientes tres líneas al llamado “eventhandler” responsable de ejecutar el código si se hace clic en el botón: output$coordinates <- renderText({

puntos

}) de modo que la sección de código completa se vea así: puntos <- eventReactive(input$recalc, {

puntos = cbind(rnorma(40) \* 2 + 13, rnorma(40) + 48)

salida$coordenadas <- renderText({

puntos

})

devolución (puntos)

}, ignorar NULL = FALSO

​ 7. Si ahora vuelves a ejecutar esta aplicación, notarás una lista de pares de coordenadas que se actualiza cada vez que haces clic en el botón “New points”

Con esto concluye este laboratorio, ¡esperamos que te hayas divertido! Si quieres saber más sobre Shiny y Leaflet, por favor visita los siguientes enlaces:

https://shiny.rstudio.com/

https://rstudio.github.io/leaflet/

**estudio Watson**

Sem 3.1: Qué es IBM Watson Studio

Todas las empresas quieren trabajar de manera más inteligente y, para hacerlo, debe aprovechar suel mayor recurso de la empresa, sus datos. Pero extraer todo el valor desus datos no siempre es un proceso fácil. Primero. terminas haciendo malabarismos increíblementecolección grande y compleja de herramientas que se utilizan para encontrar y limpiardatos, analizando y generando visualizaciones de esos datos, y usandolos datos para construir e implementar modelos de aprendizaje automático. Y para hacer las cosaspeor aún, estas herramientas son a menudo una pérdida de tiempo para administrar individualmente, y pueden serdifícil de integrar en su sistema, lo que realmente puede ralentizar el flujo de trabajo.Pero ya no más. Con Watson Studio puede simplificarsus proyectos de datos con un proceso optimizado, que le permite extraer valory conocimientos de sus datos para ayudar a que su empresa sea más inteligente y más rápida. entregaun entorno de aprendizaje automático y ciencia de datos colaborativo fácil de usarpara construir y entrenar modelos, preparar y analizar datos y compartirconocimientos, todo en un solo lugar. Watson Studios fácil de crearlas visualizaciones y el código de arrastrar y soltar ponen el poder de la base de datostoma de decisiones en manos de cualquier miembro de su organización sin necesidadpara asistencia informática. Y si necesita acceso a herramientas de código abierto, el entornoofrece algunos de los más populares y potentes disponibles. estudio watsonentorno único también crea un flujo de trabajo que es increíblemente eficiente para quelos científicos de datos pueden compartir activos y trabajar para resolver problemas dentro del sistemaen lugar de empezar de cero cada vez que surge un nuevo problema. y desarrolladorespuede usar esa eficiencia para sumergirse rápidamente en la creación de aprendizaje automático y profundoalgoritmos de aprendizaje. De hecho, en el área del aprendizaje profundo,Watson Studio admite algunos de los marcos más populares y puede implementar esoaprendizaje profundo en las últimas GPU para ayudar a acelerar el modelado haciéndolomás fácil de usar. El modelador de redes neuronales incorporado de entornos también lo ayudaconstruya modelos con una interfaz gráfica simplificada incluso si no tiene larecursos dedicados para construir un modelo desde cero, Watson's Studio puede ayudarcomienza a modelar plantillas para áreas como el reconocimiento visual,clasificación de idiomas y otras herramientas de los servicios de IBM Watson.Porque Watson Studio está perfectamente integrado con IBM Watson KnowledgeCatalog, una herramienta inteligente de descubrimiento de activos, puede transformar datos y modelosen recursos empresariales confiables y colabore con confianza, sincomprometer el cumplimiento, la seguridad o el control de acceso. Watson Studio proporcionamuchos beneficios para las organizaciones que ayudan a infundir IA en el negocio e impulsarinnovación. Puede entrenar a Watson Studio con servicios de IA incorporados, incluidosReconocimiento visual watson. Puede personalizar sus modelos e implementarlos comoAPI o Core ML mediante el uso de herramientas de código abierto como Jupyter, Notebook, Anaconday RStudio. Watson Studio es compatible con las bibliotecas de código más populares, así como conmodelado visual sin código con modelador de redes neuronales para diseñar redes neuronalesarquitecturas que utilizan los marcos de aprendizaje profundo más populares. en watsonStudio puede descubrir, limpiar y transformar sus datos de forma interactiva utilizandorefinería de datos Le ayuda a comprender la calidad y distribución de sudatos con gráficos y estadísticas incorporados, y proporciona resultados visualizados a través detableros interactivos. Watson Studio incluye una función intuitiva de arrastrar y soltarinterfaz que permite a un no programador acelerar el proceso de construcción de botellasseleccionando, configurando, diseñando y codificando automáticamente neuronalredes Desde el desarrollo y la formación hasta la producción y la evaluación, WatsonStudio realiza un seguimiento de sus modelos a lo largo del tiempo para garantizar que tenga el mejor rendimiento paracualquier tarea dada utilizando las mejores soluciones a lo largo de todo el ciclo de vida de sumodelos de aprendizaje automático.

Sem 3.2: Introducción a Watson Studio

Watson Studio es una plataforma integrada de herramientas, servicios y datos queayuda a las empresas a acelerar su cambio para convertirse en organizaciones basadas en datos que ustedpuede comenzar con una cuenta gratuita para explorar sus capacidades la ciencia de datos es un equipodeporte tenemos diferentes tipos de personas interesadas en los conocimientos que los datosla ciencia puede proporcionar esto incluye analistas de negocios ingenieros de datos datoslos administradores de datos, científicos y desarrolladores, los datos deben ubicarse en cleanselos modelos tienen que ser creados probados monitoreados y actualizados todo esto requieretrabajo en equipo por esta razon se construyo la ciudad de watson comouna plataforma colaborativa una comunidad de personas con ideas afines hay mucho paracubierta en esta introducción y solo arañará la superficie, puede encontrar másinformación sobre el sitio de participación técnica digital en ibm.com slash demosuna vez que haya iniciado sesión, puede ver la pantalla de bienvenida de inicio que puedeminimizar la pantalla haciendo clic en el botón de empezar en la esquina superior derechaun elemento importante que es fácil pasar por alto es el botón de hamburguesa en la parte superiorizquierda te da acceso directo a catálogos de proyectos y servicios entreotras cosas la galería es particularmente interesante es una colección de activosincluyendo tutoriales, cuadernos, conjuntos de datos, artículos y documentos de múltiplesfuentes se agregan constantemente nuevos activos se pueden buscar activos usando filtros paraLos temas de tecnología del idioma tailandés, etc., los resultados se pueden ordenar por característicaso por fecha administrar le brinda acceso rápido a áreas específicas para administrar finalmente nosotrostener soporte y documentación integrados en el entorno de Watsoncomo se mencionó anteriormente, el proyecto es el centro de la colaboración, es muysimple para crear un proyecto, haga clic en crear un proyecto en la pantalla de bienvenidao nuevo proyecto y crear un proyecto vacío ouno de uno existente, luego le da un nombre, posiblemente agregue una descripción yya está listo para ir a nivel de proyecto también tenemos un menú de opciones que comienzacon el resumen donde se puede ver información básica sobre el proyecto esteLa pestaña también incluye una sección Léame donde puede obtener más detalles sobre lo queel proyecto trata sobre el siguiente son los activos donde puede ver los modelos de activos de datoscuadernos y otros activos que forman parte del proyecto que puede ir a agregaractivos específicos usando el menú desplegable Agregar al proyecto en la parte superior de la pantallano entraremos en todos esos elementos del menú, pero una cosa importante que quiero saber esconexión esto le permite acceder a los datos que provienen de fuera de Watsonstudio, como puede ver, incluye una gran cantidad de servicios de datos de IBM, pero tambiénalgunos de terceros como Amazon y Microsoft volviendo a nuestro proyectoMe gustaría señalar la sección de medio ambiente como una herramienta importante para eso.la manipulación de datos de exploración y la creación de modelos es el cuaderno dependiendo dela cantidad de trabajo que hay que hacer tenemos una opción de asignación de recursostambién podemos adaptar el entorno para incluir bibliotecas adicionales para que tengamosun entorno completo desde el principio quiero señalar dos selecciones másdesde el menú superior control de acceso y configuración el control de acceso te permitepara controlar a los colaboradores y sus permisos y más en la configuraciónsección puede, entre otras cosas, agregar servicios, por ejemplo, haga clic en elmenú desplegable agregar servicio seleccione Watson y agregue un servicio de aprendizaje automático quetiene la opción de agregar un servicio existente que puede haber creado anteriormente enotro proyecto o cree uno nuevo tenga en cuenta que la mayoría de los servicios incluyen una luz gratuitaversión esto significa que puedes experimentar con todo tipo decapacidades gratisusted.

Sem 3.3: Creación de una Cuenta en IBM Watson Studio

Este video le muestra cómo probar IBM Watson de forma gratuita.IBM Watson le da acceso a IBM Watson Studio, IBM Watson KnowledgeCatálogo, la refinería de datos, aprendizaje automático y profundo, modelos de reconocimiento visual,cuadros de mando y flujos de secuencias. En https://dataplatform.cloud.ibm.com/puede registrarse para una prueba gratuita. Cuando se registra para obtener una cuenta de IBM Watsonse registra automáticamente para obtener una cuenta gratuita de IBM Cloud. Aquí ves elAplicaciones de Watson que se aprovisionarán. IBM Watson Studio e IBMCatálogo de conocimientos de Watson. Si ya tiene una cuenta en la nube de IBMluego use su ID de IBM para registrarse en IBM Watson.De lo contrario, escriba su dirección de correo electrónico que se utilizará para crear una nube de IBMcuenta para ti. En la siguiente pantalla se le redirige ala página de registro de la nube de IBM donde debe proporcionar la información básica típicainformación para una cuenta y luego haga clic en crear cuenta.Ahora revisa tu correo electrónico y confirma tu cuenta.Ahora que se registró en la nube de IBM, puede usar esas mismas credenciales pararegistrarse. A continuación, verá que el proceso crea elcuenta de usuario de IBM Watson usando sus credenciales de nube de IBM y finalmente veráque su cuenta fue creada con éxito.Esta cuenta de IBM tiene solo una cuenta de nube de IBM asociada y un grupo de recursos.Si tiene más de una cuenta asociada o una cuenta con múltiplesgrupos de recursos, luego, durante la configuración de la cuenta de Watson Studio, verá estopantalla que le da la opción de seleccionar una cuenta y un grupo de recursos para usar.IBM Cloud utiliza grupos de recursos como una forma de organizar los recursos de su cuentaen agrupaciones personalizables para que pueda asignar rápidamente a los usuarios acceso a másmás de un recurso a la vez. Ver la configuración para verificar laaplicaciones y servicios que se aprovisionan.Ahora está listo para comenzar a trabajar en IBM Watson.

Sem 3.4: Cuadernos Jupyter en Watson Studio - Parte 1

[Música]este video cubre los conceptos básicos para trabajarcon cuadernos de Júpiter en el estudio de Watsoncomience en un proyecto de estudio de Watson y agregueal proyecto un cuadernosolo proporcione un nombre en una descripcióny crear el cuadernoprimero carguemos un archivo para que tenga algunosdatos con los que trabajar desde los archivospanel deslizable navegue para seleccionar eldespués de agregar el archivo a laproyecto está disponible para trabajar eneste cuaderno simplemente haga clic en insertar para codificare inserte el marco de datos de un panda para ejecutar el portátil es lo mejorpráctica para insertar una celda en la parte superior paradescribir lo que cambia el cuadernoel tipo de celda para rebajar, por lo que esta celdano se tratará como código y luego se agregarála descripciónahora está listo para ejecutar el cuaderno elEl código insertado carga el conjunto de datos en unmarco de datos utilizando sus credenciales parasu instancia de almacenamiento de objetos en la nube yluego muestra las primeras cinco filas delconjunto de datos antes de volver al proyectoguardar el cuaderno. Soy la pestaña de activos, encontrarás elcuaderno si abres el cuadernoestará en modo de sólo lectura, pero puedeeditar el cuaderno y hacer cambiospor ejemplo puedes acceder a la infopanel y cambiar el nombre delcuaderno y en la pestaña de medio ambientepodría cambiar el entorno utilizado para ejecutarel cuadernoasí como detener o reiniciar el tiempo de ejecuciónmedioambientesi desea compartir un archivo de solo lecturaversión del cuaderno puedes hacer esodesde aquí puede seleccionar la cantidad de lacontenido que le gustaría compartir y cómoquiero compartir el cuaderno tampocoa través de un enlace o redes socialessi desea programar el cuadernopara ejecutar en un momento diferente puedecrear un trabajosimplemente proporcione un nombre para el trabajo yseleccione las opciones de programación comoespecificar una fecha para que se ejecute el trabajoy si desea que el trabajo se ejecuterepetirdespués de crear y ejecutar el trabajopuedes ver el estado en la pestaña de trabajosen el proyecto[Música]

para ejecutar el portátil es lo mejorpráctica para insertar una celda en la parte superior paradescribir lo que cambia el cuadernoel tipo de celda para rebajar, por lo que esta celdano se tratará como código y luego se agregarála descripciónahora está listo para ejecutar el cuaderno elEl código insertado carga el conjunto de datos en unmarco de datos utilizando sus credenciales parasu instancia de almacenamiento de objetos en la nube yluego muestra las primeras cinco filas delconjunto de datos antes de volver al proyectoguardar el cuaderno: agregado a la selección Presiona [CTRL + S] para guardar como nota.

Sem 3.5: Cuadernos Jupyter en Watson Studio - Parte 2

[Música]este video le muestra cómo crear unjupiter notebook comencemos agregando unactivo de datos para el proyecto, puedenavegar para seleccionar archivos o arrastrar archivosen el panelGenial, ahora el archivo de datos está subido aalmacenamiento de objetos y disponible como datosactivo en este proyectoa continuación, cree un cuadernoproporcionar un nombre y una descripcióny luego seleccione el tiempo de ejecución para usar cuandoejecutando este cuaderno aquí se ve elentornos que podrías usar, aprenderásmás sobre entornos más adelante así que por ahorasimplemente seleccione la chispa Python predeterminadaambiente y verificar el lenguaje yversión chispa cuando esté listo, cree elcomputadora portátilahora espera mientras el entorno de tiempo de ejecuciónse instancia una vez que el entorno eslisto en la libreta accede a los datosfuentes y busque el archivo haga clic en insertarcodificar y elegir cómo quieresinserte los datos las opciones en estecuadro desplegable dependen de lalenguaje utilizado en este cuaderno y eltipo de archivo observe que el código insertadoincluye las credenciales que necesitará paraleer el archivo de datos del objetoinstancia de almacenamientocuando ejecutas el código, los primeros cincovisualización de filasahora echemos un vistazo más de cercaentornos en la pestaña de entornospuede definir el tamaño del hardware yconfiguración de software para el tiempo de ejecuciónasociado con las herramientas de estudio de Watson talescomo cuadernos se puede ver que hayun tiempo de ejecución de entorno activo, a saberel tiempo de ejecución que utiliza el portátilacabas de crear y aquí están los otrosentornos predeterminados puede ver cualquiera delos entornos predeterminados para ver unresumen de la configuraciony también crear un nuevo entornodefiniciónprimero proporcione un nombre en una descripciónsi selecciona chispa para el tipo,ver alguna configuración adicionalopciones en este caso simplemente acepte elvalores predeterminados y elija Scala para elversión de software cuando esté listo para crearel nuevo entornoel entorno está listo para que lo usecon un cuadernopara cambiar una computadora portátil para usar una diferenteentorno que necesita para detener primero el coronel entonces puedes cambiar el ambientey seleccione el entorno personalizado queacaba de crear y asociar eso con elcomputadora portátilAhora abra el cuaderno en modo de edición yesperar a que el nuevo entorno seainstanciadodesde que este cuaderno se guardó por última vez usandoun núcleo diferente que necesita para configurar elnúcleo nuevoeliminemos la celda existentebusque el archivo de datos de origen e inserte unmarco de datos de la sesión de chispa cuando ejecutas el código, los primeros cincovisualización de filasahora estás listo para explorar elcomunidad y encontrar cuadernos de muestra yconjuntos de datos para comenzar a analizar datos[Música]

Sem 3.6: Obtener código de función de IBM Cloud

Para facilitar el desarrollo práctico de habilidades y permitirle completar algunos de los proyectos de este curso, le proporcionamos un código de función especial para IBM Cloud.

Este código de función le permitirá crear una cuenta de IBM Cloud sin una tarjeta de crédito, o extender la validez de su prueba de IBM Cloud existente.

Marque la casilla y haga clic en el botón Abrir herramienta a continuación para obtener un código de función único y copiarlo. Podrá aplicarlo utilizando las instrucciones explicadas en un proyecto posterior en el curso.

NOTA: Si ya ha aplicado su código de función de IBM Cloud en otro curso/laboratorio para crear una cuenta de IBM Cloud o extender su prueba, omita este elemento, ya que el código solo se puede usar una vez.

Este curso utiliza una herramientas de terceros, Obtener código de función de IBM Cloud para mejorar tu experiencia de aprendizaje. La herramienta hará referencia a información básica, como tu nombre, correo electrónico e ID de Coursera.

Laboratorio:

Laboratorio práctico: Crear una cuenta de IBM Cloud

Tiempo estimado: 15 minutes

Resumen del laboratorio:

Para acceder a los recursos y servicios que ofrece IBM Cloud necesita una cuenta IBM Cloud. Este laboratorio le ayudará paso a paso a crear una cuenta.

Objectivos:

Después de terminar este laboratorio usted será capaz de:

Crear una cuenta IBM Cloud de prueba utilizando un código especial:

Requisitos:

Antes de comenzar este laboratorio asegúrese por favor de que cuenta con un código especial obtenido en secciones anteriores del curso y que nunca haya sido utilizado.

Nota: Estas instrucciones son unicamente para crear una cuenta NUEVA IBM Cloud utilizando una dirección de correo que no haya utilizado previamente para crear otra cuenta IBM Cloud. Si ya tiene una cuenta de prueba y quiere utilizar un Código Especial para extender su periodo de prueba siga por favor las instrucciones en lab utilizando el código de su cuenta IBM Cloud.

Si ha creado anteriormente una cuenta IBM y el periodo de prueba a terminado necesitará crear entonces una cuenta nueva con una dirección de correo diferente.

Crear una Cuenta IBM Cloud

Vaya a: https://cloud.ibm.com/registration para crear una cuenta IBM Cloud.

Ingrese su dirección de correo electrónico y una contraseña segura como se solicita y haga clic en el botón Siguiente (Next).

NOTA: Por favor asegúrese de que utiliza una dirección de correo no utilizada anteriormente para crear una cuenta IBM y a la que tiene acceso para poder recibir el código de verificación requerido en el siguiente paso.

En correo electrónico es enviado a la dirección que eligió para confirmarla. Revise su correo y copie y pegue el Código de Verificación. Después haga clic en Siguiente(Next).

Una vez verificado su correo electrónico ingrese su Nombre, Apellido y País. Después haga clic en Continuar(Next)

Vaya a los Avisos de Cuenta y active las actualizaciones vía correo si lo desea, acepte las condiciones y haga clic en Continuar,

Antes de crear su cuenta revise las políticas de privacidad y acepte que las ha leído y entendido marcando la casilla, después haga clic en Continuar.

Toma unos segundos crear y configurar su cuenta.

Seleccione Personal como su tipo de cuenta y haga clic en \*Registrar mediante un código

Nota: Por favor asegúrese de escoger la opción de Código Especial (Feature Code) para poder completar este y otros cursos. No debe utilizar la opción de Tarjeta de Crédito para verificar su cuenta ya que podrían cargarse cobros innecesarios en el proceso de activación.

Copie el código de la sección Obtener código para IBM Cloud (Claim IBM Cloud Feature Code) de su curso y pegue donde indica Ingresar Código. Haga clic en Crear Cuenta.

Nota: Puede tomar algunos minutos crear su cuenta.

Una vez que haya creado con éxito su cuenta de IBM Cloud deberá poder ver el panel de control.

Ahora puede explorar el Catálogo IBM para ver los servicios y recursos que ofrece IBM.

¡Felicidades! ¡Ha creado su cuenta IBM Cloud!

**Laboratorio: Creando un Watson Studio Project con Jupyter Notebooks**

Este tutorial te guía a través de la creación de una cuenta en la IBM Cloud y un proyecto en Watson Studio para que puedas utilizar los jupyter notebooks para tu trabajo.

Por favor, tómate un momento para seguir los siguientes pasos y aprender cómo puedes usar los Jupyter notebooks de Watson Studio.

Por favor haz clic en "Agregar al proyecto"

2. Aquí puedes seleccionar una gran variedad de herramientas, pero primero vamos por los jupyter notebooks. Por favor, haz clic en "Notebook"

3. Para no agotar tus créditos de computación gratuitos mensuales, selecciona el tiempo de ejecución "Default Python 3.6 Free"

4. Por favor haz clic en "Create notebook"

5. Espera a que aparezca el notebook. En caso de que estés interesado. El jupyter enterprise gateway ha solicitado recursos en los hosts IBM de clúster de Kubernetes para servir el kernel jupyter que respalda tu notebook

6. ¡Ahora estás listo para escribir tu código!

Esto concluye este tutorial

Sem 3.7: Enlazar GitHub a Watson Studio

[Música]este video le muestra cómo conectar suCuenta de estudio de IBM Watson con sucuenta de github en Watson Studio navegara la configuración de su perfil en elpestaña de integraciones visite el enlace agenerar un token de acceso personal de githubproporcionar un nombre descriptivo para el tokeny seleccione el alcance del repositorioluego generar el token copiar el tokenvolver a la integración de githubconfiguración y pegue el token el tokense valida cuando lo guarda en suconfiguración de perfil ahora navegue a suproyectos que habilitas la integración de githuba nivel de proyecto en la pestaña de configuraciónsimplemente desplácese hasta la parte inferior y peguela URL del repositorio de github existente una vezla URL está validada, haga clic en conectarvaya a la pestaña de activos y abra ellibreta que quieres publicar observe que este portátil tiene lacredenciales reemplazadas con X's es unmejores prácticas para eliminar o reemplazarcredenciales antes de publicar en githubpor lo que este cuaderno está listo para publicarpuede proporcionar la ruta de destino a lo largocon un mensaje de confirmación también tienes laopción de publicar contenido sin ocultarcódigo lo que significa que cualquier celda en elcuaderno que comenzaba con la celda ocultael comentario no será publicado cuandoestás listo haz clic en publicar el mensajete dice que el cuaderno estabapublicadocon éxito y proporciona enlaces a lacuaderno el repositorio en el compromisoechemos un vistazo a los compromisos paraahí está el compromiso y puedes navegaral repositorio para ver la publicacióncomputadora portátilpor último, puede publicar como esencia soloson otra forma de compartir su trabajo engithub cada esencia es un repositorio de git, por lo quese puede bifurcar y clonar haydos tipos de esencias públicas y secretas sicomienzas con una esencia secreta que puedesconvertirlo en una esencia pública más tarde yde nuevo tienes la opción de eliminarceldas ocultasasí que eso es lo básico de Watson Studiosintegración de github[Música]

**Otras herramientas IBM**

Sem 3.8: Otras herramientas de IBM para la ciencia de datos

En este video, veremosvarias otras herramientas de IBM que ayudancientíficos de datos en su trabajo diario.Watson Knowledge Catalog ayuda a los científicos de datos acatalogar y gestionar todos sus recursos de datos.La refinería de datos proporciona herramientas gráficaspara analizar y preparar datos.Los productos basados ​​en SPSS incluyeninterfaces gráficas fáciles de usar paraamplias variedades dealgoritmos estadísticos y de aprendizaje automáticoy transformaciones de datos.Hablaremos sobre enfoques para la implementación de modelos,incluidos estándares abiertos y Watson Machine Learning.Las características más nuevas de Watson Studio incluyenAutoAI que calcula automáticamentela mejor canalización de datos yWatson OpenScale que ayuda a garantizarEquidad y explicabilidad de los modelos.

Sem 3.9: Catálogo de conocimiento de IBM Watson

La mayoría de las organizaciones tienen grandes cantidades de datos almacenados en muchas formas en varios lugares.Encontrar datos relevantes rápidamente y conectar fuentes de datos dispares puede ser un desafíoy consume mucho tiempo. Watson Knowledge Catalog une todos los activos de información en un únicocatálogo rico en metadatos, basado en la comprensión de Watson de las relaciones entre activosy cómo se utilizan y socializan entre los usuarios de proyectos existentes. Vamoseche un vistazo a la descripción general de las diferentes categorías de herramientas que hemos discutido anteriormente.Watson Knowledge Catalog corresponde a Data Asset Management, Code Asset Management,Gestión de datos e Integración y transformación de datos. Watson Knowledge Catalog esun catálogo de datos que está integrado con una plataforma de gobierno de datos empresariales. Tambiénfusiona las capacidades de análisis de Watson Studio. El catálogo de datos ayuda a los científicos de datospara encontrar, preparar, comprender y utilizar fácilmente los datos según sea necesario. Catálogo de conocimientos de Watsonprotege los datos del uso indebido y permite compartir activos con sistemas automatizados y dinámicos.enmascaramiento de elementos de datos confidenciales. Visualizaciones de perfil de datos, gráficos y estadísticas integradosayudar a los usuarios a comprender los activos de datos. La perfecta integración con Watson Studio ayuda a los datosciudadanos para impulsar la producción de sus datos en un conjunto de potente ciencia de datos, IA, aprendizaje automáticoy herramientas de aprendizaje profundo. Unirse a Watson Studio dirige la construcción, capacitación ydespliegue de modelos. Los usuarios pueden descubrir, limpiar y preparar datos de forma interactiva con unrefinería de datos incorporada. Posibles conexiones a más de 30 fuentes de datos de IBM y de tercerosayudar a catalogar y utilizar sus datos en las ubicaciones originales. IBM Watson Knowledge Catalog tienevarias opciones de implementación en IBM Cloud™ y se puede ejecutar en cualquier lugar con IBM Cloud Pak™para datos. Esta última es una plataforma de inteligencia artificial y datos completamente integrada basada en Red Hat® OpenShift®Fondo de contenedor. Se puede implementar fácilmente en cualquier nube pública o privada u otraplataformas empresariales. Un catálogo contiene metadatos sobre los contenidos.de activos y cómo acceder a ellos. Y un conjunto de colaboradores que necesitan usar los activospara el análisis de datos. Los metadatos se almacenan en una instancia de almacenamiento de objetos de IBM Cloud cifrada.Cualquier dato que desee almacenar en la nube, puede cargarlo en el almacenamiento de objetos en la nubede su elección y, a continuación, especifique ese almacenamiento de objetos cuando cree el catálogo. Estedividir entre dónde se almacenan los metadatos de los datos y la ubicación real de los datoses importante. Significa que puede mantener sus datos donde quiera que estén. No necesitas moverteen el catálogo porque el catálogo solo contiene metadatos. Puedes tener los datos enrepositorios de datos fuera de las instalaciones en otros servicios en la nube de IBM como Cloudant o Db2 on Cloudy en servicios en la nube que no son de IBM como Amazon o Azure, en servicios de transmisión de datos o inclusofuentes de datos oscuros como archivos PDF. En los metadatos se incluye cómo acceder al recurso de datos.En otras palabras, la ubicación y las credenciales. Eso significa que cualquiera que sea miembro deel catálogo y tiene suficientes permisos puede acceder a los datos sin conocer las credencialeso tener que crear su propia conexión con los datos. Dado que el nuevo catálogo está vacío,echemos un vistazo a un catálogo existente. En la pestaña Examinar activos puede ver "recomendaciones","activos de alta calificación" y "activos de creación reciente", así como una lista de todos los activos.Puede escribir un término de búsqueda para encontrar activos y puede filtrar por tipo de activo, comoActivo de datos o cuaderno. O filtre por etiquetas que se asignaron al recurso cuando seañadido al catálogo. Cuando ve un activo, obtiene una vista previa de los datos y otra informacióncomo una descripción, calificaciones, etiquetas, dónde se encuentra la fuente y cualquier clasificación.En la pestaña Acceso, aquellos con permiso pueden agregar miembros para ver este activo en particular.Y la pestaña Revisar muestra reseñas y le permite contribuir con una reseña. Cuando se agregan activosa un catálogo con Políticas de datos habilitadas, Watson Knowledge Catalog perfila yclasifica el contenido del activo en función de los valores de esas columnas. El perfilLa pestaña contiene información más detallada sobre las clasificaciones inferidas. Puedes verlas otras posibilidades para clasificar cada columna y los puntajes de confianza para esosotras posibilidades. En la pestaña Linaje, verá los diversos eventos que Watson KnowledgeEl catálogo ha capturado lo que ocurrió en el ciclo de vida de este activo de datos, lo que le permitepara rastrear lo que le sucedió al activo desde que se creó. En la pestaña Control de acceso,puede ver la lista actual de miembros del catálogo. también puedes agregar miembros, lo cual es bastante similarpara agregar colaboradores en un proyecto. La mayoría de los miembros del catálogo probablemente tendrán el editorrole. La función de espectador está intencionalmente restringida y solo unos pocos usuarios selectos tendrán larol de administrador Watson Knowledge Catalog incluye capacidadespara enmascarar automáticamente los datos confidenciales de acuerdo con las políticas de gobierno de su organización.Por ejemplo, puede ver en el diagrama que los datos de nombre, apellido y géneroen el conjunto de datos han sido enmascarados. Ha aprendido cómo IBM Watson Knowledge Catalog puedeayudar a las organizaciones a manejar sus numerosos datos y otros activos. En el próximo video vamos amire Data Refinery, una poderosa herramienta para analizar y preparar datos.

Sem 3.10: Refinación de Datos

Hola, soy Sonali Surange Dev.

Los científicos de datos a menudo terminan gastando muchode tiempo haciendo tareas mundanas como limpiar, dar forma y preparar datos.Por lo general, estas tareas son obstáculos para comenzar la parte más agradable deanalizando los conjuntos de datos o construyendo y entrenando modelos de aprendizaje automático.

Esto se debe a que los conjuntos de datos normalmente no están en un formato que se pueda usar fácilmente.

Primero deben limpiarse y refinarse antes de que un científico de datos pueda utilizarlos.

IBM Data Refinery soluciona este problema y simplifica la tarea de refinar datos ysus flujos de trabajo.Proporciona un entorno de preparación de datos de autoservicio en el quepuede analizar, limpiar y preparar rápidamente conjuntos de datos.La refinería de datos está disponible conWatson Studio en nube pública, nube privada y escritorio.En el resto delvideo recorreremos un escenario y veremos Data Refinery en acción.En estoescenario usaremos Data Refinery para encontrar las mejores ofertasutilizando datos sobre descuentos ofrecidos a lo largo del tiempo.A continuación, automatizaremos elanálisis para ejecutarse en un horario regular.Antes de que el científico de datos comience, mira los datosdistribución y advierte que faltan datos en la columna inSale.Ella visualiza la columna de ofertas y nota que contiene valiososinformación sobre descuentos.Muchos campos contienen el porcentaje de información,algunos contienen referencias al precio anterior indicando un nuevo precio reducidoestando disponible.Ella decide derivar la venta de la oferta.Ella usa una operación de disminución condicional para derivar si el producto está enrebaja.A continuación, utiliza una operación de filtro.para eliminar ofertas que no están en ofertaLuego quiere recoger las gangas.Ella usa la operación de subcadena de reemplazoy proporciona un patrón que extrae los descuentos de la oferta.Despuésconvertir los valores de descuento a un decimalella puede ver visualmente los descuentos que estaban disponibles.Ella necesita encontrar elmeses que ofrecieron las mejores ofertas.Visualiza la fechaActualizado y avisosque el campo de fecha tiene una variedad de formatos, algunos con guiones, otros conbarras y algunos con meses como texto.Ella espera que Data Refinery pueda normalizar ladatos y extraer un mes.Utiliza la operación de conversión de columna para convertir adate y selecciona amd.A continuación, extrae el mes y crea una columna derivada.llamado descuentoMes.Los datos ahora representan todas las marcas y productos.proporcionando las ventas y el mes en que la oferta estuvo disponible.El científico de datos es solointeresada en sus marcas preferidas.Con el tiempo ha construido una lista de favoritosmarcas y ha importado los datos en su proyecto.La refinería de datos proporcionatransformaciones relacionales como izquierda, derecha interna, completa, semi y anti-unión.Para asegurarque los datos solo contienen su marca preferida usa un semi-joinoperación que reduce las marcas para que coincidan con sus preferencias.Luego selecciona las claves para la unión y los campos resultantes.Los resultados visuales ahora confirman que las marcas coinciden con las preferencias.Para encontrar lo mejorposibles tratos que necesita para realizar algunas agregaciones.Varias características determinan un buen trato.Ella está interesada en la mejor oferta yduración cuando los descuentos están activos.Agregar los datos de venta ayudaráentender las ofertas.Agrupa las columnas por marca y descuentoMes ycalcula el descuento máximo.Finalmente ordena el resultado en orden descendente.La refinería de datos ahora muestra las mejores ofertas por preferencias de marca y elduración de la oferta disponible.El último paso es ejecutar el análisis en el conjunto de datos completo.ella comienza la completaanálisis, que ella puede monitorear para el estado de finalización.Es hora de automatizar el análisis que se ejecuta regularmente.Los datos en elLa base de datos puede crecer con el tiempo.Ella usa un tiempo de ejecución personalizado para que coincida con el más grandevolúmenes de datos y establece un cronograma para la automatización.El horario por horas dicea partir de datos actualizados de la base de datos y escribe en la tabla de destino.Data Refinery la ha ayudado a descubrir acuerdos enlos datos sin procesar a través de un pequeño conjunto deoperaciones y transformaciones con la mayor parte del trabajo realizado para ella.Graciaspara ver

Sem 3.11: SPSS Modeler Flows en Watson Studio

En este video, veremos una forma gráfica y fácil de usar de construir máquinasmodelos de aprendizaje y canalizaciones.SPSS Modeler Flows es parte de Watson Studio, que se inspiró en otro producto, IBM SPSS Modeler.

Hablaremos de ese producto en una unidad posterior.Echemos un vistazo de nuevo a la descripción general de las diferentes categorías de herramientas.Los flujos del modelador incluyen algunas capacidades de gestión de datos, así como herramientas para la preparación de datos,visualización y construcción de modelos.Todos los flujos se crean utilizando un editor de arrastrar y soltar y consisten en "nodos" de variostipos, con datos que "fluyen" de un nodo al siguiente según sus conexiones.Un flujo de modelador de muestra que se muestra aquí incluye dos nodos de fuente de datos que se muestran en púrpura a la izquierda;escriba, agregue, filtre, fusione, rellene y divida los nodos en el medio;2 nodos de construcción de modelos mostrados en pentágonos.Una vez que se ejecuta un flujo y se construyen los modelos, se crean los "nuggets de modelo" del pentágono invertido.

Se pueden usar para ver información sobre los modelos y obtener predicciones para nuevosdatos. Y los tres nodos cuadrados verdes a la derecha brindan información de evaluación del modelo en forma de tablas y gráficos.

Puede crear sus flujos de SPSS Modeler arrastrando diferentes tipos de nodos desde la izquierda,la parte de la pantalla llamada "paleta", al "lienzo", la parte principal de la pantalla.Cada flujo comienza con una o más fuentes de datos ubicadas en el grupo "Importar",y puede incluir algunos o todos los demás tipos de nodos.Watson Studio proporciona algunos flujos de muestra para ayudar a los nuevos usuarios.En el ejemplo del estudio de fármacos que se muestra aquí, estamos utilizando un pequeño conjunto de datos artificiales.La variable de destino es un campo categórico, "Droga", que tiene cinco categorías, yhay varias variables predictoras.Este flujo crea un nuevo campo "derivado" dividiendo los valores de uno de lospredictores por valores de otro, y al final construye una pequeña red neuronalmodelo y un modelo de árbol de decisión.Cuando un usuario hace clic en el botón "Ejecutar" en el panel superior, indicado por un triángulo,se ejecuta el flujo y se construyen los modelos.Esto se refleja en los nuevos nodos amarillos, llamados "nuggets de modelo", que se muestran debajocada nodo del modelo.Si hace clic en los tres puntos en la esquina superior derecha de uno de esos nodos y selecciona“Ver modelo”, verá varios tipos de información del modelo.Al conectar nuevas fuentes de datos al nugget de modelo, puede obtener predicciones sobre nuevos datos.La primera ventana del visor de modelos muestra la precisión del modelo y las medidas relacionadas, comocomo precisión y recuerdo.Este ejemplo de datos de juguetes nos permitió obtener una precisión perfecta, que normalmente no es el caso condatos de la vida real.La vista Matriz de confusión muestra cómo las predicciones del modelo en los datos de entrenamiento coincidieron con elvalores objetivo observados.Una vez más, en este ejemplo de juguete, todos los casos se clasificaron correctamente.También podemos ver Información del modelo, que muestra una tabla que nos brinda más información sobrelos detalles del modelo.Feature Importance muestra un diagrama que indica la fuerza predictiva relativade varias entradas del modelo.Finalmente, el Diagrama de red ofrece una representación visual del modelo de red neuronal.construimos.A la izquierda está la capa de entrada, con unidades correspondientes a cada predictor continuoy cada categoría de los predictores categóricos, más una unidad de sesgo que suele estar presente encada capa de una red neuronal.En el medio, vemos una "capa oculta" con 7 unidades, o neuronas, y una unidad de polarización.A la derecha está la capa de salida con 5 unidades correspondientes a las cinco categorías objetivo.Los controles a la derecha y en la parte inferior del diagrama permiten una exploración interactiva delmodelo.Los colores de las conexiones entre unidades indican los valores de los pesos en esasconexionesTambién podemos ver el modelo de árbol de decisiones creado con el algoritmo C5.Una tabla de información del modelo y un gráfico de importancia de características aparecen como antes.Además, se muestra una tabla de Reglas de decisión principales.Los modelos de árboles de decisión son populares porque tienen una estructura especial que facilitapara explicar predicciones o extraer reglas de decisión.También se muestra el diagrama de árbol.En el lado izquierdo del lienzo, vemos una parte de la paleta modelo que se puede usar en elfluyeEn la parte superior están los nodos "Clasificador automático" y "Autonumérico" que se pueden usar para clasificación categórica.y objetivos continuos, respectivamente.Esos nodos construirán varios tipos de modelos y elegirán el mejor en función de un criterio determinado.Más adelante, hablaremos sobre la función AutoAI de Watson Studio; AutoAI toma esta capacidadal siguiente nivel encontrando automáticamente no solo el mejor modelo, sino un conjunto completo de datoscanalización, que incluye varias transformaciones de datos.En este video, aprendió cómo Modeler Flows en Watson Studio puede ayudar a los analistas acree potentes canalizaciones de aprendizaje automático mediante una interfaz gráfica sin necesidadpara escribir cualquier código.Esta función se basó en IBM SPSS Modeler.Luego, después de completar un laboratorio para brindarle experiencia práctica con esta poderosa tecnología,veremos otros dos productos de IBM que se pueden usar para Data Science: IBM SPSSModelador e IBM SPSS Statistics.

Laboratorio:

Laboratorio: Modeler Flows en IBM Watson Studio

Para este laboratorio, utilizarás la cuenta de IBM Cloud y el proyecto de Watson Studio que creaste cuando trabajamos en el laboratorio en la lección anterior titulada Creando un Proyecto de Watson Studio con Jupyter Notebooks. Para crear una cuenta gratuita de Watson Studio, usa el siguiente enlace https://cloud.ibm.com/registration/

En este laboratorio haremos lo siguiente:

Cargar un flujo de ejemplo, ejecutarlo y examinar los resultados.

Agregar un Modelo Auto Numérico, ejecutarlo y examinar los resultados.

Obtener predicciones para nuevos casos usando un modelo que hemos construido.

Paso 1 – Abre tu proyecto en Watson Studio, luego haz clic en "Add Assets" haciendo clic en el botón azul en la parte superior de la pantalla. En el panel que aparece, selecciona la opción "Modeler flow".

Paso 2 – A continuación, selecciona la pestaña "From example":

Paso 3 – Seleccionando el ejemplo de Drug study a la izquierda nos daría el flujo que ya hemos visto en la sección anterior, así que escojamos el de la derecha – Sales Promotion study, luego haz clic en el botón "Create" en la esquina inferior derecha.

Una vez que el flujo se cargue, verás esto:

Paso 4 – Leyendo de izquierda a derecha, podemos examinar el flujo. El circulo purpura corresponde al conjunto de datos de algunas ventas ficticias, el siguiente nodo está derivando un campo Increase que será nuestra variable objetivo. Se basa en el aumento de ventas. Entonces vemos un tipo de nodo que especifica nuestra variable objetivo y los predictores. En este caso nuestra variable objetivo es numérica y continua. Finalmente, hay dos modelos: una red neuronal y un árbol de decisión C&RT. Podemos ejecutar el flujo tal cual o podemos modificarlo para ver otras posibilidades.

Paso 5 – Primero, veamos qué podemos obtener con los modelos existentes. Presiona el botón triangular Run en la parte superior del canvas:

Paso 6 – Después de un tiempo de ejecución, obtendrás dark gold model nuggets para los dosmodelos. Haciendo clic en tres puntos en el lado derecho de uno de ellos puedes seleccionar "View model" y examinar la información.

Por ejemplo, el diagrama de la red neuronal se ve así:

Y el árbol de decisión se ve así:

Paso 7 – Ahora, agreguemos el modelo "Auto-numeric" al flujo. Abre la paleta de modelado en el lado izquierdo haciendo clic en Modeling:

Step 8 – En la paleta de modelado resultante, elije el modelo Auto-Numeric:

Paso 9 – Haz clic en el nodo Auto Numeric y mientras mantienes el botón izquierdo del ratón presionado, arrástralo sobre el canvas.

Paso 10 – Ahora, pasa el cursor sobre el nodo de tipos con el ratón y haz clic en el círculo con una flecha que aparece en el lado derecho. Haz una conexión con el punto en el lado izquierdo del nodo Auto Numeric.

Paso 11 – Finalmente, ejecuta la nueva branch del flujo haciendo clic en los tres puntos de la parte superior derecha del nodo Auto Numeric y seleccionando la última opción del menú, "Run".

Paso 12 – Después de que la ejecución termine, obtenemos el model nugget:

y al observar el modelo, vemos la siguiente tabla:

Esto nos dice que se han construido cinco modelos diferentes y, también algunas propiedades de los modelos.

XGBost es un modelo muy popular, que representa un conjunto de árboles de decisiones potenciados por el gradiente. El algoritmo se descubrió hace relativamente poco tiempo y se ha utilizado en muchas soluciones y en competiciones ganadoras de ciencias de la información. En este caso, creó un modelo con la mayor precisión, que también "triunfó".

"C&RT" significa "Classification and Regression Tree", o en español Árbol de Clasificación y Regresión, un algoritmo del árbol de decisiones que se utiliza ampliamente. Este es el mismo árbol de decisiones que vimos antes cuando lo construimos por separado.

"LE" es "linear engine", o en español Motor Lineal, es una implementación de IBM del modelo de regresión lineal que incluye la detección automática de interacción. Los coeficientes del modelo se muestran en la tabla de Estimaciones de Parámetros. Podemos ver que varios coeficientes corresponden a una combinación de una categoría de class variable con una variable continua Promotion. Esto se denomina "un efecto de interacción". No se verán tales características en los modelos de regresión lineales simples.

A continuación, en la tabla de modelos estaba CHAID. Es otro algoritmo para construir árboles de decisiones. El acrónimo CHAID significa Chi-squared Automatic Interaction Detector, o en español, Detector de Interacción Automática Chi-cuadrado. Es uno de los más antiguos métodos de clasificación de árboles propuesto originalmente por Gordon Kass en 1980. Al hacer clic en el nombre del algoritmo podemos ver los detalles de ese modelo. La mayoría de los algoritmos de árboles de decisiones (incluyendo C&RT) construyen árboles binarios, es decir, árboles en lo que cada nodo tiene cero o dos hijos. CHAID es diferente, produce árboles con múltiples hijos para algunos nodos. Estos árboles pueden ser no tan profundos, pero sí muy anchos. Vemos un ejemplo aquí. El nodo raíz tiene cuatro hijos, con la división basada en la variable class. Cada nodo hijo tiene tres o cuatro hijos propios, todos divididos en base a la Promotion. Así que esencialmente, tenemos una interacción de class y Promotion en este modelo también.

Finalmente, el último modelo construido por el nodo Auto Numeric es "MLP Neural Network", es la misma red neuronal que en flujo original. "MLP" significa "multi-layer perceptron" o en español, perceptrón muti capas, un nombre utilizado para las redes neuronales de alimentación, completamente conectadas populares en los años 90.

Paso 13 - ¿Cómo podemos utilizar un modelo que hemos construido para poder obtener predicciones de nuevos datos (o tal vez los datos que ya hemos utilizado)? En los flujos de Modeler solo se requieren unos pocos pasos.

Crear un nodo fuente de datos para los nuevos datos, adjuntar un nodo tipo, conectar eso al model nugget, y añadir un nodo "table" después del model nugget. En nuestro ejemplo solo obtendremos predicciones para los datos originales usando el modelo C&RT, así que abrimos el grupo "Output" en la paleta de nodos, arrastramos el nodo "Table" y le añadimos la conexión desde el model nugget:

Paso 14 – Ahora ejecuta esa nueva branch, y consigue un nodo table en la parte derecha:

Haciendo doble clic en ese nodo podemos ver los datos con la nueva columna añadida $R-Increase (que contiene predicciones):

Paso 15 – Ahora puedes experimentar con flujos de Modeler por tu cuenta. Puedes usar varios nodos, pero no olvides poner un nodo tipo antes de cualquier nodo modelado.

Algunos otros posibles modelos que pueden ser usados para predecir un objetivo continuo son: Genlin, GLMM, LSVM, Regression, KNN, XGBoost Linear. Si tus datos tienen un objetivo categórico, puede que quieras probar el Auto Classifier, así como C5, C&RT, CHAID, Logistic Regression, SVM, Neural Networks y muchos otros modelos.

Aprenderás más sobre varios modelos cuando tomes un curso sobre Aprendizaje Automático.

Sem 3.13: Modelador IBM SPSS

En esta lección discutiremos dos productos que son muy útiles para los datoscientíficos. Ambos llegaron a IBM con la adquisición de SPSS en 2009. El primero es IBMModelador SPSS. Revisemos las diferentes categorías de herramientas que discutimos anteriormente.IBM SPSS Modeler incluye funciones y herramientas de gestión de datos parapreparación de datos, visualización, construcción de modelos y despliegue de modelos. Élproducto fue creado por Integral Solutions Limited en el Reino Unidoen 1994 y originalmente se llamaba Clementine. fue adquirida por una empresallamado SPSS en 1998 y SPSS a su vez fue adquirido por IBM en 2009. SPSS Modeler esuna aplicación de software de minería de datos y análisis de texto. se usa para construirmodelos predictivos y realizar otras tareas de análisis. tiene una visualinterfaz que permite a los usuarios aprovechar los algoritmos estadísticos y de minería de datossin programación. Uno de sus principales objetivos desde el principio fue crearcanalizaciones complejas de modelado predictivo que son fácilmente accesibles. Una muestraEl flujo del modelador que se muestra aquí incluye un nodo de origen de datos redondo, tresnodos gráficos, un nodo hexagonal para computación, una nueva variable y un cuadradonodo para una tabla de salida. Debajo del lienzo, podemos ver la rica paleta de nodos.con pestañas separadas para fuentes de datos, registro en operaciones de campo, gráficos, modelos,salida y así sucesivamente. Los nodos y las diferentes pestañas tienen diferentesformas con el Pentágono utilizado para modelar nodos. Examinemos el flujo de muestra.eso viene como un ejemplo con el producto. Comienza con un conjunto de datos deregistros de telecomunicaciones y el objetivo es construir un modelo para predecir quélos clientes están a punto de abandonar el servicio, también conocido como abandono. la fuente de datosse muestra por el nodo redondo en el lado izquierdo, un nodo tipo hexágono normalmente sigueun nodo de origen de datos y nos permiteespecificar roles, predictor de destino o ninguno. Y niveles de medición tales comonominal continuo o bandera para todas las variables. El término bandera se utiliza paradenota una variable con dos categorías, una de las cuales puede considerarse positivay el otro negativo. En este ejemplo, el nivel de medición para el churnel campo se establece en indicador y el rol se establece en destino. Todos los demás se establecen comopredictores y entradas. El conjunto de datos original tiene muchos campos y algunos de ellosno son relevantes para la variable de destino, por lo que primero debemos decidir qué camposson más útiles como predictores. Hay un nodo de modelado de selección de características queayuda a hacer esto. Después de ejecutar la secuencia con el nodo de selección de características,el nugget de modelo amarillo se crea debajo de él en el diagrama de flujo. Usar ese nuggetpodemos generar un nodo de filtro que filtra las variables que no sonbuenos predictores para el objetivo. El nodo de auditoría de datos ubicado debajo del filtronodo muestra varias propiedades de los datos, como el número de valores atípicos en cadavariable y el porcentaje de valores válidos. También puede ayudar a crear unnodo especial para imputación de valor faltante que reemplazavalores de una variable con algunos valores válidos que se pueden seleccionar en función deconocimiento del dominio. Aquí el peaje de registro variable tiene más del 50% de valores faltantes yespecificaremos un valor la media para reemplazarlos. Un supernodo en Modeler esun nodo especial que no se encuentra en la paleta pero es creado por el usuario confunciones especiales incluidas en él. El nodo de auditoría de datos nos permite crear unsupernodo para imputar valores faltantes. Tiene forma de estrella y se muestra en elderecha de la pantalla. Por último adjuntamos la logísticanodo del modelo de regresión a la secuencia y haga clic en ejecutar. Aparece otro nugget de modeloy al hacer clic en él podemos ver información diversa del modelo y otra salida. en la ventana de salida que se abre cuando hacemos clic en el modelo nugget el resumenLa pestaña muestra las entradas de destino y algunas configuraciones de construcción de modelos. Basado en ciertosconfiguraciones de salida avanzadas que se especificaron antes de que se construyera el modelotambién puede ver una tabla de clasificación, precisión y algunos otros generadossalidas para el modelo. Tenga en cuenta que estos resultados se basan únicamente en datos de entrenamiento.Para evaluar qué tan bien el modelo genera otros dos datos del mundo real, debeutilice siempre un nodo de partición para mantener un subconjunto de registros con el fin deprueba y validación. Luego, en la pantalla de configuración del modelo, seleccione el uso particionadocasilla de verificación de datos. Esto ayudará a detectar y evitar el modelo.sobreajuste El sobreajuste se define como tener una precisión significativamente mayor enlos datos de entrenamiento. Datos utilizados para entrenar el modelo y luego en pruebas o datos no vistos.El nugget de modelo amarillo agregado anteriormente también se puede usar para calcular predicciones,también llamados puntajes en los datos originales o en una nueva fuente de datos. Todo lo que necesitamos parahacer es conectar la fuente de datos en cuestión al nugget, asegúrese de que tengalas variables predictoras utilizadas en el modelo y crear una salida en una tabla ootra estructura para almacenar las puntuaciones. También podemos especificar ajustes para puntuardentro de la nugget de modelo. Tenga en cuenta que si el modelo se construyó sobre un predictor transformadodatos, los mismos pasos de transformación de datos se aplicarían a los nuevos datos antespuede ser puntuado por el modelo. El nodo de análisis es el nodo final en elarroyo. Se adjunta a un nugget de modelo y, cuando se ejecuta, calcula algún modelométricas de evaluación, como matriz de confusión y precisión. En este ejemplosolo hemos visto un modelo de regresión logística. IBM SPSS Modeler ofreceuna rica paleta de modelado que incluye muchas clasificaciones, regresionesagrupamiento, reglas de asociación y otros modelos. También contiene grandes selecciones.de tipos de fuentes de datos, transformaciones de datos, gráficos,y notas de salida. Y ni siquiera hemos hablado de análisis de texto, entidadresolución y muchas otras características del producto que pueden ser extremadamenteútil para los científicos de datos. Podríamos crear un curso completo en IBM SPSSModelador solo. Ha aprendido cómo IBM SPSS Modeler ayuda a los analistas a crearPotentes canalizaciones de aprendizaje automático mediante interfaz gráfica. A continuación, haremoshable sobre el producto SPSS original que ahora se llama IBM SPSS Statistics.

Sem 3.14: IBM SPSS Estadísticas

IBM SPSS Statistics evolucionó a partir de un producto original que se lanzó en 1968. Ese productose denominó “Paquete Estadístico para Ciencias Sociales” o “SPSS”.IBM SPSS Statistics es una aplicación de software de aprendizaje automático y estadística y es ampliamenteutilizado en la academia, las agencias gubernamentales y las grandes empresas. Se utiliza para construir predictivomodelos, realizar análisis estadísticos de datos y realizar otras tareas analíticas. Tiene uninterfaz visual, que permite a los usuarios aprovechar los algoritmos estadísticos y de minería de datos sinprogramación, aunque la interfaz es muy diferente a Modeler. Como puedes ver, ella sección principal de la pantalla se parece mucho a una hoja de cálculo; muestra datos y permiteedición manual. Este pequeño conjunto de datos en particular, llamado "Datos de empleados", fue creadohace algún tiempo y no representa a personas reales. Se envía con el producto parautilizar en demostraciones y tutoriales.En la parte inferior de la pantalla, podemos ver dos pestañas: Vista de datos y Vista de variables. En elVista de variables, podemos ver y editar la información sobre todas las variables, incluidos nombres, etiquetas,tipos de datos y niveles de medición. También podemos especificar etiquetas para valores de categoríasvariables y valores perdidos.En la parte superior de la ventana de datos hay un menú. En Archivo, si selecciona "Importar datos",verá una lista de una amplia variedad de formatos de datos que puede importar. El producto utilizasu propio formato de archivo de datos con la extensión “.sav” que guarda toda la informaciónsobre las variables que acabamos de ver en la vista Variable. El menú permite importar desde yexportar a muchos otros formatos.En "Datos", encontrará un amplio menú de posibles operaciones de datos. Tenga en cuenta queLa validación de datos se puede realizar utilizando reglas definidas por el usuario que especifican el comportamiento esperado devalores variables. Por ejemplo, si la fecha y el mes se mantienen en columnas separadas, ella fecha no puede exceder el "31", pero para febrero, la fecha no puede exceder el "29". Un especialPor lo tanto, la regla se puede crear y aplicar durante la validación de datos. Además, ustedpuede habilitar algunas comprobaciones, como el porcentaje de valores faltantes en un registro o en el campo.Al hacer clic en el elemento de menú "Transformar", encontrará una variedad de datos disponiblestransformaciones. En "Calcular variable..." puede escribiruna fórmula para una nueva variable basada en variables existentes. Puede usar cualquiera de las muchas matemáticasy funciones estadísticas disponibles en el producto.También tiene la opción de usar la preparación automática de datos, similar a Modeler.En el menú "Analizar", verá muchos tipos de aprendizaje estadístico y automáticoanálisis. En "Regresión", hay una variedad de modelos relacionados con la regresión. Alláhay otros tipos de regresiones que aparecen por separado en el menú Analizar, incluyendoModelo Lineal General, Modelos Lineales Generalizados, Modelos Mixtos y Loglinear.Ahora construyamos un modelo de árbol de decisiones sobre los datos. Para este ejercicio intentaremospredecir el campo "Categoría de empleo" en función de otros campos. En el menú "Analizar",seleccione “Clasificar” y luego “Árbol”. <Haga clic> En la ventana Árbol de decisiones, podemosespecifique la variable dependiente "Categoría de empleo" y use la mayoría de los demás campos, exceptoid y bdate, como predictores o variables independientes. Por lo general, la variable ID debeno debe usarse como predictor, porque no ayudará con casos nuevos, y la fecha de nacimientotampoco parece ser un predictor útil en este ejemplo. Seleccionaremos “ExhaustivoCHAID” como nuestro Método de Cultivo, aunque también hay otras tres opciones disponibles. Datoslos científicos a menudo prueban muchos modelos diferentes para ver cuál funciona mejor para sus datos.Aquí solo estamos viendo un modelo de ejemplo para ilustrar cómo funciona el producto.Haga clic en el botón "Validación" para abrir la ventana Validación del árbol de decisión. Aquí,seleccionamos "Validación de muestra dividida" para asegurarnos de probar el modelo con nuevos datos. Hacer clicHaga clic en "Aceptar" en la ventana Árbol de decisión, para <Hacer clic> generar la salida, incluido el diagrama de árbolmostrado aquí. <Haga clic> También se muestra una tabla de clasificación que muestra qué tan bienel modelo funciona con datos de entrenamiento y prueba. En este caso, la precisión es del 91,2% en el entrenamiento.datos y solo el 85,6% en datos de prueba, lo que significa que el modelo no se generaliza a nuevos datosmuy bien. Es posible que usando diferentes modelos podamos obtener mejores resultados.Pasemos al siguiente elemento del menú. Al hacer clic en "Gráficos", abrirá un versátilChart Builder, además de varias otras opciones.Chart Builder nos permite elegir un estilo de la galería y arrastrar los campos requeridosen el lienzo, seleccione colores y elija entre otras opciones.Aquí hay un ejemplo después de arrastrar "Experiencia anterior", "Salario actual" y Génerovariables a las ranuras correspondientes para definir el eje y los colores de los puntos en el gráfico.La trama en el lienzo no se basa en datos reales, este ejemplo simplemente le da una ideade que esperar.Aquí está la trama real obtenida a partir de los datos que hemos estado usando. Se muestra diferentepuntos de colores para el género y líneas de regresión que muestran la relación de la corrientesalario a la experiencia previa para cada género.A lo largo de IBM SPSS Statistics, verá un botón "Pegar". Al hacer clic en "Pegar"botón, en lugar de ejecutar la tarea de inmediato, la aplicación abrirá otra ventana,llamado editor de sintaxis. Aquí, puede ver el código llamado "sintaxis" pegado por usted.La sintaxis de SPSS es un lenguaje de programación especial.Por ejemplo, aquí está el código para el árbol de decisión que acabamos de construir. Una vez que tenemos la sintaxis,podemos ejecutarlo, editarlo manualmente, almacenarlo para su uso posterior o enviarlo a otros usuariosde IBM SPSS Estadísticas. Los usuarios experimentados de SPSS pueden escribir el código desde cero, mientras que otrosmight prefer to have it generated by the graphical interface. Remember, the option to paste syntax is available in throughout the program. If the syntax is generated by all the steps in a data analytics process -- opening the data set, applying any data transformations, building models -- and then saved as a syntax file with the extension “.sps”, it’s similar to saving a stream in IBM SPSS Modeler. However, one important difference is that it does not allow for an easy way of scoring new records with the model. We’ll talk about different ways to deploy models in the next section. You’ve learned how IBM SPSS Statistics helps data scientists to analyze their data using many statistical and machine learning techniques. Using a graphical user interface, we can create complicated analysis that can be saved in the form of syntax and reused later. A continuación, hablaremos sobre la implementación de modelos predictivos, una parte importante delciclo de vida de la ciencia de datos.

Sem 3.15: Modelo de Despliegue con Watson Machine Learning

Hasta ahora, hemos hablado sobre la creación de modelos y canalizaciones de aprendizaje automático. En lo más prácticoaplicaciones, el retorno de la inversión se obtiene cuando el modelo o tubería se poneen producción, donde se utiliza para obtener predicciones, o puntajes, para los nuevos casos.Miremos hacia atrás en nuestra descripción general de las diferentes categorías de herramientas. En esta unidad, Despliegue de modeloses nuestro enfoque.Suponga que trabajó duro para crear el mejor modelo de aprendizaje automático posible y los datostubería de preparación para ello. ¿Cómo implementará sus modelos?En muchos escenarios prácticos, los modelos son construidos e implementados por diferentes equipos, utilizando diferentesprogramación, y tal vez lenguajes humanos. Los equipos utilizarán diferentes sistemas informáticos yentornos de almacenamiento de datos, y podría resultar difícil traducir su programa y elpasos asociados de preparación y posprocesamiento de datos de un entorno a otro.Actualmente hay varios enfoques que puede utilizar para resolver este problema, algunos comerciales,algunos de código abierto. Sin embargo, cada uno generalmente admite solo un subconjunto de todos los modelos posibles, desdeconstruyéndolos para implementarlos, por lo que un usuario queda bloqueado en un marco específico.Los estándares abiertos para la implementación de modelos están diseñados para admitir el intercambio de modelos entre unvariedad de modelos propietarios y de código abierto. Lenguaje de marcado de modelo predictivo, o "PMML",fue el primer estándar de este tipo, basado en XML.Fue creado en la década de 1990 por Data Mining Group, un grupo de empresas que trabajan juntassobre los estándares abiertos para el despliegue de modelos predictivos. IBM y SPSS estuvieron entre los fundadoresmiembros del Grupo de Minería de Datos.PMML 4.4 fue lanzado recientemente.Incluye 17 modelos estadísticos y de aprendizaje automático y muchas transformaciones de datos, integradosfunciones, formas de combinar varios modelos y otras características. este estándares ampliamente conocido y utilizado. Los productos que analizamos anteriormente: Watson Studio, IBM SPSSEstadísticas, IBM SPSS Modeler: permite a los usuarios exportar la mayoría de los modelos en PMML.En 2013, creció la demanda de un nuevo estándar, uno que no describiera los modelos y suscaracterísticas, sino el procedimiento de puntuación directamente, y uno que se basó en JSON en lugar deque XML. Esto condujo a la creación de Portable Format for Analytics, o PFA. Ahora se utiliza PFApor una serie de empresas y paquetes de código abierto.Después de 2012, los modelos de aprendizaje profundo se hicieron muy populares. Sin embargo, PMML y PFA no reaccionaron rápidamente.suficiente para su proliferación. La necesidad de una representación intermedia estándar fueamplificado por la amplia variedad de marcos emergentes de aprendizaje profundo y hardware especializado.En 2017, Microsoft y Facebook crearon y abrieron Open Neural Network Exchangeo “ONNX”. Originalmente creado para redes neuronales, este formato se amplió posteriormente atambién admite el "aprendizaje automático tradicional".Actualmente hay muchas empresas trabajando juntas para seguir desarrollando y expandiendo ONNX,y una amplia gama de productos y paquetes de código abierto están agregando soporte para él.Watson Machine Learning es la oferta comercial de IBM diseñada para la implementación de modelos. Esoadmite la implementación de modelos creados con la mayoría de los paquetes de código abierto, así como los expresadosen PMML o ONNX. También es compatible con la implementación de secuencias y flujos de Modeler de IBM SPSS Modeler.de Watson Studio. La implementación se puede realizar mediante una interfaz gráfica o código Python,y puede ser para puntuación en línea a través de una API REST o puntuación por lotes.Watson Machine Learning ayuda a integrar un modelo implementado en aplicaciones en formade fragmentos de código en varios lenguajes de programación.En este video, aprendió cómo los estándares abiertos y Watson Machine Learning pueden ayudar a los usuariospara implementar sus modelos en varias aplicaciones.A continuación, hablaremos sobre AutoAI y OpenScale, dos características avanzadas de Watson Studio que ayudanpara simplificar aún más el trabajo de un científico de datos.

Sem 3.16: Auto AI en Watson Studio

En secciones anteriores, vimos cómo los flujos de IBM SPSS Modeler y Watson Studio Modeler le permitenpara crear gráficamente una secuencia o flujo que incluya pasos de transformación de datos y máquinamodelos de aprendizaje.Estas secuencias de pasos se denominan canalizaciones de datos o canalizaciones de ML.Esta sección examina una característica de Watson Studio que ayuda a automatizar la creaciónde canalizaciones de aprendizaje automático.Esto permite que los científicos de datos produzcan resultados mucho más rápido y se centren en proyectos más creativos.trabaja.Actualmente hay escasez de científicos de datos calificados.Muchas operaciones que normalmente realiza un científico de datos son repetitivas y requieren mucho tiempo.Por lo tanto, automatizar parte de ese trabajo repetitivo ayudará a liberar tanto a los nuevos como a los experimentados.científicos de datos para hacer el importante trabajo para el que están capacitados.El sistema AutoAI fue desarrollado por expertos de IBM Research en colaboración con IBM DistinguishedIngeniero y dos veces Gran Maestro de Kaggle Jean-Francois Puget.Proporciona una interfaz gráfica para crear e implementar modelos de aprendizaje automático convisualizaciones de tiempo.AutoAI realiza automáticamente los pasos típicos de aprendizaje automático, como:Preparación de datos Selección de modelosIngeniería de funciones Optimización de hiperparámetrosLos usuarios pueden ver el progreso en la interfaz gráfica.Este ejemplo muestra el entrenamiento de un modelo para predecir si un cliente es probable o nocomprar una tienda de campaña en una tienda de equipos para actividades al aire libre.Empezamos con los datos estructurados.En estos datos históricos, hay cuatro columnas de características o "predictores":GÉNERO: El sexo del cliente EDAD: La edad del clienteMARITAL\_STATUS: "Casado", "Soltero" o "Sin especificar"y PROFESIÓN: La categoría general de la profesión del cliente.profesión, como “Hostelería” o “Ventas”, o simplemente “Otro”.El modelo aprenderá a predecir el valor de la columna IS\_TENT; es decir, si ono el cliente compró una tienda de campaña.Después de elegir IS\_TENT como la columna para predecir, AutoAI analiza los datos y determina quela columna IS\_TENT contiene información Verdadero/Falso, lo que hace que estos datos sean adecuados para una clasificación binariamodelo.La métrica predeterminada para una clasificación binaria es ROC/AUC.Después de hacer clic en Ejecutar experimento, una infografía muestra el proceso de construcción de las canalizaciones.mientras el modelo entrena.Una vez que se completa la creación de la canalización, podemos ver y comparar las canalizaciones clasificadasen una tabla de clasificación.Las canalizaciones para el modelo de clasificación binaria de muestra son bastante uniformes debido a laData de muestra.Para ver las canalizaciones en acción, vuelva a ejecutar el experimento como un experimento de regresión para predecir la compraMonto.Ese experimento da una mejor variación en las canalizaciones resultantes.Después de hacer clic en "Comparación de canalizaciones", podemos ver cómo difieren las canalizaciones en variosMedidas de la calidad del modelo.Las canalizaciones se pueden guardar como recursos de aprendizaje automático en el proyecto de Watson Studio.Luego se pueden implementar y probar.Actualmente, AutoAI está disponible solo para modelos de clasificación y regresión; hay un plan paraagregar compatibilidad con modelos de series temporales en el futuro.En esta unidad, aprendió cómo AutoAI automatiza tareas típicas de ciencia de datos y ayudaobtenga canalizaciones de datos de mejor rendimiento más rápidamente, al mismo tiempo que simplifica la implementación de la canalizaciónen producción en Watson Machine Learning.En la siguiente sección, hablaremos de Watson OpenScale, que ayuda a garantizar que sulos modelos son justos, explicables y actualizados.

Sem 3.17: IBM Watson Open Scale

Este video examina las caracter�sticas Watson Studio el cual ayudan a asegurar la legitimidad y la explicabilidadde las tuber�as de datos del aprendizaje autom�tico, as� como hacer un seguimiento a su desempe�o despu�s del despliegue.IBM Watson OpenScale es un producto que incluye muchas caracter�sticas importantes.Puede probar el modelo al igual que sus predicciones para la legitimidad y aplicar formas para superar las tendencias.Tambi�n puede ayudar a proveer explicaciones acerca de las predicciones que generalmente son dif�ciles de obtenerpero que son necesarias para el cumplimiento en algunas �reas de aplicaci�n.Hace el seguimiento del funcionamiento del modelo y puede detectar su alteraci�n o "model drift",a lo largo del tiempo. Puede advertir a los usuarios cuando la drift es detectada y explica qu� predictores la est�ncausando. Podemos especificar criterios bajo los modelos que se capacitan una vez m�s autom�ticamentecon los datos nuevos. Esto tambi�n ayuda a medir c�mo el modelo ayudaal negocio.Los atributos para controlar las tendencias se recomiendan autom�ticamente en base a la experiencia previa. Ellaspueden ser editadas seg�n sea necesario. OpenScale entonces hace el seguimiento de las prediccionesdel modelo para los grupos especificados y verifica las tendencias enlas predicciones.Los aseguradores pueden usarel aprendizaje autom�ticoy OpenScalepara evaluar de manera consistente y precisa el riesgo de las reclamaciones, asegurar los resultados justos para los clientesy explicar las recomendaciones de la Inteligencia Artificial (AI) para los prop�sitos de regulaci�n e inteligencia empresarial.En este video, has aprendido c�mo OpenScale asegura la legitimidad yla explicabilidad de los modelos y hace el seguimiento delmodelo de operaciónen producci�n.Esto completa el m�dulo de los productos de IBM para los cient�ficos de datos. �Buena suerte en las pruebas!

Lectura: Insignia Digital de IBM

Obtener el certificado de Coursera para este curso le dará derecho a recibir una Insignia Digital de IBM sin ningún cargo adicional. (Nota: se requiere el pago para tener acceso completo al curso y obtener el certificado). Para este curso obtendrá la insignia de Herramientas de Código Abierto para Data Science. Los detalles completos se pueden ver aquí.

https://www.youracclaim.com/org/ibm/badge/open-source-tools-for-data-science

Las insignias digitales de IBM son una credencial en línea que valida las habilidades que adquirió al pasar este curso. Puede compartirlas en sitios populares de medios sociales, como Linked-In, Twitter o Facebook. Cada una tiene una URL única que puedes incrustar en un sitio web, correo electrónico o CV, por lo que no podría ser más fácil compartir sus insignias y sus logros. Las insignias digitales de IBM se adhieren al estándar global de insignias abiertas gestionado por la IMS Global Learning Consortium, así que también puedes compartirlas con cualquier sitio compatible con OBS, como Mozilla Backpack. IBM se ha asociado con Credly Acclaim para emitir y gestionar las insignias digitales de IBM. Si usted pasa este curso y obtiene el certificado de Coursera, se le proporcionarán las instrucciones para aceptar y reclamar la suya.