Práctica Final
Big Data AI Machine Learning
Bootcamp VI Edición

1. Descripción de la práctica

El objetivo de este proyecto final es poner en práctica todos los conocimientos adquiridos a lo largo del bootcamp, simulando una situación real: desarrollo en equipo de un producto con la arquitectura que hemos estudiado. El desarrollo se hará usando SCRUM, teniendo que coordinar no sólo vuestro propio trabajo y tiempos, sino también el del equipo del cual formáis parte.

Hoy os quitaremos los ruedines de la bicicleta y os toca pedalear solos.

Hay que vivirlo como si fuera un proyecto real, solicitado por un cliente, con un plazo muy ajustado (99% de los casos :-)), y donde se debe entregar el mejor y más completo prototipo posible.

Habrá que investigar, tomar decisiones, estudiar y aprender cosas nuevas, tal como en la vida real. ¿Lo fundamental? Aprender, consolidar conocimientos y disfrutar, ya que probablemente será la última vez que podrás enfrentarte a un desarrollo de proyecto con la posibilidad de acudir a tutorías para aclarar dudas.

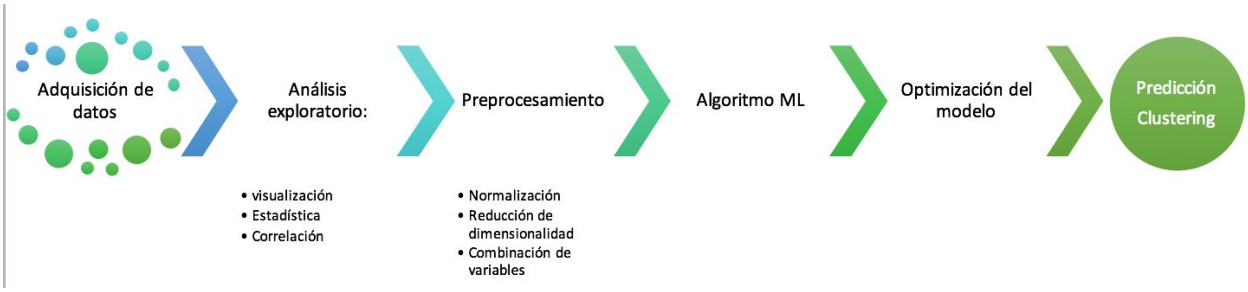
No se espera que terminéis el proyecto completo, aunque se valorará aquellos equipos que avancen más que los demás. Por ello el proyecto se abordará como un backlog de historias de usuario que tendréis que ir implementando. Eso permite que en cada etapa se toquen todos los palos. Con ese mismo objetivo, se recomienda que, tras cada sprint, los roles en equipo se vayan rotando, para que cada miembro practique todas las tecnologías.

El proyecto final consistirá en una serie de etapas que deberán ser cumplimentadas por cada grupo.

Es un proyecto que simula una situación real. Es importante tanto el desarrollo como el resultado final que obtendréis.

Es realmente importante dedicar algún tiempo al inicio de primer Sprint para definir el plan de negocio que se va a aplicar. Somos conscientes de que os pedimos un proyecto real en un corto espacio de tiempo, por lo que, una versión 1.0 se dará como válida, ¡aún así esperamos que nos sorprendáis!!

Pipeline del proyecto



1. **Definir Data Set**
    a. Elegir un dataset principal de entre las temáticas que se proponen en punto 'Selección del Dataset'.
    b. Cruzarlo, al menos, con otra fuente de datos.
    c. Si es pertinente, cruzarlo con datos de streaming, por ejemplo, redes sociales.

2. **Arquitectura y validación de los datos** (Arquitecto y Desarrollador Big Data)
    a. Muestreo y exploración inicial de los datos
    b. Definir e implementar la arquitectura
    c. Ingesta de datos
    d. Validación de la calidad de los datos (GO TO 2.c)
    e. Etc.

3. **Análisis Exploratorio** (iterativo). (Visualización de resultados de los análisis, recomendado con Tableau)(Data Scientist / Citizen Data Scientist)
    a. Detección outliers (rango de variables), imputación valores nulos.
    b. Boxplots, histogramas, etc.
    c. etc.

4. **Pre Procesamiento** (Data Scientist / Data Engineer)
    a. Normalización o escalado
    b. Reducción de la dimensionalidad (si fuera necesario)
    c. Combinación de variables
    d. Etc.

5. **Modelado** (Elegir/combinar algoritmos de data mining, machine learning, deep learning y/o NLP) (Data Scientist / Machine Learning Engineer)
    a. Definir modelo en función de tarea de aprendizaje a resolver (clasificación/regresión/agrupamiento)
    b. Definir/optimizar parámetros libres del modelo
        i. Entrenamiento / validación / test
    c. Evaluar el modelo (si resultados no satisfactorios GO TO 4)

6. **Informe** (Data Scientist/Desarrollador Big Data)

2. Selección del dataset

El grupo deberá seleccionar un dataset entre los que siguen:

En caso de querer seleccionar otro dataset, dicha elección tiene que ser aprobada por la coordinación del bootcamp.

1. Elegir un dataset principal de entre las siguientes temáticas:

    - Redes sociales
    - IMDB, Wikipedia
    - The Pirate Bay (como dataset estático) y los datos actuales de descargas como añadido real time. Posiblemente cruzado con datos de lanzamientos de películas (sacados de IMDB).
    - Cartográficos: open street maps
    - Seguridad Social, atención médica, epidemiología
        - Unidad Cuidados Intensivos, en MIT.
    - INE
    - Bioinformatics:
        - https://www.ensembl.org/index.html
        - http://www.ensembl.org/info/data/ftp/index.html
    - Instituciones públicas:
        - http://datos.madrid.es/portal/site/egob/
        - http://datos.gob.es
    - El blockchain de bitcoin: datos públicos que se pueden cruzar con noticias en redes sociales. Buscar patrones.
    - http://www.sepln.org/workshops/tass/2015/tass2015.php#corpus Corpus taggeados del 2015 para ciertas tareas de NLP.
    - Twitter.
    - https://www.bigdatanews.datasciencecentral.com/profiles/blogs/big-data-50-fascinating-and-free-data-sources-for-data
    - Clasificación binaria: Variedad de datos alternativos, que incluyen información de transacciones y de telecomunicaciones, para predecir las capacidades de pago de sus clientes (múltiples ficheros asociados a cada observación y trabajar mucho en la parte de “feature engineering”, competición abierta)
    - https://www.kaggle.com/c/home-credit-default-risk/data
    - Fuente: https://www.kaggle.com/datasets

2. Cruzarlo, al menos, con otra fuente de datos.
3. Si es pertinente, cruzarlo con datos de streaming, por ejemplo, redes sociales.

3. Arquitectura, almacenamiento e ingesta

Una vez seleccionado el dataset y teniendo en cuenta sus particularidades, habrá que:

- Definir y justificar la arquitectura elegida
- Crear un endpoint en Python o Scala que recibe los datos
- Comprobar que la ingesta se hace de forma correcta.
- Describir el stack creado (incluyendo el endpoint y el dataset) en un fichero Vagrant, Docker o Ansible playbook para poder reproducirlo
- Desplegar en la nube
- Subir el fichero en cuestión al repositorio del grupo en gitlab.keepcoding.io

4. ETL y cálculo de métricas con Spark

En esta etapa, habrá que definir y calcular las métricas según el dataset elegido.

- Crear un notebook Jupyter para llevar a cabo un análisis exploratorio de los datos.
- Hacer un estudio estadístico con R o Python y averiguar cuales son las métricas adecuadas para el dataset.
- Subir el notebook Jupyter con el estudio estadístico y sus conclusiones a gitlab.keepcoding.io
- Llevar a cabo el ETL y el cálculo de las métricas con Spark en Python o Scala.
- Actualizar el fichero vagrant, docker o ansible con Spark y los scripts de ETL y cálculo de métricas y subir a gitlab.keepcoding.io

5. Visualización de las métricas

- Crear un dashboard con Tableau dirigido a un objetivo que dependerá del data set. El dashboard debe de tener un sentido de negocio.
- Crear un report con D3 que cubra alguna necesidad importante diferente del dashboard global.

6. Resolución de problemas de DM/ML/DL/NLP

Cada dataset vendrá con al menos dos problemáticas de Data Mining, Machine Learning, Deep Learning y/o NLP asociadas (a ser posible usando técnicas diferentes).

En caso de utilizar técnicas de Machine Learning, un estudio interesante podría ser resolver dicho problema de dos formas. Primero usando técnicas más tradicionales de ML. Una vez hecho esto, se deberá de resolver el mismo problema mediante Deep Learning y se compararán los resultados.

- Crear un notebook para la resolución
- Definir el problema de ML a ser resuelto
- Resolución mediante técnicas tradicionales de ML: Rectas de regresión, Árboles de decisión,algoritmos de clasificación, etc.
- Resolución mediante Deep Learning
- Comparar los resultados e indicar cuál ha sido el método óptimo, teniendo en cuenta los resultados de cada y el coste computacional y de tiempo de cada uno.
- Actualizar el fichero que describe la arquitectura y subir a gitlab.keepcoding.io
- Subir el notebook a gitlab.keepcoding.io

7. Presentación de los resultados

En esta etapa se debe de simular la presentación de resultados en un entorno real de empresa.

Crear un notebook y presentar en él los siguientes puntos:

- Suposiciones iniciales
    - Cuales han demostrado ser válidas y cuáles no. ¿Por qué?
- Métricas seleccionadas: ¿han sido las correctas o no? ¿por qué?
- Arquitectura elegida: ¿ha sido la correcta o no? ¿por qué?
- Métodos de ML utilizados, ¿cuales han sido los mejores?
- Teniendo en cuenta lo aprendido ¿Qué cosas se harían igual y cuales se harían de otra forma? ¿Por qué?
- Información obtenida del data set.
- Conclusiones y “lessons learned”

8. Logística y Trabajo en Equipo

- Los alumnos pasarán las historias a un Scrum Board en Trello.
- Los equipos se coordinarán en una sesión de planificación para cada Sprint.
- Se seguirá el progreso de cada uno usando Trello y si es posible haciendo Scrum Meetings diarios del grupo.
- Se mantendrá actualizado un Burndown Chart.
- Para que todos toquen todos los palos se hará rotación de roles por Sprint.

9. Formación de Grupos

Todos los alumnos pueden participar en la práctica final.

Habrá 3 castas según su rendimiento pasado:

Guardia Imperial

Han entregado y aprobado mínimo 7 prácticas con calificación de APTO.

Tropa de Asalto

Han entregado y aprobado mínimo de 5 a 6 prácticas con calificación de APTO.

Milicia Gungan

Han entregado y aprobado menos de 5 prácticas con calificación de APTO.

- Las castas NO se mezclan: únicamente se formará grupo con componentes de una misma casta.

- Trabajaremos con grupos de hasta 5 personas (máx 6 bajo aprobación de KeepCoding): es el formato ideal para poder completar el desarrollo en los tiempos establecidos, y poder trabajar los conceptos fundamentales de gestión de proyectos ágiles.

- Los alumnos tendrán libertad para decidir la formación en grupos, pudiendo solicitar a KeepCoding su asignación. Una vez solicitado, se deben respetar los criterios de agrupación decididos por KeepCoding Team.

- Se permite el desarrollo global del proyecto por una sóla persona, aunque no se recomienda en absoluto: la relación volumen desarrollo / tiempo disponible es muy alta y, no permite aplicación de gestión de proyectos ágiles con SCRUM.

10. Evaluación, Presentación Final & Nota

Presentación Proyectos Grupales ante Instructores

En esta fecha se realizarán las presentaciones de los proyectos finales al grupo de Instructores. La presentación será realizada a distancia y por la tarde/noche, tal como las clases, de manera que el alumno no tendrá que hacer ningún desplazamiento para esta finalidad.

Cada grupo presentará su aplicación al grupo de instructores. Cada grupo contará con 15 minutos para presentar su App, Canvas Business Plan & Pitch, y otros 15 min para contestar a las preguntas de los profesores.

Cada grupo deberá seleccionar una historia de usuario y presentar ese código a los instructores explicando las decisiones de implementación que se han tomado y respondiendo a las preguntas que puedan hacer los instructores.

Tras las sesiones de presentación, los profesores se reunirán y darán una nota individual que será la final del Máster, teniendo en cuenta siempre las prácticas entregadas de cada asignatura y la práctica final.. Las calificaciones se darán como APTO - NO APTO + nota final del proyecto en cada área.

Los mejores proyectos saldrán publicados junto con identificación de sus autores y menciones respecto a su gran trabajo en nuestro blog y diferentes publicaciones que hagamos en parte final de Big Data Machine Learning Bootcamp - V Edición.

En la Práctica Final también se tendrá en cuenta:

- Calidad e implementación del backlog.
- Calidad del código presentado.
- Estabilidad del prototipo.
- Viabilidad económica del proyecto y calidad y claridad del “pitch”.
- **Closing Final: Presentación Proyectos ante compañeros, instructores y empresas. Pendiente confirmar fecha definitiva y enfocada totalmente a vuestra puesta en contacto con recruiters + celebración con vuestros compañeros de edición.**

Este acto lo viviremos como una gran fiesta donde celebraremos con el resto de compañeros e instructores de V Edición Big Data Machine Learning Bootcamp los excelentes objetivos alcanzados. Se trata de presentar en un muy breve plazo de tiempo (8 min) todo nuestro trabajo, compartiendo con todos los compañeros e instructores el trabajo realizado en esta Fase Final del Big Data Bootcamp. Muy importante preparar un Pitch original que nos haga sentir curiosidad por lo que hacéis. Tenemos que crear buen proyecto y además es necesario saber venderlo.

Combinaremos nuestra presentación de Proyecto Final y correspondiente Pitch de plan de negocio en un tiempo máximo de 8 min. Teniendo en cuenta que, normalmente en la vida real será de 1-2 min el tiempo disponible para hacer estas presentaciones a nuestros futuros clientes, debéis ser estrictos para que este intervalo no sea excedido bajo ningún concepto.

11. Certificados

Para obtener el certificado oficial del “Big Data Machine Learning Bootcamp - VI Edición”, es necesario haber realizado la Práctica Final y obtener una calificación final de APTO en la evaluación conjunta de prácticas entregadas y practica final.

Si el alumno no participa de la práctica final, u obtiene un NO APTO, recibirá un certificado por la realización de los módulos para los cuales haya entregado su correspondiente práctica, y obtenido un APTO.

Nuestro propósito desde KeepCoding es ayudaros a VOLAR, con vuelo firme y seguro