MÁSTER EN DATA SCIENCE

CAPSTONE: PLAN DE PROYECTO

Diseño de un modelo de *Machine Learning* para la predicción de pujas en entorno de Publicidad programática

Alumnos:
Javier Alejo Álvarez
Marco Duque García
Marta Pérez Romero

Mentora: Elena Abril Medina

Edición 2021-22 (Octubre/Abril)

IMMUNE
TECHNOLOGY INSTITUTE

INDICE

PROPUESTA DE TÍTULO

Como primer apartado del plan de proyecto del Capstone, se definirá su título respetando los requisitos: completo, claro, preciso y referido al tema principal.

De este modo, el título, que refleja el contenido del proyecto será:

"Diseño de un modelo de *Machine Learning* para la predicción de pujas en entorno de Publicidad programática".

PALABRAS CLAVE

Con las siguientes palabras clave es posible resumir el trabajo:

- Machine Learning.
- Modelos predictivos.
- Inteligencia artificial.
- Publicidad programática.
- Despliegue en AWS.

RESUMEN DE LA PROPUESTA

Hoy en día, la Ciencia de Datos es algo que forma parte de nuestra vida cotidiana. Aspectos como Inteligencia Artificial, modelos de entrenamiento, creación de *chatbots*, etc. ya no resultan tan nuevos, pese a que todavía tienen por delante un enorme potencial de crecimiento, en cantidad y calidad.

El *data mining*, concepto que existe desde hace décadas, ingeniería de datos, Big Data, *Business Intelligence*, etc., son términos cada vez más comunes y alrededor de ellos hemos experimentado un cambio en el paradigma empresarial con perspectivas como DevOps o MLOps. Todo ello, fuertemente relacionado con el desarrollo en la nube, con *Amazon Web Services, Microsoft Azure o Google Cloud Platform* como abanderados.

El *Machine Learning* puede aplicarse a distintos campos y con muy diversos propósitos. En este capstone se hará uso de ello aplicado al mundo de la publicidad por internet.

En este punto entra en juego la segunda parte de la propuesta. Desde los inicios de la publicidad online, con el modelo de pago por clic (CPC) de Espotting, MIVA, Overture (comprada posteriormente por Yahoo!), Yahoo!, Google o Microsoft, las formas de publicidad por internet han ido aumentando, refinándose, segregando el mercado en distintas soluciones. Una de ellas es la publicidad programática. Como sabemos, se basa en la existencia de anunciantes y *publishers*, los primeros desean publicitar su producto o servicio, mientras que los *publishers* (que suelen elegir qué anuncios quieren, con qué precio y qué anunciantes aceptan o no) venden su espacio publicitario.

En este contexto, las empresas de publicidad online buscan la manera de optimizar sus sistemas para obtener la mayor rentabilidad. En concreto, se contará con la ayuda de la empresa Kimia (https://kimiagroup.com/es/index.html), que facilitará un *dataset* con datos reales (ninguno de ellos sensible ni personal), a partir del cual se podrá realizar la aplicación de *Machine Learning* a este entorno, que dé respuesta a una problemática real, llegando hasta la fase final de despliegue. La problemática consiste en que la empresa envía a su red de *publishers* los anuncios disponibles. Las redes o *publishers* valoran si les interesa entrar en una subasta y pujar por esos anuncios o no. Esta decisión depende de factores como el tipo de anuncio, su segmento, el precio del que parte, etc. La empresa Kimia tiene un análisis de qué redes o *publishers* son más rentables y adecuados para el negocio (otorgan tráfico de calidad). De este modo, lo ideal es que la empresa envíe los anuncios únicamente a aquellas redes que tengan más probabilidad de aceptar ese anuncio y den mayor rentabilidad, en lugar de enviarlo a todas las redes y encontrarse posteriormente con tráfico de mala calidad y con clics fraudulentos.

JUSTIFICACIÓN

La justificación principal de este proyecto es, por un lado y tras una experiencia profesional de más de 15 años en negocios digitales y publicidad online, entender cuáles son las actuales tendencias en este campo. Por su parte, la propuesta se justifica también por el deseo de adquirir un conocimiento sustancial de los distintos modelos predictivos que existen en *Machine Learning* e Inteligencia Artificial. Con ello, se diseñará un modelo

que, dado un *dataset* con datos reales de la industria, optimice la gestión de redes que son candidatas para pujar por un determinado contenido publicitario.

OBJETIVOS

Partiendo de una motivación personal por la Inteligencia Artificial y, en concreto, por su aplicación a campos como la publicidad programática, se establecen los siguientes puntos como los auténticos objetivos marcados a nivel de grupo:

- Utilizar un *dataset* real, con datos verídicos.
- Ser capaz de construir las variables de entrada para el modelo, que serán vectoriales (cada elemento de esos vectores es una variable categórica que se deberá convertir en numérica con los diccionarios.
- Saber cómo balancear un *dataset*.
- Ejercitarse en las tareas de limpieza, exploración y visualización de los datos con Python y sus librerías.
- Abordar el estudio de posibles algoritmos y modelos, y seleccionar y aplicar el más adecuado.
- Saber evaluar y hacer test de un modelo para comprobar la validez de las predicciones.
- Aprender a crear y utilizar la máquina en AWS, y hacer el despliegue de la aplicación.

ESTRUCTURA DE LA MEMORIA

A continuación, se ofrece un resumen de lo que se expondrá en cada uno de los capítulos de la memoria, aunque este aspecto se encuentre sujeto a posibles modificaciones:

- **Capítulo 1**: El primer bloque del proyecto se dedicará al Plan de Trabajo, a una introducción y presentación de este, incluyendo un resumen de la propuesta, el

marco de trabajo, las contribuciones, la metodología que se va a seguir y una planificación temporal de las distintas tareas requeridas.

- **Capítulo 2**: En el segundo bloque se hablará del estado del arte tanto en *Machine Learning* como en publicidad online. Se repasarán los principales modelos y tendencias del aprendizaje automático, lenguajes que se pueden utilizar, etc. Por su parte, se realizará un estudio completo de las actuales formas de publicidad por internet, incluyendo la evolución del modelo CPC (Pago por Clic) a la publicidad programática.
- **Capítulo 3**: El tercer bloque incluye el trabajo sobre el *dataset*, es decir, definición de variables, limpieza, exploración, visualización, con diferentes librerías de Python. Investigar y seleccionar el modelo de *Machine Learning* más adecuado, junto a la fase de prueba.
- **Capítulo 4**: En el cuarto bloque se hará un estudio comparativo de las distintas soluciones de despliegue del proyecto, analizando ventajas e inconvenientes de utilizar *AWS, Azure* o *GCP*. Se procederá, al final de este análisis, a la implementación del modelo en la nube. Se finalizará con un análisis de los resultados.
- **Capítulo 5**: El último bloque cierra el capstone con las conclusiones extraídas de todo el trabajo realizado, incluyendo reflexiones personales y un detalle de futuras líneas de trabajo.
- **Anexos**: Se incluirá en esta sección todo el código comentado, detalles sobre la comparativa de modelos de *Machine Learning* que se pueden aplicar, y un presupuesto de puesta en marcha y despliegue del modelo en un entorno real.

METODOLOGÍA

Como se puede extraer de todo lo comentado hasta la ahora, la metodología será una mezcla de teoría, que posteriormente se aplicará a la práctica.

En concreto, hay tres puntos de atención: modelos de *Machine Learning*, publicidad programática y despliegue en la nube.

De forma paralela, pero no simultánea, se estudiará el estado del arte de cada componente citado, y con ello, se tomarán decisiones sobre las soluciones que se adoptarán.

Tras este ejercicio teórico se trabajará en la aplicación práctica en los tres núcleos de acción:

- *Machine Learning*: recepción del *dataset*, análisis de este, aplicación de limpieza, exploración y visualización, aplicación del modelo elegido, entrenamiento y *testing*.
- Publicidad programática: sin entrar a fondo en ello, dado que se encuentra fuera del alcance del proyecto, se abordará el análisis de rentabilidad de las redes y, en conjunto, el modo de funcionamiento de redes-anuncios-subastas.
- Despliegue en la nube: decidida la plataforma en la que hacer el despliegue, y tras un estudio teórico práctico de su funcionamiento, se abrirá una cuenta de prueba y se seguirá el proceso adecuado para implementar el modelo y ponerlo en funcionamiento.

  Como último paso, habrá que evaluar los resultados y explicarlos con claridad en la sección de conclusiones y próximas líneas de acción.

RIESGOS Y PLAN DE CONTINGENCIA

Es de esperar que a lo largo del *capstone* puedan surgir eventualidades no previstas. Estas incidencias se pueden clasificar en dos grupos, y para cada una de ellas se añade su plan de contingencia:

a) Incidencias técnicas

- Problemas con los recursos de *hardware*: es posible que alguno de los equipos no responda como se espera. Por ello se cuenta con tres ordenadores independientes. En caso de que los tres se estropeen, se dispone de soluciones alternativas que incluyen el uso de terminales de familiares.

- Incompatibilidad de los recursos de *software*: se puede dar el caso de que no se consiga desarrollar el proyecto completo sobre una única plataforma por cuestiones de incompatibilidad. Es por ello por lo que se dispone de los tres sistemas operativos principales, Windows, OSX y Linux, de modo que todos los programas podrán funcionar en uno u otro.
- Pérdida de datos: ante la posibilidad de que se pierdan datos, algo no deseable, se hará una copia de seguridad diaria de la carpeta de trabajo, que se guardará tanto en una localización en la nube como en un disco duro portátil y en un dispositivo USB. Se garantiza la sincronización de los datos.
- Dificultades en la utilización del *software* especializado: al ser varias librerías será necesario superar una cierta curva de aprendizaje. Para ello se han previsto unas horas de lectura de documentación que ayudarán a avanzar más rápido cuando se trabaje con ellas.

b) Incidencias personales

- Simultaneidad con el trabajo: para evitar que ambos puedan interferir en el desarrollo del *capstone* se ha aprovechado el tiempo previo al inicio oficial de este para comenzar la planificación y lectura de bibliografía, y se han establecido sesiones diarias de trabajo de 3 horas, lo cual permite simultanear el proyecto con otras actividades.
- Dificultades en la realización: este proyecto supone un reto, puesto que gran cantidad de los contenidos son nuevos. Es posible que haya puntos a los que se deba dedicar más tiempo del previsto. Para que ello no suponga un problema existe un margen de un 10% de horas que se podrán utilizar para recuperar tiempo en caso de que sea necesario.
- Enfermedad o indisposición: este 10% extra de horas contemplado en el plan será también de utilidad en caso de que alguna enfermedad obligue a replantear los horarios.
- El plan de contingencia principal para evitar cualquier riesgo se basa en hacer un seguimiento continuado del plan de proyecto. A través de ese ejercicio se puede comprobar día a día el grado de cumplimiento y, de ser necesario, hacer algún tipo de restructuración o aplicar alguna medida correctora.

ANÁLISIS DAFO

Como último ejercicio previo al desarrollo del proyecto y, más específicamente, a su planificación, conviene reflexionar sobre cuáles son los puntos fuertes y débiles, tanto internos como externos, lo cual se ha resumido en una tabla a modo de análisis DAFO:

ANALISIS INTERNO	
Fortalezas	**Debilidades**
Interés y motivación plenos para aprender sobre *Machine Learning* y nuevas tendencias de publicidad online. Experiencia profesional previa en dirección de proyectos técnicos Experiencia profesional previa y actual en publicidad online. Perseverancia Buen margen de maniobra para gestionar imprevistos	Desconocimiento de algunas librerías de *Machine Learning.* Conocimiento únicamente a nivel teórico de servicios en la nuble, como *Amazon Web Services*.
ANALISIS EXTERNO	
Oportunidades	**Amenazas**
Riqueza de recursos en Internet, tanto teóricos como prácticos Disposición de la mentora para hacer seguimiento y servir de guía Posibilidad de hacer diversos modelos de ML antes de seleccionar uno.	Riesgo de necesitar más tiempo del estimado Riesgo de enfermedad durante el proyecto Imposibilidad de prever fallos técnicos de los equipos

PLANIFICACIÓN

Calendario de trabajo

El *capstone* se desarrollará durante y después del *Máster en Data Science* (edición 2021-2022), cuyas fechas oficiales van desde el 23/02/2022 (inicio) hasta la fecha de presentación, estimada a efectos del desarrollo del trabajo en la segunda semana de junio (fecha sujeta a modificación una vez confirmada la oficial).

No obstante, puesto que durante el curso se está trabajando por cuenta ajena (y propia), se ha optado por comenzar la realización del proyecto unas semanas antes. De este modo, se dispone en realidad de aproximadamente tres meses y medio.

Para el correcto seguimiento del *capstone*, se ha dividido en cuatro bloques fundamentales: Bloque I (introducción, con la contextualización del proyecto y el plan de

trabajo), Bloque II (estado del arte y trabajo con el *dataset*) y Bloque III (soluciones de despliegue, puesta en producción y pruebas funcionales), además de conclusiones y líneas futuras de investigación. Cada uno de los bloques se irá trabajando de acuerdo con un calendario.

Visualmente, este calendario se estructurará de la siguiente manera:

Por su parte, se ha estimado una carga de trabajo de unas 125 horas.

Teniendo en cuenta lo anterior, la disponibilidad diaria será de 2 horas, tal y como queda reflejado en la tabla inferior. El número de días se ha calculado desde el 21/03/22 hasta el 12/06/22. La suma total corresponde a la suma de los días por el número de horas.

No se trabajará en días de Semana Santa (marcados en rojo) ni los viernes y sábados en los que tengamos clases del máster.

Como se puede apreciar, las horas disponibles superan a las horas calculadas. Esto permite un margen de acción suficiente en caso de que otros motivos obliguen a modificar el calendario de dedicación.

Día Semana	Nº Días	Nº Horas	Horas totales
Lunes	11	2	22
Martes	11	2	22
Miércoles	11	2	22
Jueves	11	2	22
Viernes	8	2	16
Sábado	7	2	14
Domingo	9	2	18
		Total	**136 horas**

Hitos

Los hitos corresponden a las fechas clave de finalización de cada uno de los bloques y *feedback* correspondiente por parte de la mentora. Se añaden también como hitos la realización de la pre propuesta y la fecha de presentación ante el Tribunal:

Grupo	Hito	Fecha
Bloque I	Llamada inicial con mentora	02/02/2022
	Aprobación del proyecto	23/02/2022
	Plan de proyecto	19/03/2022
Bloque II	Obtención del dataset	21/03/2022
	Validación del modelo	28/04/2022
Bloque III	Despliegue de la solución	28/05/2022
Presentación	Preparación del pptx	01/06/2022
	Evaluación del Tribunal	22/06/2022

Planificación de tareas

Considerando los contenidos del proyecto, el calendario, las horas de trabajo disponibles y las fechas de entrega, se ha desglosado el desarrollo del trabajo en las tareas y actividades que pueden consultarse en las tablas siguientes. Se muestran las fechas de realización, y su duración estimada en horas.

Igualmente, se incluye a nivel global la carga de trabajo por cada una de las tareas, en función del tiempo requerido. Es conveniente indicar que se han distribuido las tareas de manera coherente, y que los porcentajes cuadran a nivel global con la temporización general y con las fechas de entrega.

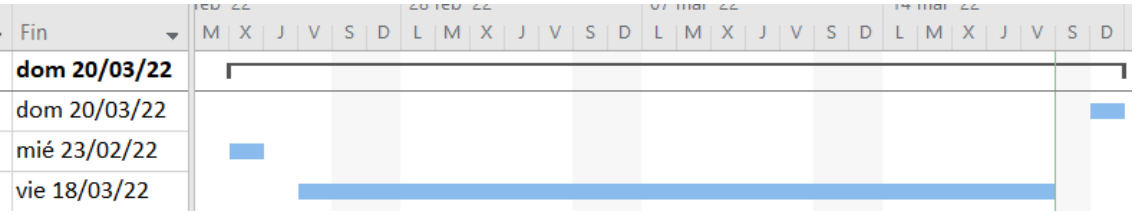
Por último, este plan estima unas 125 horas de trabajo, algo que es posible teniendo en cuenta que en el calendario de disponibilidad tenemos un total de 136 horas. En la sección de “

Incidencias personales” veremos que se han reservado deliberadamente para disponer de un margen de tiempo en caso de ser necesario.

En las próximas páginas podemos encontrar la lista completa de tareas en el calendario. Conviene destacar que seguiremos un modelo de gestión de proyectos en cascada, aunque determinadas tareas de diferentes bloques se solapen en el tiempo.

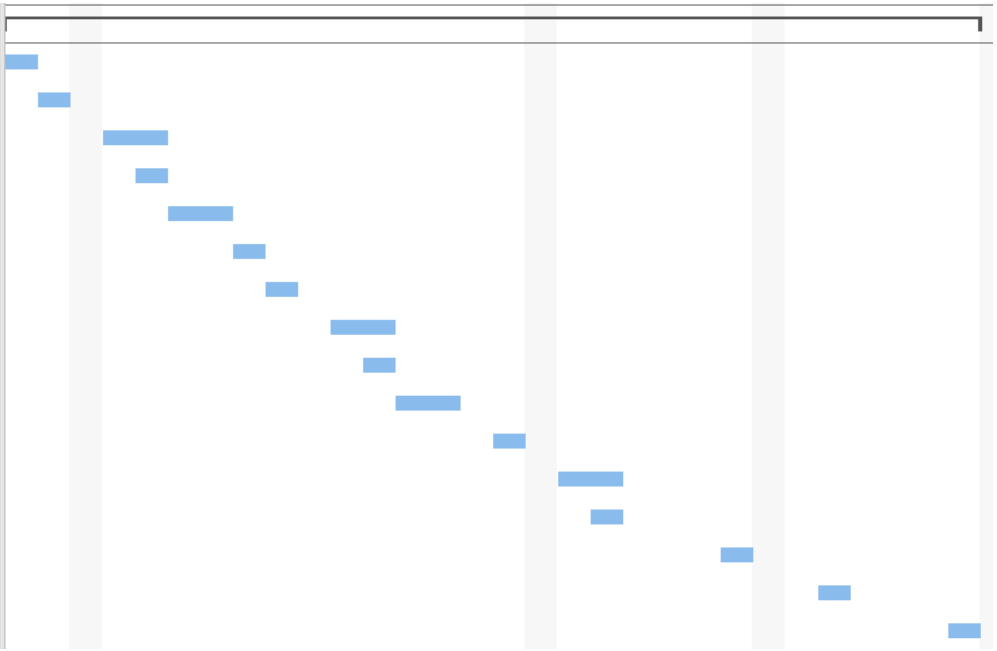
Nombre de tarea	Duración	Comienzo	Fin
Bloque I	**19 días**	**mié 23/02/22**	**dom 20/03/22**
Bloque II	**31 días**	**lun 21/03/22**	**vie 29/04/22**
Bloque III	**27 días?**	**vie 29/04/22**	**sáb 28/05/22**
Presentación	**28 días**	**jue 19/05/22**	**mié 22/06/22**

Nombre de tarea	Duración	Comienzo	Fin
Bloque I	**19 días**	**mié 23/02/22**	**dom 20/03/22**
Llamada con mentora	1 día	dom 20/03/22	dom 20/03/22
Aprobación del proyecto	1 día	mié 23/02/22	mié 23/02/22
Plan de proyecto	16 días	vie 25/02/22	vie 18/03/22

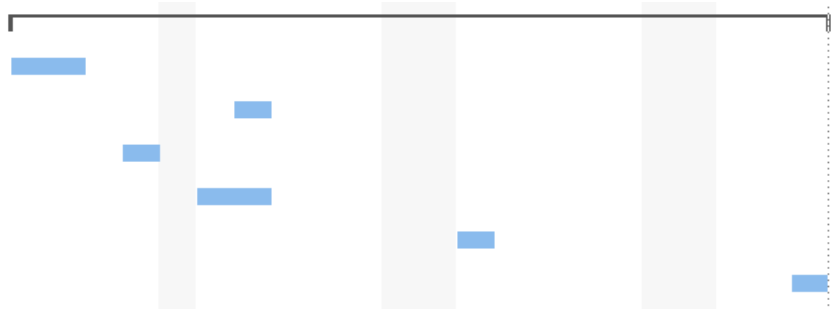


Bloque II	**31 días**	**lun 21/03/22**	**vie 29/04/22**
Creación Github de equipo	1 día	lun 21/03/22	lun 21/03/22
Obtención del dataset	1 día	lun 21/03/22	lun 21/03/22
Definición de la variable	1 día	mar 22/03/22	mar 22/03/22
Limpieza del dataset	2 días	mié 23/03/22	jue 24/03/22
Análisis exploratorio	2 días	vie 25/03/22	lun 28/03/22
Análisis visual	2 días	mar 29/03/22	mié 30/03/22
Pre selección de modelos	1 día	dom 10/04/22	dom 10/04/22
Análisis de modelos	2 días	lun 11/04/22	mar 12/04/22
Definición del modelo final	2 días	mié 13/04/22	jue 14/04/22
Entrenamiento y test	3 días	lun 25/04/22	mié 27/04/22
Llamada con mentora	1 día	mar 26/04/22	mar 26/04/22
Validación	2 días	mié 27/04/22	jue 28/04/22

Bloque III	27 días?	vie 29/04/22	sáb 28/05/22
Desarrollo dashboard Streamlit	1 día	vie 29/04/22	vie 29/04/22
Llamada grupo	1 día?	sáb 30/04/22	sáb 30/04/22
Estudio de solución AWS	2 días	lun 02/05/22	mar 03/05/22
Llamada con mentora	1 día	mar 03/05/22	mar 03/05/22
Provisión de medios AWS	2 días	mié 04/05/22	jue 05/05/22
Pruebas previas en AWS	1 día	vie 06/05/22	vie 06/05/22
Llamada grupo	1 día	sáb 07/05/22	sáb 07/05/22
Docker y Kubernetes	2 días	lun 09/05/22	mar 10/05/22
Llamada con mentora	1 día	mar 10/05/22	mar 10/05/22
Implementación final	2 días	mié 11/05/22	jue 12/05/22
Llamada grupo	1 día	sáb 14/05/22	sáb 14/05/22
Testing	2 días	lun 16/05/22	mar 17/05/22
Llamada con mentora	1 día	mar 17/05/22	mar 17/05/22
Llamada grupo	1 día	sáb 21/05/22	sáb 21/05/22
Llamada con mentora	1 día	mar 24/05/22	mar 24/05/22
Llamada grupo	1 día	sáb 28/05/22	sáb 28/05/22



Presentación	17 días	mié 01/06/22	mié 22/06/22
Presparación del pptx	2 días	mié 01/06/22	jue 02/06/22
Llamada con mentora	1 día	mar 07/06/22	mar 07/06/22
Llamada grupo	1 día	sáb 04/06/22	sáb 04/06/22
Simulacro y ensayo presentación	2 días	lun 06/06/22	mar 07/06/22
Presentación a Tribunal	1 día	lun 13/06/22	lun 13/06/22
Evaluación del tribunal	1 día	mié 22/06/22	mié 22/06/22



MEDIOS Y MATERIALES

Para la elaboración del proyecto se empleará una serie de recursos de *hardware* y *software*, que en el momento de la redacción de este plan de trabajo se encuentran ya disponibles. La lista de aplicaciones está sujeta a revisión en caso de ser necesario.

Bloque de investigación:

- Bibliografía específica sobre *Machine Learning*, incluyendo manuales en soporte papel, artículos en revistas científicas y publicaciones en Internet.
- Realización de ejercicios prácticos con *Amazon Web Services*.

Hardware:

El recurso principal es un portátil donde están instalados todos los programas. También se ha habilitado un equipo de apoyo, que servirá de alternativa en caso de que el equipo principal sufra cualquier eventualidad:

- Equipo principal: portátil HP con procesador Intel Core i7 a 2,80GHz, 16Gb de memoria RAM y SDD de 500Gb. El sistema operativo principal es Windows 10, instalado en una de sus particiones y Ubuntu 16.04LTS a través de máquina virtual.
- Equipo de apoyo: portátil MacBook Air con procesador M1 a 2,5GHz, 8Gb de memoria RAM y SSD de 512Gb. El sistema operativo principal es macOS Big Sur.

Software y herramientas:

Los equipos disponen de los siguientes programas:

- Microsoft Project 2016: utilizado para la planificación del proyecto y su seguimiento.
- Suite de Office 365: para la redacción de documentos y elaboración de presentaciones.
- Capa gratuita de AWS.

- Repositorio grupal en Github.
- Anaconda: para implementar el entorno de Python, con Visual Studio Code y PyCharm..
- Diversas librerías de Python para las fases de limpieza, exploración, visualización, etc.
- Docker y Kubernetes para la implementación del modelo en Amazon Web Services.

BIBLIOGRAFÍA PROVISIONAL

- Rob J Hyndman and George Athanasopoulos - Forecasting: principles and practice (2º edition)
- George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung - Time Series Analysis. Forecasting and Control (5th edition)
- Jerome H. Friedman, Rober Tibshirani and Trevor Hastie, The Elements of Statistical Learning.
- Chistopher Bishop, Pattern Recognition and Machine Learning.
- Aurélien Géron, O'Reilly - Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems
- Andreas C. Müller & Ssarah Guido, O'Reilly - Introduction to Machine Learning with Python
- Alice Zheng & Amanda Casari, O'Reilly - Feature Engineering for Machine Learning
- Peter Bruce, Andrew Bruce & Peter Gedeck - O'Reilly - Practical Statistic for Data Scientist (Second Edition)

- Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong - Mathematics for Machine Learning
- Jason Brownlee - Basics of Linear Algebra for Machine Learning
- Charu C. Aggarwal - Linear Algebra and Optimization for Machine Learning: A Textbook
- Jay Dawani - Hands-On Mathematics for Deep Learning: Build a solid mathematical foundation for training efficient deep neural networks
- Essentials of Business Analytics. An Introduction to the Methodology and its Applications. Bhimasankaram Pochiraju
- Mathematics for machine learning. Marc Peter Deisenroth
- Probability and Statistics for Computer Science. Springer. David Forsyth
- Data Mining Concepts and Techniques. Third Edition. Jiawei Han.
- Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data.
- Recommender Systems – An Introduction. Cambridge. Dietmar Jannach
- Aprende Machine Learning en español. Teoría + Práctica Python. Juan Ignacio Bagnato