1. Datos Generales de la asignatura

Nombre de la asignatura:	Análisis de datos y Machine Learning
Clave de la asignatura:	TDB-2405
SATCA[1]:	1-4-5
Carrera:	Ingeniería en Sistemas Computacionales

2. Presentación

Caracterización de la asignatura

Esta asignatura aporta al perfil del Ingeniero en Sistemas Computacionales la capacidad de diseñar, desarrollar y aplicar modelos computacionales para solucionar problemas, mediante la selección y uso de herramientas matemáticas.

La información obtenida de diversas fuentes permite a las empresas una mejor gestión de sus herramientas, decisiones internas y relación con los clientes.
La correcta captura, tratamiento, gestión y análisis de datos se está convirtiendo en una ventaja competitiva para las empresas, y eso se traduce en que crecen las inversiones en proyectos de Big data (terminología en idioma inglés, también llamado macro datos, datos masivos, inteligencia de datos, datos a gran escala) y aumenta la necesidad de expertos en este campo.

Los datos son, sin duda, son uno de los mayores activos de la empresa porque afectan a su modelo de negocio, a la relación, a la comunicación y a las relaciones con el cliente. En este contexto, Big data no solo permite gestionar y explotar este gran activo, sino que es el motor de las nuevas tecnologías exponenciales como son IoT, blockchain, la robótica o la nanotecnología. Tecnologías que van a marcar el escenario social, económico y empresarial del futuro.

El uso moderno del término big data tiende a referirse al análisis del comportamiento del usuario, extrayendo valor de los datos almacenados, y formulando predicciones a través de los patrones observados. La disciplina dedicada a los datos masivos se enmarca en el sector de las tecnologías de la información y la comunicación. Esta disciplina se ocupa de todas las actividades relacionadas con los sistemas que manipulan grandes conjuntos de datos. La tendencia a manipular enormes cantidades de datos se debe a la necesidad, en muchos casos, de incluir dicha información para la creación de informes

[1] Sistema de Asignación y Transferencia de Créditos Académicos

estadísticos y modelos predictivos utilizados en diversas materias, como los análisis de negocio, publicidad, los datos de enfermedades infecciosas, el espionaje y seguimiento a la población o la lucha contra el crimen organizado.

Por otro lado el aprendizaje automático (machine learning) utiliza técnicas de la Inteligencia Artificial para la detección de spam en corre y equipos de seguridad, robo de identidad, reconocimiento de voz y predicción de resultados por mencionar algunas. Las grandes empresas como Google, Microsoft, Amazon o Facebook han desarrollado herramientas para resolver los problemas antes mencionados, pero a la vez, han creado plataformas para el desarrolloo de aplicaciones que resuuelvan los problemas encontrando patrones en la información. Las aplicaciones de aprendizaje automático son actualmente las mas solicitadas debido a sus predicciones de comportamiento, por ejemplo, de compra para determinar si existe robo de identidad o para la clasificación de marketing ayudando a la optimización de publicidad.

Intención didáctica

La asignatura se conforma de cinco temas, cada tema conformado por contenidos que cotribuiran de forma gradual a la comprensión de lo que es la Ciencia de Datos y el aprendizaje de dos de sus componentes:

En el primer tema se abordan los conceptos referentes a la Ciencia de Datos, al análisis de grandes volúmenes de datos, componentes del sistema big data, los desafíos para la calidad en big data, las aplicaciones actuales del big data, así como la relación entre big data y machine learning.

En el segundo tema introduce al estudiante en el conocimiento de las fuentes de datos, en el procesamiento masivo de datos empleando la técnica de mapear y reducir, así como al procesamiento de conjunto de datos distribuidos resilientes.

Posteriormente, en el tercer tema se presentan las bases de datos relacionales y no relacionales, sus características y funcionamiento, así como el empleo de herramientas que permitan la consulta, agrupación, cálculo numérico, análisis de datos, aprendizaje automático y las librerías y herramientas necesarias para la visualización.

En el cuarto tema se presenta una introducción a machine learning, el análisis de datos empleando técnicas de estadística descriptiva y se abordan los tipos de aprendizaje supervisado, regresión, entrenamiento, algunos de sus problemas y algoritmos de clasificación.

En el quinto tema se abordan los tipos de aprendizaje no supervisado, clasificación preprocesamiento de la información y agrupamiento.

Por último, el estudiante desarrolla un proyecto final en el cual se realiza la gestión de almacenamiento, procesamiento, depuración, visualización y análisis de grandes volúmenes de datos, así como predicciones para la toma de decisiones.

3. Participantes en el diseño y seguimiento curricular del programa

Lugar y fecha de elaboración o revisión	Participantes	Observaciones
Instituto Tecnológico de la Laguna, semestre agosto-diciembre 2023	Integrantes de la Academia de Sistemas y Computación del ITL.	Diseño del módulo de especialidad de la carrera de Ingeniería en Sistemas Computacionales.

4. Competencia(s) a desarrollar

Competencia(s)específica(s)de la asignatura
Diseña y desarrolla soluciones que permitan la gestión del almacenamiento y tratamiento de grandes volúmenes de datos procedentes de diferentes contextos y problemas reales. Analiza grandes volúmenes de datos, empleando técnicas de estadística descriptiva y analítica predictiva para la creación de informes estadísticos y modelos predictivos para la toma de decisiones.

5. Competencias previas

- Analiza requerimientos y diseña bases de datos para generar soluciones al tratamiento de información basándose en modelos y estándares.
- Implementa bases de datos para apoyar la toma de decisiones considerando las reglas de negocio.
- Aplica los conceptos de la teoría de la probabilidad y estadística para organizar, clasificar, analizar e interpretar datos para la toma decisiones en aplicaciones de ingeniería biomédica, en computación y comunicaciones.
- Aplica la programación orientada a objetos para resolver problemas reales y de ingeniería.
- Aplica eficientemente estructuras de datos, métodos de ordenamiento y búsqueda para la optimización del rendimeinto de soluciones a problemas del mundo real.
- Aplica los principios lógicos y funcionales de la programación para aplicarlos en la resolución de problemas.

- Desarrolla soluciones de software para resolver problemas en diversos contextos utilizando programación concurrente, acceso a datos, que soporten interfaz gráfica de usuario y consideren dispositivos móviles.

6. Temario

No.	Temas	Subtemas
1	Introducción al Análisis de datos y Machine Learning	1.1. Ciencia de datos y Análisis de datos 1.2. Conceptos de Machine Learning y Deep Learning 1.3. Habilidades del analista de datos y científico de datos 1.4. Ciclo de vida del análisis de datos 1.5. Lenguajes de programación utilizados en el análisis de datos 1.5.1 Fundamentos 1.5.2 Librerías utilizadas en el análisis de datos y machine learning 1.5.2 Ejemplos 1.6. Aplicaciones.
2	Procesamiento de Datos (data wrangling)	2.1 Introducción al procesamiento 2.2 Preprocesamiento: Carga, almacenamiento y formato de los datos 2.3 Combinación de conjuntos de datos 2.4 Transformación de los datos. 2.5 Agregación y agrupamiento 2.6 Series de tiempo.
3	Visualización de Datos	3.1 Introducción a la visualización de datos. 3.2 Generalidades y selección de los gráficos según los datos y su uso (comparación, relación composición o distribución). 3.3 Nativas 3.4 Software y librerias de visualización para análisis de datos. 3.5 Gráficas estáticas y dinámicas.

4	Introducción a Machine Learning y Aprendizaje supervisado.	4.1. Introducción a Machine Learning 4.2. Minería de datos 4.3. Aprendizaje supervisado 4.3.1. Los k-vecinos más cercanos. 4.3.2. Modelos lineales 4.3.3. Clasificadores Bayes 4.3.4. Árboles de decisión 4.3.5. Conjuntos de árboles 4.3.6. Máquinas apoyadas en vectores 4.3.7. Redes neuronales artificiales.
5	Aprendizaje no supervisado	5.1 Tipos de aprendizaje no supervisado. 5.2 Preprocesamiento y escalamiento. 5.3 Reducción de la dimensionalidad 5.4 Agrupamiento (clustering) 5.4.1 Agrupamiento k-media 5.4.2 Agrupamiento aglomerativo. 5.4.3 DBSs

7. Actividades de aprendizaje de los temas

1. Introducción al análisis de datos y machine learning	
Competencias	Actividades de aprendizaje

Específica(s): Conoce y comprende los conceptos análisis de datos, ciencia de datos, machinel learning y Deep learning así como el ámbito de su aplicación. Genéricas: • Capacidad de abstracción, análisis y síntesis • Capacidad de comunicación oral y escrita • Habilidad para trabajar en forma autónoma • Habilidades para buscar, procesar y analizar información procedente de fuentes diversas • Conocimientos sobre el área de estudio y la profesión	• Elaborar un cuadro sinóptico de la Ciencia de Datos y disciplinas afines. • Búsqueda de información sobre los conceptos, características y aplicaciones reales de big data y machine learning (análisis de datos). • Búsqueda de información acerca de las empresas regionales, nacionales y extranjeras que empleen aplicaciones con big data y/o machine learning y como les apoya en sus organizaciones. • Elabora tabla de clasificación de aplicaciones de las empresas. • Instala la herramienta Anaconda y aprende sobre su entorno y aplicación. • Aprende y aplica fundamentos de las herramientas IPython y Jupyter notebook.
2. Procesamiento de Datos	
Competencias	Actividades de aprendizaje
Específica(s): Crea un conjunto de datos, realiza consultas, muestra información y elabora gráficos. Genéricas: • Capacidad de abstracción, análisis y síntesis • Habilidad para trabajar en forma autónoma • Habilidades para buscar, procesar y analizar información procedente de fuentes diversas • Capacidad de aplicar los conocimientos en la práctica • Capacidad para identificar, plantear y resolver problemas	• Búsqueda de información sobre fuentes de datos. • Búsqueda de información de los principales conjuntos de datos además de métodos en Scikit-learn. • Instalar un Dataset a HDFS en línea de comandos. • Codificar un histograma de calificaciones con MapReduce. • Realizar varias consultas sobre un Dataset para ver funcionamiento de MapReduce. • Mostrar información utilizando RDD. • Mostrar información utilizando DataFrames.

	• Desarrollar la solución a problemas con las bibliotecas numpy, pandas y scipy. • Desarrolla una aplicación para la exploración de hojas en excel incrustación de comandos SQL y manejo de archivos CSV.
3. Visualización de Datos	
Competencias	Actividades de aprendizaje
Específica(s): Comprende tendencias, valores atípicos y patrones en los datos utilizando herramientas que permitan visualizar los datos Genéricas: • Capacidad de abstracción, análisis y síntesis • Habilidad para trabajar en forma autónoma • Habilidades para buscar, procesar y analizar información procedente de fuentes diversas • Capacidad de aplicar los conocimientos en la práctica • Capacidad para identificar, plantear y resolver problemas	• Utiliza y experimenta con conjuntos de datos incluidos en scikit-learn la biblioteca matplotlib para la visualización de dichos datos en forma de diagramas de barras, histogramas, gráficos de dispersión. • Utiliza la biblioteca pandas para la visualización de datos en forma de tabla. • Utiliza la biblioteca cartopy para visualizar los datos sobre mapas. • Experimenta varios tipos de graficas con las librerías seaborn y bokeh
4. Introducción a Machine Learning y Aprendizaje supervisado	
Competencias	Actividades de aprendizaje
Específica(s): Conoce y aplica ténicas de aprendizaje supervisado y no supervisado. Genéricas: • Capacidad de abstracción, análisis y síntesis • Habilidad para trabajar en forma autónoma • Habilidades para buscar, procesar y analizar información procedente de fuentes diversas • Capacidad de aplicar los conocimientos en la práctica	• Investiga la importancia y aplicaciones de machine learning. • Realiza un análisis comparativo de las ventajas y desventajas de usar machine learning. • Implementa algoritmos básicos para resolver problemas reales. • Utiliza un conjunto de datos de un repositorio libre para ejemplificar un problema de clasificación binaria y clasificación múltiple, utilizando un algoritmo de aprendizaje supervisado.

• Capacidad para identificar, plantear y resolver problemas	• Utiliza un conjunto de datos de un repositorio libre para ejemplificar un problema de regresión utilizando un algoritmo de aprendizaje supervisado. • Analiza las fortalezas, debilidades, parámetros y evaluación del modelo en las dos prácticas anteriores.
5. Aprendizaje no supervisado	
Competencias	Actividades de aprendizaje
Específica(s): Conoce y aplica ténicas de aprendizaje supervisado y no supervisado. Genéricas: • Capacidad de abstracción, análisis y síntesis • Habilidad para trabajar en forma autónoma • Habilidades para buscar, procesar y analizar información procedente de fuentes diversas • Capacidad de aplicar los conocimientos en la práctica • Capacidad para identificar, plantear y resolver problemas	• Utiliza un conjunto de datos incluidos en scikit-learn para ejemplificar un problema de transformación, utilizando preprocesamiento y escalamiento. Mostrar un encadenamiento con un algoritmo de aprendizaje automáticos supervisado. • Utiliza un conjunto de datos incluido en la biblioteca scikit-learn que encuentre una mejor visualización, comprensión y representación de los datos para un posterior procesamiento. Utilizar el algoritmo de análisis de componentes principales (PCA). • Utiliza un conjunto de datos incluido en la biblioteca scikit-learn donde se aplique el agrupamiento (clustering). Utilice cualquiera de los algoritmos para agrupamiento: agrupamiento k-media, agrupamiento aglomerativo o el DBSACN. • Analiza la representación y características de los datos además de la evaluación del modelo y su mejoramiento en las prácticas anteriores.

8.Práctica(s)

Se sugieren las siguientes prácticas:

- Búsqueda de información acerca del uso de big data y Machine Learning en aplicaciones comerciales.
- Instala bases de datos relacionales y no relacionales.
- Instalación de Anaconda y las librerias numpy scipy scikit-learn matplotlib pandas pillow graphviz preamble watermark seaborn bokeh
- Fundamentos de uso de las herramientas IPython y Jupyter Notebook
- Investigar y utilizar los principales modelos de conjuntos de datos además de métodos en scikit-learn
- Crea grandes conjuntos de datos y realiza consultas empleando herramientas de software para big data.
- Analiza grandes volúmenes de datos empleando técnicas de estadística descriptiva.
- Analiza grandes volúmenes de datos empleando analíticas predictiva y web.
- Realiza consultas y las presenta en interfaces gráficas para usuario adecuadas (visualización de datos).
- Utilizar con conjuntos de datos incluidos en scikit-learn o de un repositorio libre, la biblioteca Matplotlib para visualización de dichos datos en forma de diagramas de barras, histogramas, gráficos de dispersión.
- Utilizar un conjunto de datos incluidos en scikitlearn o de un repositorio libre, para ejemplificar un problema de clasificación binaria y clasificación múltiple, utilizando uno de los algoritmos de machine learning supervisado.
- Utilizar un conjunto de datos incluidos en scikitlearn o de un repositorio libre, para ejemplificar un problema de regresión, utilizando uno de los algoritmos de machine learning supervisado.
- Analizar la representación y características de los datos además de la evaluación del modelo y su mejoramiento, para las dos prácticas anteriores.
- Utilizar un conjunto de datos incluidos en scikitlearn o de un repositorio libre, para ejemplificar un problema de transformación, utilizando Preprocesamiento y escalamiento.
- Utilizar un conjunto de datos incluido en la biblioteca scikitlearn o de un repositorio libre, que encuentre una mejor visualización, compresión y representación de los datos para un posterior procesamiento. Utilizar el algoritmo de análisis de componentes principales (PCA)
- Utilizar un conjunto de datos incluido en la biblioteca scikit- learn o de un repositorio libre, donde apliquemos el agrupamiento (clustering).
- Utilice cualquiera de los algoritmos para agrupamiento: agrupamiento k-media, agrupamiento aglomerativo o el DBSCAN.

9. Proyecto de asignatura

Crear un proyecto que solucione un problema real utilizando los conceptos aprendidos en la asignatura. El proyecto deberá comprender la gestión de almacenamiento, procesamiento, depuración, visualización, análisis de grandes volúmenes de datos y/o predicción para la toma de decisiones.

10. Evaluación por competencias

La evaluación debe ser continua y formativa por lo que se debe considerar el desempeño en cada una de las actividades de aprendizaje, haciendo énfasis en:

- Investigaciones documentadas, reuniéndose después para realizar una lluvia de ideas o bien mesas de trabajo, donde los estudiantes interactúan y presentan la información investigada por cada equipo.
- Reportes escritos de las soluciones planteadas durante las actividades, así como las conclusiones obtenidas de dichas soluciones.
- Elaborar mapas conceptuales por equipo de los temas explicados en el aula con el fin de reforzar el aprendizaje adquirido.
- Usar tecnología de información (internet, libros electrónicos, artículos, revistas electrónicas, etc.) para efectuar una recopilación de información de temas afines a los contenidos temáticos del curso.
- Exámenes escritos y prácticos por unidades de aprendizaje.
- Descripción de otras experiencias concretas que se obtendrán al participar en eventos, conferencias, paneles de discusión, o cualquier otro medio didáctico profesional que trate sobre la materia y que deberán realizarse durante el curso.
- Presentación y exposición de actividad de aprendizaje. Algunas se evaluarán por equipo.
- Implementar una dinámica para solucionar un problema en el que se requiera establecer estrategias para su solución por medio del liderazgo efectivo.
- Planear, organizar y ejecutar una mesa redonda sobre el impacto de las habilidades directivas en el logro de los objetivos organizacionales.

11. Fuentes de información

Aven, J. (2018). *Data Analytics with Spark Using Python*. Addison-Wesley Professional.

Deitel, P., & Harvey, D. (2019). *Intro to Python for Computer Science and Data Science: Learning to Program with Ai, Big Data and the Cloud*. Pearson.

Feasel, K. (2020). *Polybase Revealed: Data Virtualization with SQL Server, Hadoop, Apache Spark, and Beyond*. Apress.

Golerik, A. (2019). *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reilly.

Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Archivo Aurélien Géron, O'Reilly Media, Inc. (2019).

Introduction to Machine Learning with Python. Andreas C. Müller and Sarah Guido. O'Reilly Media, Inc. (2016).

Karau, H., & Warren, R. (2017). *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*. O'Reilly.

Mittal, M., Balas, V. E., Mohan Goyal, L., & Kumar, R. (2019). *Big Data Processing Using Spark in Cloud*.Springer.

Mongo DB Inc. (2019). *MongoDB*. Retrieved from La base de datos líder para aplicaciones modernas: https://www.mongodb.com/es

Oracle Corporation. (2019). *MySQL*. (Oracle, Editor) Retrieved from The world's most popular open source database: https://www.mysql.com

The Apache Software Foundation. (2014). *APACHE HIVE TM*. (T. A. Foundation, Editor) Retrieved from https://hive.apache.org/

The Apache Software Foundation. (2016). *Apache CASSANDRA*. (T. A. Foundation, Editor) Retrieved from Manage massive amounts of data, fast, without losing sleep: http://cassandra.apache.org/

The Apache Software Foundation. (2018). *Apache Hadoop*. Retrieved septiembre 23, 2019, from The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.: https://hadoop.apache.org

The Apache Software Foundation. (2018). *Apache Spark™ Lightning-fast unified analytics engine*. Retrieved from Apache Spark™ is a unified analytics engine for large-scale data processing: https://spark.apache.org

The Apache Software Fundation. (2019). *Apache ZooKeeper™*. (T. A. Fundation, Editor) Retrieved from https://zookeeper.apache.org