UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

Maestría en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software

SÍLABO

Nombre de la asignatura	:	MACHINE LEARNING Y BIG DATA
Profesor responsable	:	Mg. Juan Gamarra Moreno
Correo electrónico	:	juan.gamarra@unmsm.edu.pe

2024-2

1. INFORMACIÓN GENERAL

1.1	Nombre de la asignatura	:	MACHINE LEARNING Y BIG DATA
1.2	Tipo de asignatura	:	Obligatoria
1.3	Profesor	:	Mg. Juan Gamarra Moreno
1.4	Programa	:	Maestría en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software
1.5	Código de asignatura	:	R3P2A111, R3S2A111
1.6	Créditos	:	4
1.7	N° de horas semanales	:	3
1.8	N° de horas por semestre	:	48
1.9	Semestre académico	:	2024-2
1.10	Duración	:	16 semanas
1.11	Fecha de inicio	:	14 de septiembre de 2024
1.12	Fecha de finalización	:	02 de enero de 2025
1.13	Horario	:	Sábados 3:00 p.m.-6:00 p.m.

2. FUNDAMENTOS DE LA ASIGNATURA

2.1. Sumilla.

Asignatura que corresponde al periodo Profundización, es de naturaleza Teórica-practica de modalidad presencial. Tiene el propósito de brindar los conocimientos para la predicción de patrones en grandes volúmenes de información estructurados y no estructurados, aglomerados en bases de datos o dispersos en diversos formatos en internet. Abarca los siguientes aspectos: Modelado de datos, Datawarehouse, Análisis multidimensional, Técnicas Data Mining, aprendizaje automático supervisado y no supervisado, Cloud Computing, Big data. Culmina con una aplicación práctica. Al finalizar el estudiante deberá presentar un informe aplicando los tópicos tratados.
Las unidades son:
• Patrones
• Modelado de datos
• Análisis multidimensional
• Técnicas Data Mining

2.2. Competencias del programa.

El curso contribuye al logro de competencias, tales como:

2.2.1. Realizar investigación de alto nivel académico en al menos una de las líneas de investigación del programa.

2.2.2. Identificar y adoptar tecnologías disruptivas basadas en Gestión de la Información y del Conocimiento para la transformación digital de la organización.

2.2.3. Aplicar el análisis crítico y pensamiento creativo para la identificación y solución de problemas de la organización con tecnologías emergentes de Gestión de la Información y del Conocimiento.

2.3. Competencias de la asignatura.

2.3.1. Comprende los fundamentos teóricos de Machine Learning y Big Data, aplicando técnicas de preprocesamiento y algoritmos de regresión, clasificación y agrupamiento para resolver problemas, mientras se actúa con responsabilidad ética al manejar datos masivos, asegurando la privacidad y minimizando sesgos en los modelos.

2.3.2. Usa algoritmos avanzados como redes neuronales profundas, Deep Learning y modelos Transformers, implementando soluciones efectivas con herramientas y tecnologías adecuadas, manteniendo una actitud innovadora y colaborativa en la resolución de problemas complejos mediante tecnologías emergentes.

2.3.3. Despliega modelos de Machine Learning en producción, optimizando su rendimiento en entornos reales utilizando plataformas adecuadas y la nube, y promoviendo una mentalidad crítica, ética y proactiva en la automatización, el monitoreo continuo y la mejora de modelos de inteligencia artificial.

3. CONTENIDO TEMÁTICO

3.1. **Unidad de Aprendizaje 1:** Fundamentos y Modelos Clásicos de Machine Learning

Competencias de la asignatura

- Comprende los fundamentos teóricos de Machine Learning y Big Data, aplicando técnicas de preprocesamiento y algoritmos de regresión, clasificación y agrupamiento para resolver problemas, mientras se actúa con responsabilidad ética al manejar datos masivos, asegurando la privacidad y minimizando sesgos en los modelos.

Logros de aprendizaje:

- Explicar los fundamentos teóricos de Machine Learning y Big Data, comprendiendo sus principios, tipos de aprendizaje (supervisado y no supervisado) y sus aplicaciones en diferentes industrias.
- Preprocesar datos para su uso en modelos de Machine Learning, aplicando técnicas como la limpieza, normalización, estandarización y codificación de variables utilizando herramientas como Python y bibliotecas como Pandas y Scikit-learn.
- Implementar algoritmos clásicos de Machine Learning supervisado, como regresión lineal y clasificación (Regresión Logística, KNN), evaluando su rendimiento mediante métricas adecuadas en datasets reales.
- Aplicar algoritmos no supervisados, como K-Means y clustering jerárquico, para analizar datos no etiquetados, evaluando la calidad de los clusters generados a través de métodos como Elbow y Silhouette Score.
- Desarrollar redes neuronales artificiales básicas (ANN), aplicando funciones de activación y retropropagación, y evaluando su rendimiento en problemas de clasificación de datos.
- Actuar con responsabilidad ética en el manejo de datos, asegurando la privacidad y minimizando los sesgos durante el preprocesamiento y la implementación de modelos de Machine Learning.

Contenido Temático:

Semana 1: Introducción a Machine Learning y Big Data
Teoría:
- Definiciones de Machine Learning (ML) y Big Data.
- Diferencias y similitudes entre ML y Big Data.
- Aplicaciones en el mundo real.

Laboratorio:
- Configuración del entorno de trabajo con Python.
- Introducción a las bibliotecas esenciales (Pandas, NumPy, Scikit-learn).

Semana 2: Preprocesamiento de Datos
Teoría:
- Limpieza de datos: manejo de datos faltantes, duplicados, y transformación de datos.
- Normalización, estandarización y codificación de variables.

Laboratorio:
- Preprocesamiento de un dataset real utilizando Pandas y Scikit-learn.

Semana 3: Aprendizaje Supervisado - Regresión Lineal
Teoría:
- Introducción a los modelos de regresión (Regresión Lineal, Regularización: Ridge, Lasso).
- Evaluación del modelo: métricas de rendimiento (MSE, R^2).

Laboratorio:
- Implementación de un modelo de regresión lineal en Python con Scikit-learn.
- Evaluación y ajuste del modelo con diferentes técnicas de regularización.

Semana 4: Aprendizaje Supervisado - Clasificación
Teoría:
- Modelos de clasificación (Regresión Logística, KNN, Support Vector Machines).
- Métricas de evaluación: Precisión, Recall, F1-Score.

Laboratorio:
- Implementación de un modelo de clasificación utilizando Regresión Logística.
- Interpretación de resultados con métricas de evaluación.

Semana 5: Árboles de Decisión y Random Forest
Teoría:
- Introducción a Árboles de Decisión.
- Conceptos de sobreajuste y técnicas para evitarlo.
- Introducción a Random Forest y su uso en problemas de clasificación y regresión.

Laboratorio:
- Implementación de Árboles de Decisión y Random Forest.
- Comparación de rendimiento entre ambos modelos.

Semana 6: Aprendizaje No Supervisado - Clustering
Teoría:
- Introducción al Clustering: K-Means, DBSCAN, Clustering Jerárquico.
- Métodos de evaluación: Silhouette Score, Elbow Method.

Laboratorio:

• Implementación de K-Means Clustering en un dataset no etiquetado. • Visualización y análisis de los clusters generados.
Semana 7: Redes Neuronales Artificiales (ANN) Teoría: • Fundamentos de las redes neuronales: estructura, funciones de activación, y retropropagación. • Redes Neuronales frente a otros modelos de ML. Laboratorio: • Implementación de una red neuronal básica usando TensorFlow/Keras. • Entrenamiento y evaluación de una ANN para clasificación.
Semana 8: Evaluación Parcial • Examen teórico y práctico sobre los temas cubiertos en las semanas anteriores. • El examen práctico consistirá en implementar y ajustar un modelo de ML basado en un dataset proporcionado.

3.2. **Unidad de Aprendizaje 2:** Técnicas Avanzadas y Aplicaciones en Producción

Competencias de la asignatura

- Usa algoritmos avanzados como redes neuronales profundas, Deep Learning y modelos Transformers, implementando soluciones efectivas con herramientas y tecnologías adecuadas, manteniendo una actitud innovadora y colaborativa en la resolución de problemas complejos mediante tecnologías emergentes.
- Despliega modelos de Machine Learning en producción, optimizando su rendimiento en entornos reales utilizando plataformas adecuadas y la nube, y promoviendo una mentalidad crítica, ética y proactiva en la automatización, el monitoreo continuo y la mejora de modelos de inteligencia artificial.

Logros de aprendizaje:

- Explicar los principios avanzados de Deep Learning, comprendiendo el funcionamiento de redes neuronales profundas (ANN, CNN) y su aplicación en tareas de clasificación de imágenes y otros datos complejos.
- Implementar redes neuronales convolucionales (CNN) para la clasificación de imágenes, utilizando bibliotecas como TensorFlow y Keras, y evaluar su rendimiento utilizando datasets como MNIST.
- Utilizar herramientas de procesamiento de datos masivos (Big Data), como Hadoop y Spark, para procesar grandes volúmenes de datos de manera eficiente, aplicando técnicas de paralelización y procesamiento distribuido.
- Desarrollar modelos de aprendizaje automático distribuidos utilizando Spark MLlib, implementando algoritmos escalables y optimizando el rendimiento en entornos de Big Data.
- Aplicar modelos de procesamiento de lenguaje natural (NLP) basados en Transformers, como BERT o GPT, ajustando modelos preentrenados y evaluando su rendimiento en tareas como clasificación de texto o análisis de sentimientos.
- Desplegar modelos de Machine Learning en producción, utilizando herramientas como Flask o FastAPI para crear APIs y monitorear el rendimiento de los modelos en entornos reales.

- Adoptar prácticas de MLOps para gestionar el ciclo de vida de los modelos de Machine Learning en producción, incluyendo el monitoreo, mantenimiento y actualización continua de los modelos.
- Demostrar una actitud crítica y ética en el uso de modelos avanzados de Machine Learning y Big Data, asegurando la equidad, transparencia y minimización de sesgos en los resultados, y considerando el impacto de estos modelos en la sociedad.

Contenido Temático:

Semana 9: Introducción a Deep Learning y Redes Convolucionales (CNN) Teoría: • Introducción a Deep Learning y redes neuronales profundas. • Introducción a las CNN: estructura y casos de uso. Laboratorio: • Implementación básica de una CNN para clasificación de imágenes (MNIST dataset).
Semana 10: Procesamiento de Datos Masivos con Big Data Teoría: • Introducción a las arquitecturas de Big Data: Hadoop, Spark. • Modelos de procesamiento: Batch Processing vs Stream Processing. Laboratorio: • Introducción a PySpark para el procesamiento de grandes volúmenes de datos. • Análisis de un dataset masivo con Spark.
Semana 11: Aprendizaje Automático Distribuido Teoría: • Escalabilidad en Machine Learning. • Técnicas de paralelización y procesamiento distribuido. • Introducción a MLlib de Apache Spark. Laboratorio: • Implementación de algoritmos de ML distribuidos con Spark MLlib. • Comparación de rendimiento en datasets masivos.
Semana 12: Procesamiento de Lenguaje Natural (NLP) Teoría: • Introducción a NLP: tokenización, lematización, y modelos de lenguaje. • Aplicación de redes neuronales en NLP (Word2Vec, embeddings). Laboratorio: • Implementación de modelos básicos de NLP usando NLTK y SpaCy. • Análisis de sentimientos en textos con ML.
Semana 13: Redes Generativas Antagónicas (GANs) Teoría: • Introducción a GANs: generadores y discriminadores. • Aplicaciones de GANs en la generación de datos sintéticos. Laboratorio: • Implementación básica de una GAN usando TensorFlow/Keras. • Generación de imágenes sintéticas a partir de ruido.

Semana 14: Introducción a Transformers y Modelos Preentrenados

Teoría:

- Introducción a la arquitectura Transformer: mecanismo de atención, auto-atención, capas de codificador y decodificador.
- Ventajas de los Transformers frente a modelos anteriores (RNN, LSTM) en procesamiento de secuencias.
- Modelos preentrenados basados en Transformers: BERT, GPT, T5, entre otros.
- Aplicaciones prácticas: clasificación de texto, generación de texto, traducción automática, y análisis de sentimientos.

Laboratorio:

- Uso de la biblioteca Hugging Face para trabajar con modelos preentrenados.
- Carga de un modelo preentrenado como BERT o GPT.
- Ajuste fino (fine-tuning) en un dataset específico, como una tarea de clasificación de texto o análisis de sentimientos.
- Evaluación de métricas de rendimiento y comparación con otros modelos.

Semana 15: Implementación de Modelos de ML en Producción

Teoría:

- Ciclo de vida de un modelo de ML en producción: desde el desarrollo hasta el despliegue.
- Consideraciones prácticas: rendimiento, escalabilidad, y mantenimiento de modelos en entornos reales.
- Introducción a MLOps: cómo gestionar el desarrollo, el despliegue y el mantenimiento de modelos a gran escala.
- Herramientas para despliegue: Flask, FastAPI, Docker, y plataformas en la nube (AWS, Google Cloud, Azure).

Laboratorio:

- Desarrollo de una API sencilla para servir un modelo de ML utilizando Flask o FastAPI.
- Despliegue de un modelo en un entorno local o en la nube (puede usarse Google Cloud o AWS).
- Creación de endpoints para la predicción en tiempo real.
- Implementación de un pipeline para monitorizar el rendimiento del modelo en producción.

Semana 16: Evaluación Final

- Examen teórico y práctico sobre los temas de las semanas 9 a 15.
- El examen práctico consistirá en desarrollar y desplegar un modelo de ML o Transformer aplicado a un problema real, incluyendo preprocesamiento, entrenamiento y despliegue en producción.

3.3. Bibliografía.

- Navlani A., Fandango A., Idris I. (2021), "Python Data Analysis", 3rd Ed, Packt.
- Zhang A. (2017), Data Analytics. Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life
- James G., Witten D, Hastie T. Tibshirani R. (2013), "An Introduction to Statistical Learning with Applications in R", Springer.

- Montgomery D., Peck E., Vining G. (), "Introduction to Linear Regression Analysis", 5th Ed, Wiley.
- Bobadilla J. (2020). "Machine Learning y Deep Learning. Usando Python, Scikit y Keras", Ra-Ma.
- Albon C. (2018) "Machine Learning with Python Cookbook. Practical Solutions from Preprocessing to Deep Learning", 1st, Ed. O'Reilly.
- Fuentes A. (2018), "Hands-On Predictive Analytics with Python", Packt.
- Tan P., Steinbach M., Kumar V., Karpatne A. (2019), "Introduction to Data Mining", 2nd Ed., Pearson Education.
- Grus J. (2019). "Data Science from Scratch", 2nd Ed., O'Reilly
- Müller A., Guido S. (2017). "Introduction to Machine Learning with Python", 1st Ed. O'Relly.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
- Chollet, F. (2021). Deep Learning with Python (2nd ed.). Manning Publications.
- Zaharia, M., & Wendell, P. (2020). Learning Spark: Lightning-Fast Data Analytics (2nd ed.). O'Reilly Media.
- Vaswani, A., et al. (2017). Attention is All You Need. NeurIPS Conference.
- Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Recursos Electrónicos

- **TensorFlow**
  Recurso oficial de Google para aprender y aplicar Machine Learning y Deep Learning con TensorFlow. Incluye tutoriales, guías y ejemplos.
  https://www.tensorflow.org/

- **Scikit-learn**
  Biblioteca de Python para Machine Learning. La documentación oficial proporciona guías prácticas, ejemplos y referencias sobre algoritmos de aprendizaje supervisado y no supervisado.
  https://scikit-learn.org/stable/

- **Apache Spark**
  Página oficial de Apache Spark, una plataforma de procesamiento distribuido utilizada en Big Data. Incluye documentación, guías de instalación y tutoriales.
  https://spark.apache.org/

- **Hugging Face**
  Plataforma líder en modelos preentrenados basados en Transformers (BERT, GPT, etc.). Proporciona acceso a modelos, datasets y guías para ajustar y desplegar modelos de NLP.
  https://huggingface.co/

- **Kaggle**
  Plataforma para ciencia de datos que ofrece datasets gratuitos, tutoriales y competiciones de Machine Learning. Excelente recurso para obtener datasets y practicar con modelos.

https://www.kaggle.com/

- **Google Cloud - Big Data and Machine Learning Products**
  Documentación y guías sobre cómo utilizar las herramientas de Big Data y Machine Learning en Google Cloud, incluyendo BigQuery, AutoML, y AI Platform.
  https://cloud.google.com/products/ai

- **Coursera - Machine Learning by Stanford University**
  Curso en línea impartido por Andrew Ng, uno de los cursos más recomendados para aprender los fundamentos de Machine Learning.
  https://www.coursera.org/learn/machine-learning

4. ESTRATEGIAS METODOLÓGICAS

La metodología de enseñanza es semi presencial, los contenidos y material de enseñanza se publicarán en el Aula Virtual de la Unidad de Posgrado, la sesión de clase se desarrollará mediante videoconferencias en línea (síncronos) y con sesiones presenciales según la programación horaria del curso. Entre las estrategias de enseñanza a utilizar se tiene: La clase magistral, discusión grupal, aprendizaje basado en problemas ABP, aprendizaje basado en proyectos, asesoría, y talleres. El presente curso utilizará Microsoft Teams para videoconferencias y para pizarra electrónica Microsoft Whiteboard.

5. ESTRATEGIAS DE EVALUACIÓN

5.1. Modalidades de evaluación:

La Modalidad de evaluación comprende; solución a problemas propuestos, entregables del artículo que deberán ser presentados mediante la plataforma del Aula Virtual, asimismo las exposiciones de trabajos deberán realizarse mediante videoconferencias en línea cuyos enlaces a sus contenidos deberán registrase en la plataforma del Aula Virtual

5.2. Criterios de evaluación

Se consideran como trabajos los ejercicios propuestos en las sesiones de clase.

Ponderaciones:

Modalidades	Porcentaje
N1 Trabajos desde la semana 1 hasta la semana 7.	30%
N2 Trabajos desde la semana 9 hasta la semana 15	30%
N3 Evaluación Parcial	20%
N4 Evaluación Final	20%
Total	100%

NF= 30% N1+ 30% N2 +20% N3 + 20% N4

Nota aprobatoria mínima: 14

5.3. Requisitos para aprobar la asignatura:

- Es un requisito contar con asistencia mayor o igual al 70% las cuales serán registradas por el docente en el SUM (Sistema Único de Matricula) según la programación académica.
- Obtener un promedio final mayor o igual a 14.

Ciudad Universitaria, Septiembre del 2024