Diplomatura Superior en Ciencia de Datos

ESTRUCTURA DEL PLAN DE ESTUDIOS

El diseño curricular de la Diplomatura comprende el cursado de 8 materias agrupados en 3 áreas: la primera denominada Fundamentos (100hs) está compuesta por materias introductorias. El Área Núcleo (140 hs) incluye asignaturas enfocadas a conceptos básicos en Ciencia de Datos y el Área Específicas (70 hs) que consiste en temas de aplicación. Todas distribuidas en dos cuatrimestres. **Carga horaria total 310 horas.**

Área Fundamentos		
Asignatura	Carga horaria (horas reloj)	Cuatrimestre
Análisis Estadístico	40	1º
Programación y Bases de Datos	60	1º

Área Núcleo		
Asignatura	Carga horaria (horas reloj)	Cuatrimestre
Captura de la Información	40	1º
Aprendizaje Automático	40	2º
Arquitecturas en la Nube	30	1º
Minería de Datos	30	2º

Área Específicas		
Asignatura	Carga horaria (horas reloj)	Cuatrimestre
Aplicaciones de Inteligencia de Datos	40	2º
Procesamiento para Grandes Datos	30	2º

CONTENIDOS MÍNIMOS

Área Fundamentos

Análisis Estadístico

Carga horaria total: 40 horas

Contenidos:

1-Distribuciones de probabilidades. Estadística descriptiva. Introducción a la programación en R. Importación de datos y scripting (archivo de instrucciones) para procesamiento intensivo de datos. Análisis de estadística descriptiva en R.

2- Técnicas de Contrastes de hipótesis paramétricas. Test de hipótesis para la media, la varianza y cociente de varianzas. Test de hipótesis para proporciones.

3- Métodos de regresión lineal. Regresión lineal simple. Test de hipótesis en regresión lineal simple. Intervalos de confianza para la respuesta media y predicciones futuras.

Análisis de residuos y coeficiente de determinación. El modelo de regresión múltiple. Estimación de los coeficientes de regresión múltiple. Inferencias en la regresión múltiple. Inferencias basadas en el coeficiente de determinación. Predicciones basadas en la regresión múltiple.
4- Análisis de varianza. Diseño de experimentos en ingeniería. Comparación de Medias Múltiples. Análisis de varianza. ANOVA para Poblaciones Dependientes.
5- Estadística no paramétrica. Test del signo. Test de rangos signados de Wilcoxon. Métodos no paramétricos en ANOVA. Kruskal-Wallis.

Programación y Bases de Datos
Carga horaria total: 60 horas
Contenidos:
1. Programación.
Paradigmas de programación. Lenguajes de programación.
Características, ventajas y desventajas, casos de uso.
Programación. Tipos de datos. Estructuras. Condicionales. Ciclos. Recursividad. Programación Orientada a Objetos. Frameworks.
2. Base de datos relacionales: características, casos de uso, ventajas y desventajas.
Mysql, Oracle, PostgreSQL, SQL Server. Instalación y uso.
SQL. Operaciones CRUD.
3. Base de datos no relacionales (NoSQL): características, casos de uso, ventajas y desventajas.
MongoDB. Instalación y uso. Estructura. Operaciones CRUD. Operadores.
Cassandra. Instalación y uso. Estructura. CQL. Operaciones CRUD.
Redis. Instalación y uso. Estructura. Comandos.
4. Integración de las bases de datos con programación.
Python con bases de datos: SQLAlchemy. PyMongo. Cassandra driver. Redis py.

Área Núcleo

Captura de la Información
Carga horaria total: 40 horas
Contenidos:
− Big Data. Características y desafíos del Big Data. Fuentes de Datos.
− Digitalización de Datos. Tipos de Datos y sus Características: Ópticos: Imágenes, videos y otros. Sonoros: voz, música. Provenientes de Otros Sensores.
− Dispositivos de Captura: Manuales y OCR. Escáneres. Cámaras. Lectores de Marcas: OMR, MICR. Micrófonos. Sensores Varios: Radares, Nucleares, etc.
− Lenguaje de marcado. Metadatos.
− Captura de datos. Extracción de datos de las redes sociales. Obtención de datos históricos y datos en tiempo real. Utilización de APIs y herramientas para capturar datos.
− Captura de información de la web. Web scraping y web crawling. Tipos de web crawlers. Uso de expresiones regulares. Librerías y frameworks específicos.
− Bases de Datos relacionales y no relacionales o NoSQL. Breve historia de NoSQL. Descripción y tipos de bases de datos NoSQL: orientadas a columnas, documentos,

claves-valores, grafos, objetos o híbridas. ACID vs. BASE. Teorema de Brewer. Ventajas y desventajas de NoSQL.
− Orientadas a documentos - MongoDB

Aprendizaje Automático
Carga horaria total: 40 horas
Contenidos:
Introducción al Aprendizaje Automático. Los orígenes del aprendizaje automático. Usos del aprendizaje automático. Aciertos y límites. Aspectos éticos. Aprendiendo patrones a partir de los datos. Buenas prácticas de diseño y evaluación de performance.
− Almacenamiento y estructura de datos. Abstracción. Generalización. Evaluación. Aprendizaje supervisado y no supervisado. Tipos de datos de entrada y tipos de algoritmos de aprendizaje.
− Preprocesamiento y generación de características. Selección de atributos. Reducción de la dimensión del espacio de entrada. Covarianza. Análisis de componentes principales.
− Regresión. Regresión lineal en una y varias variables. Método del gradiente. Regresión logística.
− Máquinas de soporte Vectorial (SVM). Definición. Hiperplano óptimo. Clasificación lineal y no lineal. Máximo margen y vectores soporte. Formulación Dual. Optimización cuadrática. Kernels usuales. SVM multiclase.
− Redes Neuronales Feedforward. Descripción de la arquitectura. Regla delta generalizada. Algoritmo de entrenamiento backpropagation. Incorporación del término de momento. Capacidad de generalización de la red. Resolución de problemas de clasificación y predicción. Aprendizaje profundo.
− Redes Neuronales Competitivas. Técnicas de Agrupamiento partitivas. Agrupamiento utilizando redes neuronales. Red CPN y red SOM. Similitudes y diferencias con el agrupamiento producido por k-medias.
− Algoritmos de aprendizaje automático para Big Data.

Arquitecturas en la Nube
Carga horaria total: 30 horas
Contenidos:
1- Conceptos Básicos: Definición del Cloud Computing. Raíces. Riesgos y desafíos. Características. Seguridad. Capas. Tipos de Cloud. Aplicaciones.
2- IaaS: Infraestructura como servicio. Definición. Alcance. Ventajas de su implementación. Ejemplos. Amazon Web Services. Oracle Storage.
3- PaaS & SaaS: Plataforma como servicio. Software como servicio. Definición. Alcance. Ejemplos. Microsoft Azure. Google cloud. Google Drive. Dropbox.
4- Despliegue de Cloud privados: Despliegue, administración y configuración de Cloud privados a través de herramientas Open Source como OpenStack y OpenNebula. Contenedores: Definición. Alcance. Ventajas y desventajas. LXC (Linux Containers). Dockers.

Minería de Datos
Carga horaria total: 30 horas

Contenidos:
Introducción. Obtención de conocimiento a partir de los datos. El concepto de patrón. El proceso KDD. Fases del proceso de extracción del conocimiento. La Minería de Datos como fase del proceso KDD. Relación con otras disciplinas.
− Recuperación de información vs recuperación de datos. Proceso de recuperación de información.
− Preparación de Datos. Metadatos. Análisis de la información de entrada. Construcción y análisis de representaciones gráficas. Limpieza y transformación. Transformación y creación de atributos. Discretización y Numerización, Normlización de rango, escalado y centrado. Exploración mediante visualización y selección de datos.
− Técnicas de Minería de Datos. Extracción de Patrones. Introducción. Tareas y Métodos. Tareas predictivas y descriptivas. Aprendizaje supervisado y aprendizaje no supervisado. La Minería de Datos y el aprendizaje inductivo. Comparación de las técnicas de Minería de Datos.
− Arboles de decisión. Métricas de selección de atributos. Entropía. Ganancia de Información. Tasa de Ganancia. Indice Gini. Poda y Sobreajuste. Algoritmos Id3, C4.5 y Random Forest. Construcción de árboles para grandes volúmenes de datos.
− Reglas de clasificación. Partición vs cobertura. Métodos ZeroR, OneR, PRISM y PART. Métricas de una regla: soporte, cobertura, confianza, interés y convicción.

Área Específicas

Aplicaciones de Inteligencia de Datos

Carga horaria total: 40 horas
Contenidos:
1. Customer Relationship Management (CRM). a. Uso de la minería de datos para descubrir patrones y relaciones del comportamiento del consumidor.
2. Predicción de precios o rendimientos usando redes neuronales. a. Hipótesis de Mercados Eficientes y sus limitaciones. b. Uso de datos a distintas frecuencias para la predicción.
3. Análisis de la memoria de largo plazo en series temporales financieras.
4. Uso de Inteligencia de datos para la mejora de las estadísticas públicas: a. Datos de redes sociales. b. Datos de uso de celulares. c. Colaboración ciudadana mediante aplicaciones móviles. 5. Inteligencia de datos en ciencias de la salud: a. Las redes sociales como arma para la planificación en la lucha contra epidemias. b. Datos online que nos permiten inferir el perfil de salud de una población.

Procesamiento para Grandes Datos

Carga horaria total: 30 horas
Contenidos:
Unidad 1: Conceptos básicos de paralelismo Procesamiento paralelo. Arquitecturas paralelas. Servidores, Clusters y Cloud. Modelos de programación. Métricas. Herramientas.
Unidad 2: Introducción a Big data. Relación con Paralelismo Fundamentos. Objetivos. Modelos de datos y modelos de procesamiento. Paradigma Map-Reduce. Apache Hadoop. Por qué paralelismo sobre Big Data?

Unidad 3: Sistemas de almacenamiento para Big Data Sistemas de archivos distribuidos. Clasificación. Apache HDFS. Bases de datos relacionales. Bases de datos NoSQL. Hive, Shark, MongoDB, Cassandra.
Unidad 4: Procesamiento paralelo para Big Data en la Nube
Los alumnos deberán desarrollar un proyecto que involucre construir una infraestructura en la nube y combinar los conocimientos adquiridos.