

Maestría en Analítica de Datos

Descripción de espacios formativos del núcleo básico

Fundamentos de estadística para analítica de datos

La necesidad de evaluar los fenómenos sociales, económicos, culturales, ambientales, endémicos y demás, a través de una mirada analítica, hace que los profesionales, científicos e investigadores de diversas áreas del conocimiento recurran a una gran variedad de herramientas estadísticas. Entre estas se pueden mencionar las técnicas del análisis exploratorio de datos, estadística inferencial y multivariada y algunas relacionadas con técnicas de muestreo. La finalidad de este curso es abordar algunas de estas técnicas vitales en el análisis de datos a partir de problemas reales. El curso se desarrollará en las clases teóricas en las salas de cómputo donde se trabajará en Python y el *software* estadístico R.

Métodos estadísticos para analítica de datos

Los métodos de análisis multivariado se ocupan de identificar y señalar de forma resumida comportamientos atípicos, tendencias y/o patrones de comportamiento de una gran variedad de fenómenos sociales, biológicos, económicos, entre otros. De forma conjunta, permiten interpretar muchas variables en conjuntos con gran volumen de datos, reducir sus dimensiones originales, agrupar datos de acuerdo con características representativas, encontrar relaciones multidimensionales entre ellas; en síntesis, se encarga de la comprensión e interpretación de información que requiere tratamiento analítico multidimensional para que sus hallazgos propicien una toma de decisiones objetiva, en situaciones afectadas por la incertidumbre.

En el curso se construyen los fundamentos teóricos de los principales métodos de la estadística exploratoria multidimensional y se trabajan sus aplicaciones empleando un *software* especializado de uso libre y comercial (Python y R). Las prácticas se orientan a la aplicación de los métodos a situaciones reales que requieren apropiación de los contextos específicos, a partir de la revisión de artículos, para realizar interpretaciones adecuadas de los datos y del fenómeno en general. Para el aprendizaje de la aplicación de los métodos, el estudiante deberá realizar talleres utilizando el paquete estadístico.

Por otro lado, el uso de herramientas estadísticas junto a métodos computacionales ha permitido la creación de técnicas que se constituyen en lo que hoy se llama el aprendizaje

de máquina. En este curso se abordarán las diferentes técnicas de aprendizaje supervisado y no supervisado y su aplicación a situaciones particulares.

Programación para analítica de datos

El análisis de datos, especialmente aquel aplicado a grandes conjuntos o colecciones, requiere del uso de las herramientas de la computación para ser factible y efectivo en las aplicaciones actuales. Ello implica la necesidad permanente de desarrollar nuevos métodos y procesos de gestión y análisis de los datos disponibles para cada aplicación o uso, de modo que el resultado de dicho procesamiento conduzca a la obtención de resultados pertinentes y oportunos para la toma de decisiones.

En este sentido, el curso de programación para analítica de datos se enfoca en el desarrollo de competencias y habilidades técnicas y conceptuales para la construcción de algoritmos consistentes y eficaces al enfrentar problemas de exploración, caracterización y análisis de datos mediante los lenguajes de programación adecuados y en función de la extracción de información relevante en su contexto particular.

Bases de datos

En todas las aplicaciones de análisis de datos a escala empresarial o de investigación y desarrollo tecnológico, es deseable que los datos que soportan el proceso de análisis se encuentren organizados de tal manera que su registro, acceso, navegación y extracción sea tan eficiente como la estructura intrínseca de los datos lo permita, incluso si dicha estructura no se encuentra previamente establecida (datos no estructurados).

Este curso propone una aproximación formal orientada a la solución de problemas, a los contenidos más relevantes en el campo de las bases de datos, tanto en su formulación tradicional encaminada al álgebra relacional y su aplicación en los sistemas basados en el lenguaje SQL, así como los enfoques más recientes en los que se requieren estrategias alternativas dado el creciente tamaño de las colecciones de datos disponibles para cada aplicación. En este curso se estudiarán metodologías de procesos ETL (extracción, transformación y carga) además de la introducción a modelos No-SQL.

Big data

Con la generación de grandes cantidades de datos provenientes de sensores, nuevas técnicas de adquisición de datos, modelamiento y simulación de fenómenos, redes sociales etc., es vital la generación de nuevos algoritmos y tecnologías de administración de datos, que permitan desarrollar de manera óptima procesos claves en el tratamiento de los datos, a saber, procesos de extracción, interpretación de datos, entre otros.

En este curso se tratarán técnicas de análisis y gestión de datos que permitan la toma de decisiones en entornos distribuidos, la inteligencia empresarial y el descubrimiento científico. Se presentarán los principales conceptos y aplicaciones de Big Data Analysis y mediante el aprendizaje basado en problemas, los estudiantes integrarán sus conocimientos adquiridos en los cursos de fundamentos y métodos estadísticos y bases de datos aunado a los conocimientos computacionales para diseñar e implementar marcos y sistemas de información destinados a manejar grandes cantidades de datos y generar conocimiento útil para la toma de decisiones en las organizaciones.

Adicionalmente, se trabajará en el sistema Hadoop y la arquitectura Mapreduce para almacenamiento de información y procesamiento de base de datos de manera rápida y eficiente.

Inteligencia de negocios

En la actualidad, las organizaciones necesitan aprovechar la información extraída de los grandes volúmenes de datos. Para lograr este propósito es necesario contar con herramientas que le permitan utilizar y analizar la información para dar soporte a la toma de decisiones y, en este sentido, definir estrategias de negocios que lleven al cumplimiento de las metas planteadas.

Este curso desarrollará en el estudiante las competencias necesarias para generar conocimiento a partir de los grandes volúmenes de datos usando técnicas de big data, con el fin de dar soporte a la toma de decisiones en las organizaciones. Igualmente, se abordarán los sistemas de información relacionados con inteligencias de negocios, sistemas de soporte a la decisión y su aplicación en algunos aspectos relevantes de las organizaciones como son el *marketing*, los CRM y los sistemas de control. Además, se abordará la arquitectura del Data Warehouse; sus usos y aplicaciones en la organización.

Seminario

El seminario de analítica de datos abordará los siguientes aspectos:

1. En la primera parte se abordará, de manera general, la función de la analítica en otras disciplinas, como su importancia, desafíos y dificultades sin entrar en detalle al procedimiento de la aplicación respectiva, con el fin de tener un panorama de la importancia y uso del procesamiento y análisis de grandes volúmenes de datos en contextos prácticos. Este espacio le ampliará la visión al estudiante y generará ideas para su posterior trabajo de grado.
2. Una vez estudiado el panorama general de la analítica de datos a partir de otras disciplinas, surge un aspecto fundamental: ¿cuál el papel de la ética en el big data? El manejo de información personal trae consigo unos interrogantes y desafíos que se abordarán en esta parte del seminario.
3. Por último, se abordará el marco legal de la protección de datos personales desde la legislación colombiana y otros países de Latinoamérica. El desarrollo de este espacio formativo estará basado en exposiciones por parte de los estudiantes, del profesor e invitados expertos en algunos temas.

Maestría en Analítica de Datos

Descripción de espacios formativos proyectos

Proyecto de investigación I

En este espacio se propende por que el estudiante desarrolle las competencias necesarias para la elaboración del anteproyecto de investigación, teniendo en cuenta la definición de una problemática sobre la cual se realizará la investigación. El estudiante deberá establecer el estado del arte, la fundamentación teórica y el diseño metodológico que le permitan abordar el problema de investigación.

Proyecto de investigación II

Este espacio tiene como finalidad el desarrollo de la fundamentación metodológica y demás elementos de un proyecto de investigación.

Proyecto de investigación III

Este espacio tiene como finalidad la entrega del informe final de investigación y la estructuración del artículo que será sometido a una revista indexada para su posterior publicación.

Proyecto de profundización I

El estudiante deberá desarrollar las competencias necesarias para la elaboración del anteproyecto de investigación aplicada, es decir, la formulación de una problemática a la luz de los conceptos y herramientas adquiridas en los demás espacios formativos para dar solución a una situación particular del mundo real. Igualmente, el estudiante deberá establecer el estado del arte, la fundamentación teórica y el diseño metodológico que le permitan abordar el problema planteado.

Maestría en Analítica de Datos

Descripción de espacios formativos optativas

Computación avanzada para analítica de datos

Se propone una introducción a la computación de alto desempeño orientada a la analítica de datos y campos afines. En su desarrollo se considerarán los aspectos más relevantes de computación numérica y el diseño y construcción de *software* científico de mayor complejidad, según los estándares y marcos de trabajo de cómputo avanzado, aplicado a la analítica de datos y el big data.

El lenguaje de programación Python será la herramienta principal del curso, dada su amplia difusión, el tamaño de su comunidad y la implementación de múltiples bibliotecas de análisis y procesamiento de datos en dicho lenguaje. También se usarán otros lenguajes como R, C++ y Java para casos específicos de uso y aplicación.

Del mismo modo, se presentarán las principales arquitecturas de cómputo para el procesamiento paralelo y distribuido como soporte al tratamiento y análisis de grandes volúmenes de datos.

Diseño de experimentos

En los procesos de investigación que requieren realizar experimentos, es una práctica frecuente que estos se realicen durante su desarrollo. Aquí las decisiones se parametrizan según la experticia o la intuición de quienes toman las decisiones. Este proceso de ensayos y error le imprime al estudio un fuerte sesgo sobre las conclusiones y decisiones finales. Sin embargo, si se sigue un plan experimental adecuado que garantice una buena respuesta a las interrogantes planteadas, donde se haga un análisis riguroso de toda la información obtenida y que se considere la variación, será posible generar mejoras significativas en los experimentos, y, así, fortalecer el proceso. El diseño estadístico de experimentos es precisamente la forma más eficaz de llevar a cabo este proceso, determinando cuáles pruebas y de qué manera se debe plantear la experimentación para obtener estadísticas objetivas que respondan a las necesidades del mismo estudio.

Fundamentos de minería de datos aplicados a publicidad

Este programa capacita a los estudiantes en la integración de conceptos y herramientas de análisis para el uso de los datos generados por la web desde la minería. Este curso es innovador en el campo publicitario porque está orientado a ver, entender y visualizar de una manera no convencional el uso de los datos, es decir, permite conocer información no trivial ni evidente, lo que redundará en decisiones estratégicas innovadoras.

Por otra parte, el curso acerca al estudiante a la realidad de la industria, en tanto las decisiones empresariales en la actualidad se sustentan en los datos y no solamente en la intuición.

Este curso tiene una naturaleza teórico-práctica con un dispositivo de **estudio de caso**. El proceso educativo se orienta en sala de sistemas, de manera que se presenten los conceptos teóricos del análisis de datos y puedan ser aplicados en cada sesión, generando un ambiente idóneo para el procesamiento de datos, generación de información y producción de conocimiento que ayude a la toma de decisiones estratégicas.

Se trabaja un caso real y se configura partiendo de un hecho o conjunto de hechos determinados que requieren ser atendidos, por tanto, se identificará el contexto, los relatos, informaciones, problemáticas, relacionando conceptos y desarrollando preguntas que refieran a posibles acciones para el caso de estudio.

Reconocimiento de patrones

El tratamiento y análisis de conjuntos de datos de mayor complejidad, bien sea por su gran tamaño, su baja estructuración o la heterogeneidad de sus fuentes, exige el uso de estrategias algorítmicas más elaboradas que permitan extraer la información relevante para cada aplicación a partir de especificaciones definidas de manera general sobre las características de las piezas, segmentos o paquetes de información contenidos de manera implícita en los datos de entrada. Dichas especificaciones cuantitativas implican una etapa de exploración y caracterización de los patrones que deben ser obtenidos a partir del conocimiento existente *a priori* sobre el problema en cuestión y los indicadores estadísticos relacionados.

De este modo, los objetos, regularidades o secuencias de información podrán ser hallados dentro del conjunto original de datos a partir del uso de dicha caracterización (mediante rangos de valores, categorías, distancias, etc.), en procesos de exploración y búsqueda complejas por medio de algoritmos capaces de identificar la presencia de los patrones especificados en los datos disponibles. En el estado del arte esto suele implicar el uso de técnicas estadísticas avanzadas o de métodos modernos de aprendizaje de máquina supervisado o no-supervisado, según el grado de conocimiento previo sobre las clases o categorías bajo estudio.

En este curso se presentarán los fundamentos conceptuales y operativos del reconocimiento de patrones enfocado en la búsqueda de información relevante en grandes volúmenes de datos, incluyendo métodos generativos como la teoría de la decisión de Bayes y otras técnicas relacionadas con la estimación de parámetros y densidad de probabilidad. También se presentan métodos discriminativos, como la clasificación de vecino más próximo, las máquinas de soporte vectorial, las redes neuronales artificiales modernas (*deep learning*) y la combinación de clasificadores.

Series de tiempo

Al finalizar el curso el estudiante deberá estar en capacidad de manejar y aplicar conceptos necesarios en el estudio de técnicas modernas para el análisis básico de series de tiempo, como los métodos de suavizamiento, metodología ARIMA y ARIMAX (Box- Jenkins). Al interiorizar dichas técnicas, el estudiante entenderá, criticará y seleccionará modelos y técnicas usadas y reportadas en revistas de talla internacional, así como seleccionar modelos o técnicas alternativas según la capacidad de pronóstico de las mismas.