TEC | Tecnológico de Costa Rica

Programa de capacitación

Ciencias de los Datos

CETIEC

Centro de Especializaciones TI de la Escuela de Computación TEC

Introducción

Este nuevo plan de estudios se trata de cubrir transversalmente grandes áreas como (Tecnologías de la Información, Minería de Datos, Aprendizaje Estadístico, Toma de Decisiones, Economía, Empresa e Inteligencia de Negocio) de modo que los titulados quedarían en la mejor disposición para su futura actividad, tanto a nivel profesional en empresas de diversos sectores como a nivel investigador en centros o equipos privados o públicos. El objetivo fundamental del título propuesto es preparar profesionales con una formación transversal y muy versátil, que abarque un amplio espectro y de fácil adaptación a entornos de trabajo significativamente diferentes. En la actualidad, existe una gran demanda de profesionales con capacidades en este ámbito. Dependiendo de la intensidad de su especialización en las tres áreas fundamentales (Computación, Predicción o Negocio) encontramos distintos tipos de nuevos profesionales y científicos. Programadores Big Data (o “Data Developers”), Analistas o Científicos de Datos (“Data Analysers” o “Data Scientists”), y profesional de empresa experto en datos (“Data Businessman”) son algunas de las nuevas profesiones surgidas alrededor del Big Data (Harris, H., Murphy, S., Vaisman, M. “Analyzing the Analyzers. An Introspective Survey of Data Scientists and Their Work.” O'Reilly, 2013).

Algunos hechos fácticos sobre datos producidos en la actualidad.

1. A diario en el mundo se generan 2.5 trillones de bytes de información.
2. El 90% de los datos a nivel mundial se han creado solamente en los últimos 2 años
3. Empresas como Google cuenta 600 personas dedicadas al estudio del Big Data
4. The Economist Intelligence Unit (EIU) entrevistó a 600 ejecutivos globales y el 54% de los empresarios estadounidenses aseguraron que encontrar a los profesionales adecuados para un exitoso proyecto de Big Data es el obstáculo más importante para no hacerlo.

1. El mayor número de profesionales en el área de Big Data provienen de países asiáticos
2. De 2012-2013 más del 60% de los artículos de opinión de tecnología avanzada hablan de Big Data como la nueva estrategia indispensable para las empresas de cualquier sector, declarando, poco menos , que aquellos que no se sumen a este nuevo movimiento, quedarán obsoletos.

Requisitos de Ingreso

- Personas con algún tipo de grado en Estadística, Ingeniería en Computación, Ingeniería en Sistemas de Información, Matemáticas, Economía, Física.
- Es importante que las personas cuenten con experiencia previa en programación de algoritmos.

Perfil de Ingreso

- Programa dirigido a personas egresadas o que se desarrollen en alguna de las siguientes áreas:
  - Ingeniería en Computación
  - Ingeniería en Sistemas
  - Estadística
  - Matemática
  - Ingeniería

Perfil de Salida

- Una persona especialista en ciencias de los datos, es aquella capaz de crear y utilizar algoritmos de aprendizaje automático, diseñar modelos, determinar el dominio aplicable de cada uno de ellos; de forma tal que pueda realizar análisis de datos y análisis cuantitativos, predicciones e interpretación.

Sobre el programa

- Programa de Capacitación compuesto por 5 módulos, impartidos a lo largo de un año.
- Cada módulo se impartirá de forma bimestral
- En el Desarrollo de cada uno de los módulos, se utilizará como principal y de manera trasversal, un lenguaje de alto uso en la indistria, el cuál se verá complementado con algunos otros lenguajes de prototipado.
- Entre las técnicas a desarrollar durante el desarrollo de los cinco módulos, se encuentran:
  - Técnicas de procesamiento de lenguaje natural

	○ Técnicas de procesamiento de imágenes ○ Técnicas de procesamiento de sonido ○ Temas de minería de opinión ● Todos los módulos del programa, estarán desarrollados bajo una metodología basada en proyectos y con un alto contenido en visualización.
Inversión por estudiante	● Cada módulo tendrá un costo de $500
Secuencia de Módulos	1. Matemática para ciencias de los datos 2. Aprendizaje automático 3. Estadística para ciencias de los datos 4. Big Data 5. Inteligencia de negocios y minería de datos
Aprobación del Programa	Con la aprobación de cada uno de los cursos, se otorgará un certificado de aprovechamiento; sin embargo para obtener el certificado del Programa de Ciencias de los Datos, es necesario cumplir con la aprobación de los 5 módulos.

Módulos del Programa

Nombre del Curso	**Matemática para Ciencias de los Datos**
Descripción	El curso introducirá conceptos matemáticos fundamentales en las ciencias de los datos, como el cálculo multivariable, el álgebra lineal y el cálculo matricial, sirviendo como base para plantear modelos de clasificación y regresión de datos en cursos posteriores.
Tipo de Curso	Aprovechamiento Tipo teórico/práctico Para obtener el certificado correspondiente, es necesario tener una asistencia efectiva de más del 80% a las lecciones, y sus evaluaciones con un promedio mayor o igual a 70.
Cantidad de Horas Lectivas	32 Horas lectivas presenciales
Objetivos	**Objetivo General:** Al finalizar el curso, el estudiante contará con todas las bases matemáticas necesarias para ingresar al mundo de Ciencias de Datos. **Objetivos Específicos:** Al finalizar el curso el estudiante será capaz de:

- Modelar problema en términos matemáticos, y múltiples variables.
- Encontrar la solución óptima a un problema planteado según un modelo matemático, utilizando herramientas básicas del cálculo multivariable.

Metodología de la enseñanza

Se abordarán clases magistrales por parte del profesor, como introducción a las actividades y conceptos que se desarrollan en cada sesión.
El curso utilizará una metodología de Aprender Haciendo, basado en el desarrollo de proyectos de forma tal que mediante el desarrollo de casos de estudio y proyectos en el laboratorio, se pueda afianzar los conocimientos adquiridos durante el transcurso de las diferentes lecciones.

Contenidos del programa

1. **Algebra lineal:**
   a. Operaciones en vectores y matrices
   b. Sistemas lineales e independencia lineal.
   c. Autovectores y análisis de componentes principales.
2. **Cálculo matricial**
   a. Introducción al cálculo multivariable
   b. Funciones multivariable
   c. Derivadas parciales
   d. Integrales de superficie
   e. Introducción al cálculo matricial
3. **Optimización**
   a. Optimización convexa y no convexa
   b. Optimización de funciones con restricciones y sin restricciones
   c. Optimización de funciones diferenciables y no diferenciables

d. Algoritmos estocásticos de optimización

TEC | Tecnológico de Costa Rica

Nombre del Curso	**Aprendizaje Automático**
Descripción	El curso tiene como objetivo desarrollar las habilidades necesarias para la adecuada representación de problemas de reconocimiento de patrones, así como la implementación de soluciones concretas a problemas reales. El estudiante aprenderá sobre la elección e implementación de diferentes algoritmos de clasificación y regresión, populares en el aprendizaje automático. El curso pretende aportar las bases del reconocimiento de patrones y el aprendizaje automático, para además generar curiosidad en el estudiantado para para participar más adelante en proyectos de investigación científica, así como de valor práctico para la industria, en temas de alta complejidad.
Tipo de Curso	Aprovechamiento Tipo teórico/práctico Para obtener el certificado correspondiente, es necesario tener una asistencia efectiva de más del 80% a las lecciones, y sus evaluaciones con un promedio mayor o igual a 70.
Cantidad de Horas Lectivas	32 Horas lectivas presenciales

Objetivos

Objetivo General:

Al finalizar el curso el estudiante habrá adquirido las destrezas para construir un sistema de reconocimiento de patrones en general, a través de la experiencia del reconocimiento de patrones basado en imágenes

Objetivos Específicos:

Al finalizar el curso el estudiante será capaz de:

- Analizar los datos de entrada de un sistema de de aprendizaje automático
- Diseñar las etapas de preprocesamiento y extracción de características relevantes, en un sistema de reconocimiento de patrones.
- Implementar un clasificador bajo un enfoque de aprendizaje automático, con base en las características extraídas de los datos de entrada del sistema.
- Evaluar la efectividad de un sistema de aprendizaje automático usando una métrica cuantitativa adecuada.

Metodología de la enseñanza	Se abordarán clases magistrales por parte del profesor, como introducción a las actividades y conceptos que se desarrollan en cada sesión. El curso utilizará una metodología de Aprender Haciendo, de forma tal que mediante el desarrollo de casos de estudio y proyectos en el laboratorio, se pueda afianzar los conocimientos adquiridos durante el transcurso de las diferentes lecciones.

Contenidos del programa

1. **Introducción al Reconocimiento de Patrones**
    a. Ejemplos y aplicaciones.
    b. Etapas de un sistema de reconocimiento de patrones.
    c. Tipos de aprendizaje.
1. **Etapa de preprocesamiento** Categorización y discretización de datos
    a. Normalización
    b. Eliminación de sesgos, redundancia y ruido
2. **Etapa de extracción de características**
    a. Descriptores e invariantes
3. **Etapa de clasificación:** Métodos de clasificación supervisada y no supervisada.
    a. Ajuste polinomial de curvas.
        i. Modelos paramétricos lineales de regresión: mínimos cuadrados y mínimos cuadrados regularizados.
        ii. Selección del modelo (sobreajuste)y validación cruzada.
        iii. La maldición de la dimensionalidad.
    b. Métodos supervisados.
        i. PCA y K-vecinos más cercanos.
        ii. Mínimos cuadrados.
        iii. Discriminante lineal de Fisher.
        iv. Perceptrón.
        v. Análisis de componentes principales (PCA, por sus siglas en inglés).
        vi. Redes neuronales de retropropagación y con entrenamiento de descenso de gradiente.
        vii. Redes convolucionales

viii. Máquinas de soporte vectorial.

c. Métodos no supervisados.
   i. Algoritmo BSAS
   ii. Algoritmo K-medias.
   iii. Algoritmos de aprendizaje competitivo.

4. **Validación**
5. **Evaluación de modelos**
   a. AUC ROC
   b. Precision / Recall
   c. Particiones de datos para entrenamiento y pruebas
   d. Retroalimentación y actualización de modelos

TEC | Tecnológico de Costa Rica

Nombre del Curso	**Estadística para Ciencias de Datos**
Descripción	El curso introducirá a los estudiantes conceptos de estadística descriptiva, pruebas de hipótesis y experimentación usando herramientas estadísticas. Los conceptos estudiados en el curso permitirán proponer preguntas en el ámbito de las ciencias de los datos, y responderlas con rigurosidad estadística.
Tipo de Curso	Aprovechamiento Tipo teórico/práctico Para obtener el certificado correspondiente, es necesario tener una asistencia efectiva de más del 80% a las lecciones, y sus evaluaciones con un promedio mayor o igual a 70.
Cantidad de Horas Lectivas	32 Horas lectivas presenciales
Objetivos	Al finalizar el curso, el estudiante será capaz de: Utilizar herramientas estadísticas básicas para modelar, resolver e interpretar problemas en ámbitos de aplicación y utilizar herramientas computacionales para realizarlo. Objetivos específicos:

	-Utilizar herramientas de muestreo, diseñar un experimento simples y recolectar datos de manera apropiada. - Analizar datos de forma descriptiva e interpretar los resultados. - Utilizar técnicas inferenciales adecuadas para cada problema específico -Utilizar un lenguaje de programación de alto nivel para implementar modelos estadísticos y analizar los resultados. - Implementar las herramientas estadísticas estudiadas en la resolución de problemas en distintas áreas
Metodología de la enseñanza	Se abordarán clases magistrales como introducción a las actividades que se desarrollan en cada sesión. El curso utilizará una metodología de Aprender Haciendo, de forma tal que se pueda desarrollar el o los proyectos para ejemplificar el aprendizaje

Contenidos del programa

1. **Análisis de datos**
    a. Medidas de posición y variabilidad
    b. Relaciones entre variables cualitativa-cualitativa (tablas de contingencia), cualitativa-cuantitativa (agregación, estadísticas por grupo), cuantitativa-cuantitativa)
    c. Distribucion de Frecuencia
    d. Histogramas, diagramas de caja por grupo, gráficos de dispersión, coeficientes de correlación (Pearson, Spearman)
    e. Entropía y teoría de información
    f. Correlación y causalidad
2. **Correlación y causalidad**
3. **Distribuciones comunes y propiedades**
    a. Discretas: Binomial y Poisson
    b. Continuas: Normal, Beta, Exponencial
4. **Pruebas de hipótesis**
    a. Paradigma frecuentista vs. Bayesiano
    b. Intervalos y p values
    c. Pruebas de normalidad, T student, Kolmogorov Smirnoff
    d. Pruebas no paramétricas: Wilcoxon
5. **Series de tiempo**
    a. Modelos ARIMA
    b. Cadenas de Markov
    c. Redes de Elmann

6. **Experimentación en ciencias de los datos**
    a. Diseños de experimentos
    b. Muestreo: Bagging, remuestreo
    c. A/B Testing
    d. Pruebas de hipótesis en experimentos

7. **Experimentación**
    a. Diseño de experimentos
    b. Poblaciones y muestreo
    c. A/B Testing

d. Prueba de hipótesis en experimentos

TEC | Tecnológico de Costa Rica

Nombre del Curso	Big Data
Tipo de Curso	Aprovechamiento Tipo teórico/práctico Para obtener el certificado correspondiente, es necesario tener una asistencia efectiva de más del 80% a las lecciones, y sus evaluaciones con un promedio mayor o igual a 70.
Cantidad de Horas Lectivas	32 Horas lectivas presenciales
Objetivos	· **Objetivo General:** Entender y aplicar técnicas de análisis de grandes cantidades de datos para la resolución de problemas concretos a través de tecnologías de manipulación, extracción y sintetización estadística. **Objetivos Específicos:** ● Aplicar bibliotecas para la transformación de datos a gran escala para poder sintetizar el conocimiento para futuro análisis. ● Aplicar técnicas de análisis de datos para extraer patrones que mejoren el entendimiento de un problema concreto. ● Aplicar técnicas para aprendizaje automatizado de patrones, basado en datos existentes, para mejorar la certeza de la solución aplicada a problemas concretos.

Metodología de la enseñanza

Se abordarán clases magistrales por parte del profesor, como introducción a las actividades y conceptos que se desarrollan en cada sesión.
El curso utilizará una metodología de Aprender Haciendo, de forma tal que mediante el desarrollo de casos de estudio y proyectos en el laboratorio, se pueda afianzar los conocimientos adquiridos durante el transcurso de las diferentes lecciones.

Contenidos del Progama

1. **Fuentes y repositorios de datos**
2. **Diferencias entre fuentes**
    a. SQL vs NoSQL
    b. Archivos simples o distribuídos
    c. Schema vs No Schema
    d. Escalabilidad
3. **Procesamiento de fuentes (data frames)**
    a. Uniones
    b. Agregaciones
    c. Transformaciones
    d. Filtros
4. **Procesamiento de atributos**
    a. Selección por importancia y redundancia
    b. Normalización
    c. Discretización
5. **Organización de datos procesados**
    a. Tablas de estadísticas
    b. Data warehousing
    c. Vistas (dinámicas y materializadas)
    d. Escalabilidad
6. **Análisis de datos**
    a. Estadística descriptiva
    b. Correlación y causalidad
    c. Simulación
    d. Generación de reportes
        i. Consultas a bajo nivel
        ii. Business Intelligence

iii. Cubos
iv. Entities / Facts
v. Híbridos

7. **Uso de modelos de aprendizaje automático en big data**
    a. Sin supervisión
        i. Clustering
    b. Supervisado
        i. Regresión lineal
        ii. Árboles Aleatorios

TEC | Tecnológico de Costa Rica

Nombre del Curso	**Minería de Datos e Inteligencia de Negocios**
Descripción	Este curso pretende estudiar las técnicas de minería de datos y de inteligencia de negocios que persiguen como objetivo esencial el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en conjuntos de datos que pueden tener dimensión elevada. Estas técnicas tienen por objeto descubrir patrones, perfiles, tendencias y otras relaciones presentes en la información, pero ocultas si no se trata adecuadamente. Cuando la dimensión de la información es muy alta, proviene de múltiples orígenes y no está estructurada.
Tipo de Curso	Aprovechamiento Tipo teórico/práctico Para obtener el certificado correspondiente, es necesario tener una asistencia efectiva de más del 80% a las lecciones, y sus evaluaciones con un promedio mayor o igual a 70.
Cantidad de Horas Lectivas	32 horas presenciales

Objetivos

· **Objetivo General**

Capacitar a los estudiantes para alcanzar una mentalidad crítica y analítica dentro de la empresa, mediante el conocimiento de los diferentes sistemas de información de empresa, los métodos y técnicas de análisis de datos, la formulación de preguntas e hipótesis, la obtención y la visualización de datos para conclusiones útiles en la toma de decisiones.

Objetivos Específicos

- Diseñar el proceso de obtención y preparación de datos a alto nivel, en las diferentes fuentes, tipos de datos para métodos cuantitativos y cualitativos de análisis.
- Conocer las técnicas de análisis multidimensional (estructurado y no estructurado (durante todo el proceso de construcción y mantenimiento) de alto nivel (tableau, powerBI).
- Conocer técnicas, herramientas y algoritmos de análisis en minería de datos con el fin de adquirir capacidades superiores de comprensión de problemas, formulación de hipótesis, interrogación e interpretación de la información en procesos empresariales como: gestión económico-financiera, marketing y ventas y operaciones y logística. Modelar los procesos de negocio
- Formulación de preguntas de ciencias de ciencias de los datos
- Construir informes y cuadros de mando para la toma de decisiones de los empleados y directivos con técnicas y herramientas de visualización de datos.

Metodología de la enseñanza	Se abordarán clases magistrales por parte del profesor, como introducción a las actividades y conceptos que se desarrollan en cada sesión. El curso utilizará una metodología de Aprender Haciendo, de forma tal que mediante el desarrollo de casos de estudio y proyectos en el laboratorio, se pueda afianzar los conocimientos adquiridos durante el transcurso de las diferentes lecciones.

Contenidos del programa

1. **Fundamentos de Inteligencia de negocio**
    a. Concepto de inteligencia de negocio.
    b. Tipos de datos explotados.
    c. Recogida y preparación de datos.
    d. Metodologías (KDD, CRISP-DM)
    e. Casos de Estudio
2. **Sistemas de inteligencia de negocio: Data Warehouse (DW)**
    a. Diseño del almacén de datos
    b. Integración, limpieza y transformación del almacén de datos
    c. Explotación y administración de sistemas de DW
3. **Sistemas de inteligencia de negocio: Data Lake**
    a. Diseño y construcción del lago de datos
    b. Explotación y administración de sistemas estructurados y no estructurados
4. **Análisis y minería de datos**
    a. Minería de datos complejos (espacial, temporal, web mining)
    b. Modelos de clasificación
    c. Modelos de relación
    d. Análisis de textos
5. **Visualización de datos**
    a. Visualización (Gráficos, Mapas, Dashboard)
    b. Evaluación
    c. Conclusiones