Universidad Peruana Unión

Carret. Central Km. 19.5 Ñaña. Telf. 01-6186300 Fax 01-6186-339 Casilla 3564 Lima 1,Peru

S Í L AB O

I. Información General

1.	**Facultad:** **Unidad de Posgrado:** **Asignatura:**	Escuela de Posgrado Ingeniería y Arquitectura Big data
2.	**Ciclo:**	4
3.	**Número de Créditos:**	3
4.	**Horas Teóricas:**	9
5.	**Horas Prácticas:**	22
6.	**Nota Aprobatoria:**	14
7.	**Nombre del profesor:**	Harvey Alférez, Ph.D. http://www.harveyalferez.com, https://icd.um.edu.mx, harveyalferez@um.edu.mx
8.	**Semestre Académico:**	Cuarto
9.	**Fecha (Inicio - Final):**	6-9 de enero y 17 de febrero de 2018

II. Sumilla

Cada día las organizaciones generan datos nuevos. Pero ¿podrías utilizar estos datos más eficientemente? En este curso descubriremos como convertir Datos Masivos (*Big Data* en inglés) en resultados de alto impacto para tu organización. Para lograrlo, aplicaremos técnicas avanzadas de Aprendizaje Automático (*Machine Learning* en inglés) en Datos Masivos. Para lograrlo, el curso comenzará con una descripción de Ciencia de Datos y del rol del Aprendizaje Automático y de los Datos Masivos en esta ciencia. Luego se introducirán diferentes aproximaciones de aprendizaje supervisado y no supervisado. A continuación, se estudiará Apache Spark, un motor de análisis unificado para Datos Masivos. En el proyecto final, se explotará la funcionalidad de MLib de Apache Spark. La asignatura “Big Data” es de naturaleza teórica práctica, perteneciente al área de especialización.

III. Competencia de la Asignatura

Crear programas computacionales que utilicen técnicas avanzadas de Aprendizaje Automático dentro del marco de Ciencia de Datos para descubrir conocimiento en Datos Masivos.

IV. Unidades de Aprendizaje

Unidad 1 Fundamentos de Ciencia de Datos, Aprendizaje Automático y Datos Masivos

Resultado de aprendizaje: Aprender los conceptos de ciencia de datos, aprendizaje automático y datos masivos.

Sesión	Fecha	Contenidos a Tratarse en el Aula	HA	HNP	Aprendizaje Autónomo	Estratégicas Metodológicas
1	6 de enero	Introducción a la ciencia de datos, el aprendizaje automático, y los datos masivos	3.5	4	- **Informe de Lectura 1 - Entregar el 7 de enero:** Informe de lectura de los siguientes artículos científicos (se encuentran en “primer reporte de lectura”: https://drive.google.com/file/d/1DeLNYWBINu7Sco4H2rIWnZvFd-n1AqMP/view?usp=sharing): 1) Dynamic Adaptation of Service Compositions with Variability Models; 2) Achieving Autonomic Web Service Compositions with Models at Runtime; 3) Detection of Melanoma through Image recognition and Artificial Neural Networks; 4) Proactive Control of Traffic in Smart Cities; 5) Software Architecture	Análisis reflexivo. Informe de lectura. Escritura de ensayo.

					Evolution in the Open World through Genetic Algorithms; 6) Interpreting the Geochemistry of Southern California Granitic Rocks Using Machine Learning; 7) Application of Data Science to Discover the Relationship between Dental Caries and Diabetes in Dental Records; 8) Prediction of Glaucoma through Convolutional Neural Networks; y 9) Dynamic Evolution of Simulated Autonomous Cars in the Open World Through Tactics. Cada informe de lectura será de dos páginas espacio sencillo. Incluirá el resumen del artículo y una reflexión al final de uno o dos párrafos. **- Ensayo - Entregar el 8 de enero:** Ensayo de 7 páginas espacio sencillo acerca del panorama actual de la ciencia de datos, el aprendizaje automático y los datos masivos en la industria. Debes utilizar por lo menos 10 referencias bibliográficas de artículos de revistas científicas en: https://ieeexplore.ieee.org, https://www.elsevier.com, http://springer.com	

Unidad 2 Aprendizaje Supervisado

Resultado de aprendizaje:

Sesión	Fecha	Contenidos a Tratarse en el Aula	HA	HNP	Aprendizaje Autónomo	Estratégicas Metodológicas
2	6 de enero	Presentación de K-Nearest Neighbors, Regresión Logística, Support Vector Machines, redes neuronales, y evaluación de la efectividad de clasificadores	3.5	3	**Laboratorio 1 - Entregar el 7 de enero:** Implemente mediante Sci-kit Learn o TensorFlow dos clasificadores, cada uno con un algoritmo de clasificación diferentes, con un conjunto de datos de: https://archive.ics.uci.edu/ml/index.php. Si fuera necesario, normalice los datos antes de construir el clasificador. Pruebe los resultados. Al introducir los siguientes valores de las características (variables independientes), el sistema deberá decir la clase (la variable dependiente). ¿Cuál clasificador dio el mejor	Resolución de problemas.

					resultado con respecto a la exactitud (*accuracy*)?	

Unidad 3 Aprendizaje No Supervisado

Resultado de aprendizaje:

Sesión	Fecha	Contenidos a Tratarse en el Aula	HA	HNP	Aprendizaje Autónomo	Estratégicas Metodológicas
3	7 de enero	K-Means	1.5	2.5	**- Laboratorio 2 - Entregar el 8 de enero:** Descubrir mediante la aplicación de K-Means en Weka por lo menos 3 patrones interesantes en un conjunto de datos que usted elija. En la presentación describirá los resultados del descubrimiento.	Resolución de problemas.

Unidad 4 Manejo de Datos Masivos Mediante Computación Paralela

Resultado de aprendizaje:

Sesión	Fecha	Contenidos a Tratarse en el Aula	HA	HNP	Aprendizaje Autónomo	Estratégicas Metodológicas
4	7 de enero	Una introducción a Apache Spark	2	4	**- Informe de Lectura 2 - Entregar el 8 de enero:** Informe de lectura de 5 artículos científicos que se encuentran en "segundo reporte de lectura": https://drive.google.com/file/d/1DeLNYWBINu7Sco4H2rlWnZvFd-n1AqMP/view?usp=sharing. Cada informe de lectura será de dos páginas espacio sencillo. Incluirá el resumen del artículo y una reflexión de uno o dos párrafos.	Informe de lectura. Análisis reflexivo.
5	7 de enero	Laboratorio 3 y Laboratorio 4 de Apache Spark	3.5	3	**- Laboratorios 3 y 4 - Entregar el 9 de enero:** Laboratorio 3 (ejercicios de New York University) y Laboratorio 4 (conteo de palabras – Berkely University) en Apache Spark. Realizaré preguntas individuales. Estos laboratorios están disponibles en: http://201.134.41.15:8889/tree/PhDCourse	Resolución de problemas.
6	8 de enero	Laboratorio 5 y Laboratorio 6 de Apache Spark	7	3	**- Laboratorios 5 y 6 - Entregar el 9 de enero:** Laboratorio 5 (análisis de log de servidor Web – Berkely University) y Laboratorio 6 (aplicación en Mlib) en Apache Spark. Estos laboratorios están disponibles en: http://201.134.41.15:8889/tree/PhDCourse. Realizaré preguntas individuales.	Resolución de problemas.

Unidad 5 Aplicación de Aprendizaje Automático en Datos Masivos

Resultado de aprendizaje:

Sesión	Fecha	Contenidos a Tratarse en el Aula	HA	HNP	Aprendizaje Autónomo	Estratégicas Metodológicas
7	9 de enero	Sesión de Asesoría Acerca del Proyecto Final de la Materia	7	2	Buscar un problema relevante para el proyecto final y el conjunto de datos para realizar los experimentos.	Aprendizaje basado en proyectos.
8	17 de febrero	Sesión Virtual de Presentación de los Proyectos Finales	3	70	**- Proyecto Final - Entregar el 17 de febrero:** Este proyecto se compone de dos partes: 1) 50%: Proyecto que utilice MLib de Apache Spark para descubrir patrones ocultos en un conjunto de datos abiertos de gran tamaño (una lista de ejemplo se encuentra en: https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#54b7fbeb54db). 2) 50%: Ensayo científico en inglés que describa en 3 páginas el problema que se resolvió y los descubrimientos que se encontraron con el programa creado.	Aprendizaje basado en proyectos. Elaboración de ensayo.

V. Asesoría, monitoreo y consultoría docente

Las asesorías, el monitoreo y la consultoría docente se realizarán durante las fechas y los horarios de clase de manera presencial.

VI. Aspectos y Técnicas de Evaluación

N°	Fecha	Unidades	Estrategia - Descripción	Ponderado
1.	8 de enero	1	**Ensayo.** Ensayo acerca del panorama actual de la ciencia de datos, el aprendizaje automático y los datos masivos en la industria	8.0000 %
2.	7 de enero	1	**Informe de lectura 1.** Informe de lectura de artículos científicos	10.0000 %
3.	8 de enero	4	**Informe de lectura 2.** Informe de lectura de artículos científicos	10.0000 %
4.	6-9 de enero	2, 3, 4, 5	**Laboratorios 1-6.** Realización y comprensión de laboratorios (cada laboratorio equivale a 7.0000 %)	42.0000 %
5.	17 de febrero	5	**Proyecto final (primera parte).** Proyecto que utilice MLib de Apache Spark para descubrir patrones ocultos en un conjunto de datos abiertos de gran tamaño	15.0000 %
6.	17 de febrero	5	**Proyecto final (segunda parte).** Ensayo acerca de los resultados de los experimentos	15.0000 %
			Total:	**100.0000 %**

VII. Bibliografía

- Ryza, S., Laserson, U.,Owen, S., and Wills, J. (2017). “Advanced Analytics with Spark”. 2nd, ed. O’Reilly.
- Garillot, F. and Maas, G. (2017). “Stream Processing with Apache Spark”. E. O’Reilly.
- Géron, A. (2017). “Hands-on Machine Learning with Scikit-Learn & TensorFlow”. O’Reilly.
- Erl, T., Khattak, W. and Buhler, P. (2016). “Big Data Fundamentals: Concepts, Drivers & Techniques”. Prentice Hall.
- Cielen, D., Meysman, A. and Ali, M. (2016). “Introducing Data Science.” Manning Publications.
- Brink, H., Richards, J. and Fetherolf, M. (2016). “Real-World Machine Learning.” Manning Publications.

- Kelleher, J.D., Mac Namee, B. and D'Aarcy, A. (2015). "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies". The MIT Press.
- Grus, J. (2015). "Data Science from Scratch". O'Reilly Media Inc.
- Marz, N. and Warren J. (2015). "Big Data: Principles and Best Practices of Scalable Realtime Data Systems". Manning Publications.
- Ryza, S., Laserson, U., Owen, S., and Wills, J. (2015). "Advanced Analytics with Spark". O'Reilly.
- Gurin, J. (2014). "Open Data Now". McGraw-Hill.
- Kitchin, R. (2014). "The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences". SAGE Publications.
- Richert, W. and Pedro-Coelho, L. (2013). "Building Machine Learning Systems with Python". Packt Publishing.
- Harrington, P. (2012). "Machine Learning in Action". Manning.
- Poole, D.L. and Mackworth, A.K. (2010). "Artificial Intelligence – Foundations of Computer Agents". Cambridge University Press.

VIII. Enlaces en internet

- http://ieeexplore.ieee.org/Xplore/home.jsp
- http://www.springer.com/
- http://www.acm.org/
- https://www.elsevier.com
- https://www.kdnuggets.com/
- https://spark.apache.org
- https://www.technologyreview.com
- https://www.data.gov
- http://archive.ics.uci.edu/ml/index.php
- http://www.harveyalferez.com