Construcción de un espacio virtual con material de apoyo y ejercicios reproducibles para las prácticas de Machine Learning e Inteligencia Artificial

Autor
Raul Alejandro Buitrago Castellanos

Universidad Distrital Francisco José de Caldas
Maestría en Ciencias de la Información y las Comunicaciones
Énfasis en Ingeniería de Software
Bogotá, Colombia
Septiembre 2020

Construcción de un espacio virtual con material de apoyo y ejercicios reproducibles para las prácticas de Machine Learning e Inteligencia Artificial

Autor
Raul Alejandro Buitrago Castellanos

Director
Cesar Andey Perdomo Charry
Msc. Ciencias de La Informacion y Las Comunicaciones

Universidad Distrital Francisco José de Caldas
Maestría en Ciencias de la Información y las Comunicaciones
Énfasis en Ingeniería de Software
Bogotá, Colombia
Septiembre 2020

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Esta pasantía de investigación bajo la dirección del ingeniero ***Cesar Andrey Perdomo Charry*** pretende realizar y/o consolidar material de apoyo a los docentes del grupo de investigación ***LASER*** adscrito a la *Universidad Distrital Fransisco José de Caldas* para las practicas y/o laboratorios de *Machine learning e Inteligencia Artificial* por medio de ejemplos que aborden las siguientes temáticas:

1. Estadística descriptiva,
2. Preparación de datos,
3. Regresión lineal (con una, y múltiples variables),
4. Aprendizaje supervisado
5. Aprendizaje no supervisado
6. Redes neuronales
7. Redes neuronales convolucionales

Para la realización de esta pasantía se utilizaran diversos conjuntos de datos disponibles en UC Irvine Machine Learning Repository, Kaggle, entre otros.

PALABRAS CLAVE

Algoritmos, Regresión lineal, Machine Learning, Inteligencia Artificial, Naive Bayes, Random Forest, Algoritmos de vecindad, Redes neuronales,

1 PROBLEMA DEL ENTORNO ACADÉMICO

En consecuencia de la evolución de las tecnologías de información y comunicaciones diversas prácticas y procesos se han adaptado a nivel industrial, social, cultural, entre otros. De los cuales se puede resaltar la transmisión del conocimiento por medio de canales virtuales redefiniendo la forma y el alcance en que se comparte la información haciendo uso de diferentes tecnologías y herramientas de comunicación [1].

A pesar que hay varias corrientes en cuanto a la percepción del término educación virtual donde en ocasiones se le confunde con la sustitución de un libro impreso y un aula de clase por un recurso electrónico. Hay quienes la interpretan como una alternativa en el camino de la formación ya que conlleva los siguientes desafíos [2]

- Contexto
- Definición
- Metodología
- Contenido
- Herramientas
- Expectativas

Teniendo en cuenta que la ciencia de los datos *(Data Science)* es un área de estudio del grupo de investigación e inteligencia computacional ***LASER*** [3] y la adopción el uso de las tecnologías de la información y comunicaciones. El grupo de investigación necesita generar un espacio virtual de apoyo a los estudiantes por medio de material didáctico y ejercicios reproducibles en temas de estadística, machine learning e inteligencia artificial.

1.1 PLANTEAMIENTO DEL PROBLEMA

En el mercado actualmente hay varias herramientas de formación virtual con diferentes enfoques de las que se destacan:

1. ***Coursera*** Plataforma con cursos en vídeo ofrecidos por más de 115 universidades e instituciones educativas del mundo [4].
2. ***Udemy*** Plataforma con más de 150000 cursos en vídeos hechos por particulares sobre diversos temas entre ellos el aprendizaje de las máquinas *machine learning*, inteligencia artificial y blockchain [5].

3. ***Jupyter*** Es una aplicación web de código abierto para la ejecución de secuencias de código en alrededor de 40 lenguajes de programacion y soporta tecnologías relacionadas con *Big Data* [6].

4. ***Google Colab*** Es una aplicación web de Google similar a Jupyter pero solo admite el uso del lenguaje de programación Python, no requiere configuración para usarlo y brinda acceso gratuito a recursos computacionales, incluidas GPU, pero tiene algunas restricciones excepto cuando se utiliza la versión de pago [7].

Esta pasantía puede convertirse en una aproximación al proyecto del grupo de investigación asociado a la construcción de un espacio virtual de apoyo a los estudiantes en el área de ciencia de los datos utilizando las tecnologías consultadas.

1.2 FORMULACIÓN DE PREGUNTA COMO SOLUCIÓN PROPUESTA

¿Cómo construir un espacio virtual que permita a los estudiantes la interacción con diferentes prácticas dedicadas al análisis y tratamiento de información?

1.3 SISTEMATIZACIÓN DEL PROBLEMA

1. ¿Qué conjunto de tecnologías se van a utilizar?
2. ¿Cuáles son los conjuntos de datos que se van a utilizar?
3. ¿Qué tratamiento se le dará a los conjuntos de datos?
4. ¿Qué algoritmos se van a utilizar?
5. ¿Cómo será la interacción de los estudiantes con el material?

2 OBJETIVOS

2.1 OBJETIVO GENERAL

Documentar el procedimiento de análisis y tratamiento a diferentes conjuntos de datos para las practicas y/o laboratorios de *Machine learning e Inteligencia Artificial* en los procesos de formación dirigidos por los docentes del grupo de investigación e inteligencia computacional ***LASER*** a sus estudiantes.

2.2 OBJETIVOS ESPECÍFICOS

- Documentar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos.
- Crear documentos reproducibles que muestren la ejecución de los diferentes algoritmos aplicados.
- Consolidar los procedimientos realizados en un formato amigable y de fácil acceso para los estudiantes y el publico en general.

3 MARCO DE REFERENCIA

3.1 MARCO TEÓRICO

3.1.1 Algoritmo

Es una sucesión finita de instrucciones diseñadas para un propósito específico [8], además contiene una descripción de los datos para resolver el proposito en mención [9].

3.1.2 Python

Es un lenguaje de programación de proposito general, muy popular en el área cientifica que permite incluir código escrito en otros lenguajes. Además al ser un lenguaje interpretado, requiere un interprete para ser traducido a lenguaje maquina por lo tanto puede ejecutarse en diferentes sistemas operativos [10].

3.1.3 R

Es un paquete estadístico muy popular en la comunidad científica para la manipulación y tratamiento de datos. Además cuenta con una variedad de extensiones que permiten realizar análisis especializado de información, ya sea desde un punto de vista gráfico, hasta modelos de regresión, entre otros [11].

3.1.4 Estadística Descriptiva

Es el conjunto de técnicas para la presentación y reducción de los datos, ya sea desde el punto de vista gráfico, a través de la obtencion de las variables estadísticas (moda, media, mediana, varianza, entre otros), e inclusive utilizando técnicas que estudian la dependencia entre dos o mas características (regresión y correlación) [12].

3.1.5 Módelo de Regresión lineal

Cuando se puede determinar que existe una correlación entre dos variables, es decir que una variable pueda describir el comportamiento de otra el módelo de regresión lineal consiste en una prueba de hipótesis para determinar si existe una recta cuyo coeficiente de error sea aceptable. Aunque este módelo no esta limitado a dos variables, ya que se puede extender al modelo de regresión lineal multivariado [13].

3.1.6 Aprendizaje Supervisado

Es cuando en el proceso de aprendizaje hay un maestro o supervisor, que se encarga de alimentar a por medio de datos marcados, validar las respuestas esperadas y hacer ajustes al modelo [14].

3.1.7 Aprendizaje No Supervisado

Es cuando en el proceso de aprendizaje no hay un maestro o supervisor, es decir el modelo se encarga de identificar patrones, hacer correlaciones, y agrupar características que arrojen similitudes en la variable de interés [14].

3.1.8 Redes Neuronales

Son módelos matemáticos que teorizan el comportamiento del cerebro, explorando y reproduciendo información de una forma similar al cerebro. Su uso frecuentemente es el análisis de datos, detección de patrones, entre otros [15].

3.1.9 Redes Neuronales Convolucionales

Es un caso especial de las redes neuronales, en el cual se asume que las entradas son imágenes o estan codificadas como si lo fuesen, y las neuronas están marcadas por pesos permitiendo resaltar características [16].

3.2 MARCO TEMPORAL

El proyecto será desarrollado durante 5 meses a partir de septiembre del año 2020 hasta enero del año 2021.

4 METODOLOGÍA

Fase	Método	Objetivo Específico	Actividades	Resultado
Contexto	Análisis de literatura	Detallar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Revisión de las tecnologías y lenguajes de programación más usados para el procesamiento de datos en los ultimos años	Determinar las tecnologias y lenguajes de programación que se utilizaran en el proyecto
Análisis	Análisis de literatura	Documentar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Revisión de los diferentes conjuntos de datos candidatos para el análisis	Determinar los conjuntos de datos que se utilizaran en el proyecto de pasantia
Análisis	Análisis de datos	Documentar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Realizar el análisis exploratorio de los conjuntos de datos	Determinar la correlación de variables y la selección de las mismas para su respectivo análisis y procesamiento
Implementación	Construcción y validación de modelos	Detallar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Aplicar el método de regresión lineal (si aplica)	Plantear modelos basados en la regresión lineal que describan la(s) variable(s) objetivo
Implementación	Construcción y validación de modelos	Detallar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Utilizar algoritmos de clasificación (si aplica)	Plantear modelos basados en algoritmos de clasificación que describan la(s) variable(s) objetivo
Implementación	Construcción y validación de modelos	Detallar los pasos y decisiones tomadas en el proceso efectuado a cada conjunto de datos	Realizar algún tratamiento por medio de redes neuronales al conjunto de datos	Plantear modelos basados en redes neuronales que describan la(s) variable(s) objetivo
Documentación	Recopilar resultados	Crear documentos reproducibles que muestren la ejecución de los diferentes algoritmos aplicados	Documentar los detalles del análisis realizado a cada conjunto de datos	Un documento reproducible que contenga el análisis realizado a cada conjunto de datos
Documentación	Recopilar resultados	Consolidar los procedimientos realizados en un formato amigable y de fácil acceso para los estudiantes y el publico en general	Validar y comparar los resultados de modelos generados	Habilitar un sitio web para compartir los procedimientos y resultados obtenidos del análisis de los datos

Tabla 1: Metodología de trabajo

5 IMPACTO Y RESULTADOS ESPERADOS

Dimensión	Item Colciencias	Elemento tangible
Generación de conocimiento.	Producción Bibliográfica. Articulo.	Un (1) artículo con los procedimientos realizados a los datos, resultados del análisis de los mismos y las lecciones aprendidas.
Apropiación social del conocimiento.	Circulación de conocimiento Especializado. Evento Científico.	Publicar en un sitio web administrado por el grupo de investigación ***LASER*** los procedimientos realizados a los datos, resultados del análisis de los mismos y las lecciones aprendidas.

Tabla 2: Impacto y resultados esperados de la pasantia de investigación

6 CRONOGRAMA

Las siguientes estimaciones de tiempos están sujetas a ajustes puesto que corresponden a una estimación de alto nivel por lo tanto no se contemplan los imprevistos.

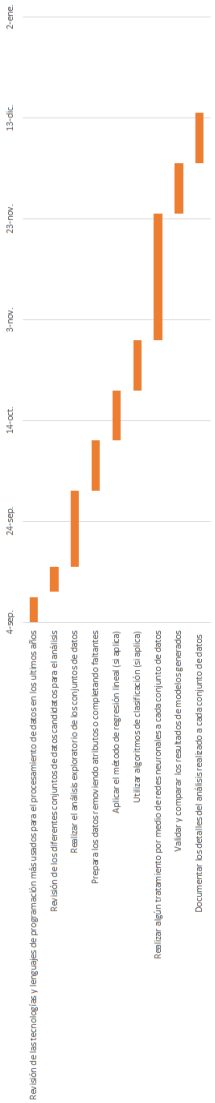

Figura 1: Cronograma del actividades de la pasantía

REFERENCIAS

[1] F. Díaz-Barriga, "César coll y carles monereo (eds). psicología de la educación virtual. aprender y enseñar con las tecnologías de la información y la comunicación. madrid: Morata." *REIRE. Revista d'Innovació i Recerca en Educació; Vol.: 2 Núm.: 3*, vol. 2, 10 2009.

[2] M. A. Unigarro Gutierrez, *Educación virtual: encuentro formativo en el ciberespacio*. UNAB, 2004. [Online]. Available: https://books.google.com.co/books?id=C03hWjUL9OAC&printsec=frontcover&dq=educacion+virtual&hl=es-419&sa=X&ved=2ahUKEwjRzsy9493rAhUpxVkKHUdXA2kQ6AEwAXoECAUQAg#v=onepage&q=educacion%20virtual&f=false

[3] G. de Investigación LASER, "Laser," *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://comunidad.udistrital.edu.co/laser/

[4] *Coursera Inc.*, "Coursera," *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://play.google.com/store/apps/details?id=org.coursera.android&hl=es_CO

[5] Udemy, "Udemy," *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://play.google.com/store/apps/details?id=com.udemy.android&hl=es_CO

[6] Jupyter, "Jupyter," *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://jupyter.org

[7] Google, "Colaboratory," *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://research.google.com/colaboratory/faq.html

[8] A. Marquina Vila, *Elogio del algoritmo: Las matemáticas del cálculo científico*. Universitat de Valencia, 2016. [Online]. Available: https://books.google.com.co/books?id=b9A6DgAAQBAJ&pg=PT9&dq=algoritmo&hl=es-419&sa=X&ved=2ahUKEwiUnvXlj97rAhXixlkKHeLTDdQQ6AEwA3oECAYQAg#v=onepage&q=algoritmo&f=false

[9] C. Bertolotti Zuñiga and J. J. Flores, *Método de las 6'D. modelamiento - algoritmo - programación. (enfoque orientado a las estructuras lógicas)*. MACRO, 2014.

[10] A. Cuevas Alvarez, *Programar con Python 3*. RA-MA Editorial, 2016.

[11] *The R Foundation*, "What is r?" *Sitio web consultado el 01 de septiembre de 2020.* [Online]. Available: https://www.r-project.org/about.html

[12] S. Fernández Fernández, J. M. Cordero Sánchez, and A. Córdoba Largo, *Estadística Descriptiva*. Escuela Superior de Gestion Comercial y Marketing. [Online]. Available: https://books.google.com.co/books?id=31d5cGxXUnEC&printsec=frontcover&dq=estadistica+descriptiva&hl=es-419&sa=X&ved=2ahUKEwjcm-Ptl97rAhXMwFkKHZ9-BVMQ6AEwAXoECAEQAg#v=onepage&q=estadistica%20descriptiva&f=false

[13] S. M. Ross, *Introducción a la estadística*. Editorial Reverte, S.A. [Online]. Available: https://books.google.com.co/books?id=pPM2TgQsx8wC&pg=PA527&dq=regresion+lineal&hl=es-419&sa=X&ved=2ahUKEwjj8qSpmt7rAhUyo1kKHYDUAggQ6AEwA3oECAUQAg#v=onepage&q=regresion%20lineal&f=false

[14] R. Flórez López and J. M. Fernández Fernández, *Las Redes Neuronales Artificiales. Fundamentos teóricos y aplicaciones prácticas*. Netbiblo. [Online]. Available: https://books.google.com.co/books?id=X0uLwi1Ap4QC&pg=PA33&dq=aprendizaje+supervisado&hl=es-419&sa=X&ved=2ahUKEwjKq4uint7rAhXLxlkKHWr3DpAQ6AEwA3oECAQQAg#v=onepage&q=aprendizaje%20supervisado&f=false

[15] M. Redondo Fonseca, *Simulación de redes neuronales como herramienta Big Data en el ámbito sanitario*. Lulu Press, Inc. [Online]. Available: https://books.google.com.co/books?id=9vSBDgAAQBAJ&pg=PA17&dq=redes+neuronales&hl=es-419&sa=X&ved=2ahUKEwihra7dod7rAhXwuFkKHa6CCVMQ6AEwCXoECAAQAg#v=onepage&q=redes%20neuronales&f=false

[16] J. Torres, *Deep Learning. Introducción práctica con Keras*. Lulu Press, Inc. [Online]. Available: https://books.google.com.co/books?id=ju1mDwAAQBAJ&pg=PA147&dq=redes+neuronales+convolucionales&hl=es-419&sa=X&ved=2ahUKEwj6gLOLpd7rAhXntlkKHR0KCbEQ6AEwBXoECAQQAg#v=onepage&q=redes%20neuronales%20convolucionales&f=falsee