UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS ECONÓMICO ADMINISTRATIVAS

MAESTRÍA EN CIENCIA DE LOS DATOS

Protocolo de Investigación

Técnicas de Machine Learning para estudios de epigenética: caso de estudio Sanvite

P R E S E N T A

Eduardo Carlos Fletes Aréchiga

Director del proyecto

Iván Alejandro Toledano Juárez

14 Febrero 2024

eduardo.fletes1892@alumnos.udg.mx

CUCEA
El mejor lugar para el talento

MCD MAESTRÍA EN CIENCIA DE LOS DATOS

Comentarios sobre trabajo o proyecto de alumno

Ficha de evaluación de la presentación en coloquios

Profesor comentarista	Dr. Sergio Rogelio Tinoco Martínez
Nombre del alumno	Eduardo Fletes
Tema de trabajo	Predicción en epigenética
Período académico	**2023B**

ASPECTOS POR CONSIDERAR

	DESCRIPCIÓN	(E)	(B)	(R)	(D)
1	Estructura de la presentación	X			
2	Apoyos visuales y gráficos		X		
3	Capacidad para transmitir las ideas		X		
4	Seguridad e interacción con el auditorio (manejo escénico)		X		
5	Dominio del tema		X		
6	Ortografía y gramática (en diapositivas o materiales de apoyo)	X			
7	Apariencia y formalidad (vestimenta adecuada, vocabulario empleado)		X		
8	Fluidez y agilidad de la presentación	X			
9	Manejo del tiempo		X		
10	Capacidad para responder las preguntas	X			

ASPECTOS A EVALUAR DESEMPEÑOS

E= Excelente, B= Bien, R=Regular, D=Deficiente

Comentarios y/o recomendaciones

Forma	Fondo

ATENTAMENTE

Zapopan, Jalisco a 09 de enero de 2024

Dr. Sergio Rogelio Tinoco Martínez

Nombre y firma del

Profesor comentarista

Comentarios sobre trabajo o proyecto de alumno
Ficha de evaluación de la presentación en coloquios

Profesor comentarista	Dr. Alex Guillén Bonilla
Nombre del alumno	Fletes Aréchiga Eduardo Carlos
Tema de trabajo	Técnicas de Machine Learning para Predicción de Patrones en Epigenética: Caso de Estudio Sanvite
Período académico	2023-B

ASPECTOS POR CONSIDERAR

	DESCRIPCIÓN	(E)	(B)	(R)	(D)
1	Estructura de la presentación		x		
2	Apoyos visuales y gráficos		x		
3	Capacidad para transmitir las ideas		x		
4	Seguridad e interacción con el auditorio (manejo escénico)		x		
5	Dominio del tema		x		
6	Ortografía y gramática (en diapositivas o materiales de apoyo)		x		
7	Apariencia y formalidad (vestimenta adecuada, vocabulario empleado)		x		
8	Fluidez y agilidad de la presentación		x		
9	Manejo del tiempo		x		
10	Capacidad para responder las preguntas		x		

ASPECTOS A EVALUAR DESEMPEÑOS
E= Excelente, B= Bien, R=Regular, D=Deficiente

Comentarios y/o recomendaciones

Forma	Fondo
Las recomendaciones y comentarios fueron hechos en su presentación.	

ATENTAMENTE
Zapopan, Jalisco a 9 de enero de 2024

Dr. Alex Guillén Bonilla
Profesor comentarista

Contenido

Título del protocolo de investigación

Técnicas de machine learning para análisis de estudios de epigenética: caso de estudio Sanvite

Planteamiento del problema

La epigenética es el estudio de los cambios heredables en la expresión de los genes que no implican alteraciones en la secuencia del ADN (NIH Biblioteca Nacional de Medicina, 2021). Estos cambios pueden estar influenciados por factores ambientales, como la dieta, el estrés o la exposición a sustancias químicas. La epigenética tiene implicaciones importantes para la salud humana, ya que puede estar relacionada con el desarrollo de enfermedades como el cáncer, la diabetes o el alzhéimer. Por ello, existen empresas que se dedican a realizar exámenes de epigenética para detectar posibles alteraciones en el perfil epigenético de las personas y ofrecer recomendaciones personalizadas para mejorar su calidad de vida.

Sin embargo, la interpretación de los resultados de los exámenes de epigenética no es una tarea sencilla, ya que implica analizar una gran cantidad de datos complejos y variables que pueden tener diferentes significados según el contexto. Además, los médicos que leen los resultados de epigenética pueden tener limitaciones de tiempo, conocimiento o experiencia para identificar todos los patrones relevantes que pueden indicar un riesgo o una oportunidad para la salud del paciente. Esto puede generar una pérdida de información valiosa que podría ayudar a prevenir o tratar enfermedades.

Ante esta situación, surge la necesidad de utilizar las herramientas y técnicas de la ciencia de los datos para apoyar la interpretación de los resultados de epigenética. La ciencia de los datos es una disciplina que combina el uso de métodos estadísticos, computacionales y matemáticos para extraer conocimiento útil a partir de grandes volúmenes de datos (IBM, s.f.). Una de sus ramas es el aprendizaje automático, que permite crear modelos que aprenden de los datos y hacen predicciones o clasificaciones sobre nuevos casos. El aprendizaje automático podría ser útil para descubrir posibles patrones en los resultados de epigenética que pueden pasar desapercibidos a los médicos, como, por ejemplo, la asociación entre ciertos marcadores epigenéticos y el desarrollo de enfermedades específicas.

El objetivo general de esta investigación es diseñar e implementar un algoritmo de aprendizaje automático que ayude a interpretar los resultados de epigenética y a descubrir posibles patrones que puedan mejorar la salud de las personas.

Objetivo de la tesis

El objetivo de esta tesis es diseñar e implementar un algoritmo de aprendizaje automático que ayude a interpretar los resultados de epigenética y a descubrir posibles patrones que puedan mejorar la salud de las personas. Para ello, se pretende:

- Revisar la literatura científica sobre la epigenética y el aprendizaje automático, así como los trabajos previos que hayan aplicado estas técnicas al análisis de datos epigenéticos.
- Seleccionar el tipo de algoritmo de aprendizaje automático más adecuado para el problema planteado, teniendo en cuenta los criterios de eficiencia, precisión y explicabilidad.
- Obtener y procesar los datos epigenéticos de una empresa que realiza exámenes de epigenética, asegurando su calidad, integridad y confidencialidad.
- Entrenar y validar el algoritmo de aprendizaje automático con los datos epigenéticos, utilizando diferentes medidas de evaluación y comparando sus resultados con los obtenidos por los médicos que interpretan los exámenes.
- Analizar e interpretar los patrones descubiertos por el algoritmo de aprendizaje automático, identificando su significado biológico y clínico, así como sus implicaciones para la prevención o el tratamiento de enfermedades.
- Demostrar que el algoritmo de aprendizaje automático propuesto es capaz de apoyar la interpretación de los resultados de epigenética y de descubrir posibles patrones que puedan mejorar la salud de las personas.

Justificación

La epigenética es una rama de la biología que estudia los cambios heredables en la expresión génica que no implican alteraciones en la secuencia del ADN. Estos cambios pueden estar influenciados por factores ambientales, nutricionales, psicológicos y sociales, y pueden tener efectos en la salud y el desarrollo de los individuos. Sin embargo, la interpretación de los resultados de los exámenes de epigenética que se realizan a los pacientes depende en gran medida de la experiencia y los sesgos de los médicos, lo que puede limitar la precisión y la eficacia del diagnóstico y el tratamiento.

El propósito de esta investigación es proporcionar una herramienta a los médicos que reciben los resultados de exámenes de epigenética y que les ayude a dar un mejor diagnóstico y tratamiento a los pacientes. Para ello, se pretende diseñar modelos predictivos (AWS, s.f.) basados en algoritmos de aprendizaje automático y redes neuronales que puedan detectar patrones patológicos o relaciones entre distintos sistemas del cuerpo humano que tengan relación directa con la patología del paciente (IBM, s.f.). Se espera que esta investigación contribuya al avance del conocimiento científico en el campo de la epigenética y la ciencia de datos, así como a la mejora de la calidad de vida de los pacientes que sufren enfermedades relacionadas con la expresión génica. Los resultados se difundirán mediante publicaciones académicas y conferencias, y se aplicarán mediante el desarrollo de una interfaz gráfica y una guía de uso para los modelos predictivos

Marco teórico

El término Epigenética fue acuñado en la década del cincuenta para describir el mecanismo por el cual los organismos multicelulares desarrollan múltiples tejidos diferentes a partir de un único genoma. En la actualidad reconocemos que este proceso se logra mediante marcas moleculares detectables; dichas marcas generan modificaciones que afectan la actividad transcripcional de los genes y una vez establecidas son relativamente estables en las siguientes generaciones. El uso actual del término consiste en indicar cambios heredables en la estructura y organización del ADN que no involucran cambios en la secuencia y que modulan la expresión génica. Estos cambios en la expresión génica implican, entonces, cambios heredables en el fenotipo. Los mecanismos tradicionales de regulación epigenética incluyen metilación del ADN y modificaciones de histonas (Rodriguez Dorantes, Téllez Ascencio, Cerbón, Lopez, & Cervantes, 2004), entendiendo a estas

proteínas como las encargadas de empaquetar el ADN y considerando que los dos tipos de mecanismos participan en la modulación de los complejos remodeladores de la cromatina (García Robles, Ayala Ramírez, & Perdomo Vazquez, 2012).

Según el National Human Genome Research Institute (NIH National Human Genome Research Institute, 2023), el término epi significa por encima. Es un prefijo griego. También se define como por encima de la secuencia base de ADN. En términos generales se puede comparar con los acentos de las palabras donde el ADN es el lenguaje y las modificaciones son los acentos. Las marcas epigenéticas, cambian la forma como se expresan los genes. La promesa de la epigenética es que nos cuenta acerca de la célula, es una manera de definir la célula que es diferente que si simplemente miramos los niveles de expresión génica. Cualquier tipo de célula que miremos tiene patrones epigenéticos especializados. Hay dos tipos de modificaciones: la metilación del ADN y la modificación de las histonas. La metilación del ADN se ve alterado en el cáncer por lo que si sabemos cuál es el patrón normal de metilación y luego observamos el patrón de metilación en un tumor podríamos ver los cambios que estaban teniendo lugar y cuáles son los genes afectados.

Un estudio de epigenética puede arrojar resultados sobre cómo los cambios epigenéticos pueden ayudar a determinar si los genes están activados o desactivados y pueden influir en la producción de proteínas de ciertas células, asegurando que solo se produzcan las proteínas necesarias. Por ejemplo, las proteínas que promueven el crecimiento de los huesos no se producen en las células musculares. También puede arrojar resultados sobre cómo los factores ambientales, como la dieta, el estrés, el ejercicio o las infecciones, pueden generar un impacto en la célula y modificar la expresión de algunos genes. (NIH Biblioteca Nacional de Medicina, 2021)

Hipótesis

El algoritmo de aprendizaje automático propuesto mejora la interpretación de los resultados de epigenética y descubre posibles patrones que pueden mejorar la salud de las personas, dependiendo del tipo de datos epigenéticos y del tipo de enfermedad asociada.

Donde para ello se tienen contempladas las siguientes variables:

- Variable independiente: es la variable que se manipula o controla en la investigación, y que se supone que tiene un efecto sobre la variable dependiente. En este caso, el algoritmo de aprendizaje automático propuesto.
- Variable dependiente: es la variable que se mide o evalúa en la investigación, y que se supone que es afectada por la variable independiente.

En el presente documento es la interpretación de los resultados de epigenética y el descubrimiento de posibles patrones.

- Variable interviniente: es la variable que modera o media la relación entre la variable independiente y la variable dependiente, y que puede influir en el resultado de la investigación. Aquí analizo el tipo de datos epigenéticos o el tipo de enfermedad asociada.

Metodología de investigación

La metodología de investigación que se utilizará para llevar a cabo la tesis es la siguiente:

- Tipo de investigación: se trata de una investigación aplicada, ya que busca resolver un problema práctico mediante el uso de la ciencia de los datos y el aprendizaje automático. También se trata de una investigación cuantitativa, ya que se basa en el análisis de datos numéricos y estadísticos.

- Diseño de investigación: se trata de un diseño experimental, ya que se manipula la variable independiente (el algoritmo de aprendizaje automático) y se mide su efecto sobre la variable dependiente (la interpretación de los resultados de epigenética y el descubrimiento de posibles patrones). Se utilizará un diseño pretest-postest con grupo control, es decir, se comparará el rendimiento del algoritmo propuesto con el de un algoritmo existente o con el de los médicos que interpretan los exámenes, antes y después de aplicar el algoritmo a los datos epigenéticos.

- Población y muestra: la población de estudio está conformada por los resultados de epigenética que la empresa Sanvite recopila y almacena. La muestra será una selección aleatoria y representativa de dichos resultados, teniendo en cuenta el tipo de datos epigenéticos y el tipo de enfermedad asociada. Se calculará el tamaño muestral adecuado para garantizar la validez y la confiabilidad de los resultados.

- Instrumentos y técnicas de recolección de datos: los instrumentos que se utilizarán para recolectar los datos son los exámenes de epigenética realizados por la empresa Sanvite, que contienen información sobre los marcadores epigenéticos y las características clínicas de los pacientes. Las técnicas que se utilizarán para recolectar los datos son la solicitud y el acceso

a la base de datos de la empresa, previa autorización y consentimiento informado, y la extracción y el almacenamiento de los datos en un formato adecuado para su posterior análisis.

- Procesamiento y análisis de datos: los datos se procesarán mediante técnicas de limpieza, transformación, normalización y reducción de dimensionalidad, para asegurar su calidad e integridad. El análisis de datos se realizará mediante técnicas de aprendizaje automático, utilizando diferentes algoritmos y parámetros para entrenar y validar el modelo propuesto. Se utilizarán diferentes medidas de evaluación para comparar el rendimiento del modelo propuesto con el del modelo existente o con el de los médicos, tales como la precisión, la sensibilidad, la especificidad, el valor predictivo positivo, el valor predictivo negativo y la curva ROC. También se realizará un análisis e interpretación de los patrones descubiertos por el modelo propuesto, identificando su significado biológico y clínico, así como sus implicaciones para la prevención o el tratamiento de enfermedades.

Recursos

Los recursos que se utilizarán para la elaboración del proyecto de investigación son los siguientes:

Recursos humanos: el proyecto contará con la participación de un investigador principal, que será el autor de la tesis, y de un asesor académico, que será el director de la tesis. El investigador principal se encargará de diseñar e implementar el algoritmo de aprendizaje automático, de obtener y procesar los datos epigenéticos, de entrenar y validar el modelo propuesto, de analizar e interpretar los resultados y de redactar el informe final. El asesor académico se encargará de orientar y supervisar el desarrollo del proyecto, de revisar y retroalimentar el informe final y de avalar la calidad y la originalidad del trabajo. Además, se contará con el apoyo de la empresa que realiza los exámenes de epigenética, que facilitará el acceso a la base de datos y a los médicos que interpretan los exámenes.

Recursos materiales: el proyecto requerirá de las siguientes instalaciones, equipos y componentes:

- Una computadora personal con conexión a internet, que servirá para realizar la revisión bibliográfica, el diseño e implementación del algoritmo, el procesamiento y análisis de los datos y la redacción del informe final.
- Un software especializado para el desarrollo y ejecución del algoritmo de aprendizaje automático, que puede ser Python, R, MATLAB o cualquier otro lenguaje o plataforma adecuada para este tipo de aplicaciones.
- Una base de datos con los resultados de epigenética de la empresa, que contendrá información sobre los marcadores epigenéticos y las características clínicas de los pacientes. La base de datos se accederá mediante una solicitud y una autorización previa, y se extraerá y almacenará en un formato adecuado para su posterior análisis.
- Un dispositivo de almacenamiento externo, como una memoria USB o un disco duro portátil, que servirá para respaldar la información generada durante el proyecto y evitar posibles pérdidas o daños.

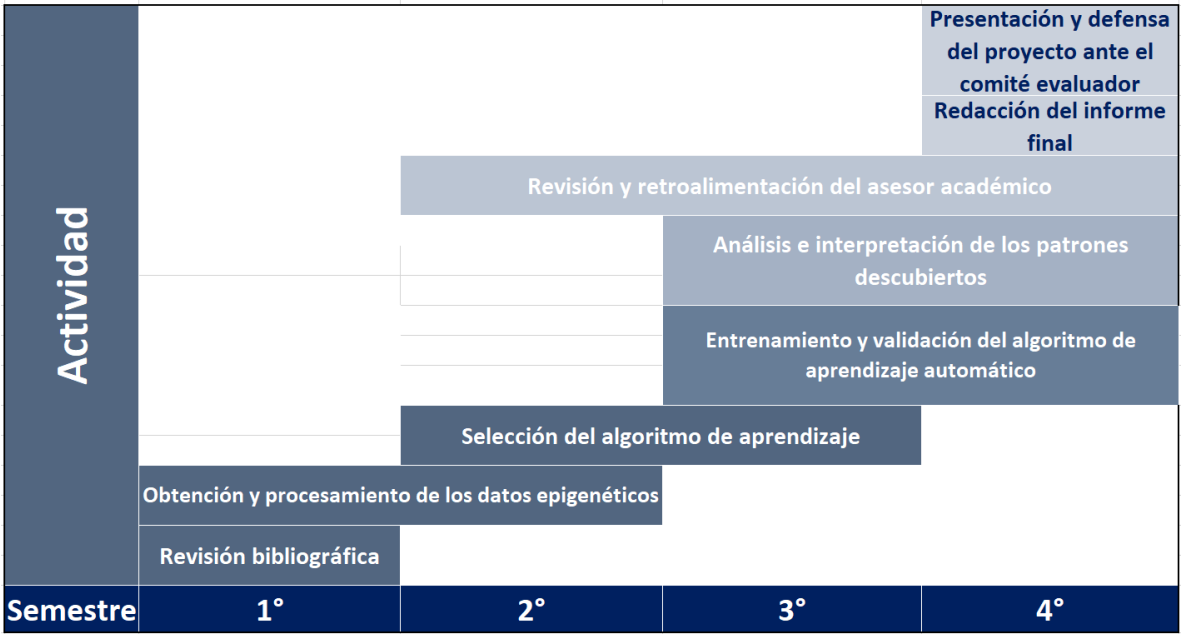
Recursos económicos: el proyecto tendrá un costo estimado de acuerdo con los siguientes conceptos:

- Honorarios del investigador principal: por ser parte del proyecto de titulación de la maestría no se consideran honorarios.
- Honorarios del asesor académico: se considera una remuneración por hora equivalente al promedio del mercado académico para un profesor o investigador con nivel doctoral en ciencia de los datos. Se estima que el asesor dedicará 4 horas semanales al proyecto durante 12 meses, por lo que se multiplica la remuneración por hora por 4 por 52.
- Costo del software: se considera el precio del software especializado para el desarrollo y ejecución del algoritmo de aprendizaje automático. Se asume que se utilizará una licencia gratuita o académica, por lo que no se genera ningún costo adicional.
- Costo de la base de datos: se considera el costo por acceder a la base de datos con los resultados de epigenética de la empresa. Se asume que la empresa colaborará con el proyecto sin cobrar ningún costo adicional.
- Costo del dispositivo de almacenamiento externo: se considera el precio del dispositivo de almacenamiento externo que se utilizará para respaldar la información generada durante el proyecto. Se asume que se utilizará una memoria USB o un disco duro portátil con una capacidad suficiente para almacenar los datos y los archivos del proyecto.
- Costo total: se suma el costo de cada uno de los conceptos anteriores para obtener el costo total estimado del proyecto.

Cronograma

El cronograma del proyecto de investigación se establece de acuerdo con la duración y la organización del plan de estudios de la maestría en ciencia de los datos, que tiene una duración de dos años y se divide en cuatro semestres. Cada semestre tiene una duración de 16 semanas, con un periodo vacacional entre cada semestre. El proyecto se desarrollará paralelamente a las asignaturas del plan de estudios, dedicando un tiempo estimado de 10 horas semanales al proyecto.

El cronograma se presenta en la siguiente tabla



Referencias

AWS. (s.f.). *AWS Amazon*. Recuperado el 23 de Junio de 2023, de https://aws.amazon.com/es/what-is/predictive-analytics/#:~:text=Los%20modelos%20de%20an%C3%A1lisis%20predictivo%20entrenados%20pueden%20ingerir,en%20datos%20de%20forma%20m%C3%A1s%20r%C3%A1pida%20y%20precisa.

García Robles, R., Ayala Ramírez, P. A., & Perdomo Vazquez, S. P. (2012). Epigenética: definición, bases moleculares e implicaciones en la salud y evolución humana. *Revista Ciencias de la salud*, 10 (1):59-71. Recuperado el 23 de Junio de 2023, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-72732012000100006%20Con%20acceso%2022/6/2023

IBM. (s.f.). *IBM.COM*. Recuperado el 23 de Junio de 2023, de ¿Qué son las redes neuronales?: https://www.ibm.com/es-es/topics/neural-networks

IBM. (s.f.). *IBM.COM*. Recuperado el 27 de Junio de 2023, de IBM Topics Data Science: https://www.ibm.com/es-es/topics/data-science

NIH Biblioteca Nacional de Medicina. (11 de Agosto de 2021). *Medline Plus*. Obtenido de https://medlineplus.gov/spanish/genetica/entender/comofuncionangenes/epigenetica/

NIH National Human Genome Research Institute. (23 de 06 de 2023). *genome.gob*. Recuperado el 23 de 02 de 2023, de https://www.genome.gov/es/genetics-glossary/Epigenetica

Rodriguez Dorantes, M., Téllez Ascencio, N., Cerbón, M. A., Lopez, M., & Cervantes, A. (2004). Metilación del ADN: un fenómeno epigenético de importancia médica. *Revista de investigación clínica, 56*(1), 56-71. Recuperado el 23 de Junio de 2023, de https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0034-83762004000100010