**Universidad Peruana de Ciencias Aplicadas**

**UPC**

“Año de la Universalización de la Salud”



PROYECTO

Predicción de diagnóstico de diabetes en mujeres de pueblos aborígenes

CURSO

Programación concurrente y distribuida

CARRERA

Ciencias de la Computación

SECCIÓN

CC65

ALUMNO

Jose Ysique

Braulio Baldeón

Monterrico, noviembre 2020

Contenido

[Introducción](#_heading=h.gjdgxs) 3

[Objetivos](#_heading=h.30j0zll) 3

[Descripción del dataset](#_heading=h.c9d8jxeacnbk) 3

[Diseño](#_heading=h.hshe0994n05q) 4

[Desarrollo](#_heading=h.3znysh7) 4

[Conclusiones](#_heading=h.2et92p0) 5

[Referencias](#_heading=h.tyjcwt) **5**

# Introducción

La diabetes es un problema de salud pública muy grave en nuestro país que afecta a los costos directos (atención de la familia y medicamentos) e indirectos (días no laborables) de la población. Según los datos de vigilancia epidemiológica (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades, 2018), en el año 2018 se registraron más de 15 000 casos de diabetes de todo tipo. De los cuales, el mayor porcentaje corresponde a personas entre los 50 y 70 años de edad; mientras que, el porcentaje más bajo, a personas entre 20 y 30 años. En el cuadro presentado en (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades, 2018 pág. 8) se puede observar que el número de pacientes femeninos es el doble del sexo opuesto. Por otro lado, según Nicolaisen (2006) se ha vivido un sesgo profundo de atención hacia las poblaciones aborígenes que provienen de la discriminación y de la evaluación del status económico. Debido a esto “la sanidad de los pueblos indígenas es insuficiente” (Nicolaisen, 2006 pág. 2), lo que en general ocasiona que enfermedades como la diabetes no sean detectadas ni tratadas en una etapa temprana.

Según el informe del Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (2018) las complicaciones son más frecuentes de acuerdo a la duración de la diabetes. Debido a esto, se propone el presente proyecto, cuyo principal objetivo es predecir si un paciente tiene diabetes o no, basándose en determinadas medidas de diagnóstico incluidas el conjunto de datos *Pima Indians Diabetes Database[[1]](#footnote-0)*. El cual consta de 768 instancias y 9 columnas; donde las 8 primeras columnas son las medidas de diagnóstico y la última es el diagnóstico en sí. El presente documento está estructurado de la siguiente forma, en primer lugar, mencionaremos los objetivos del proyecto. Luego presentaremos brevemente la descripción de la dataset. En la sección siguiente, explicaremos el diseño de la arquitectura empleada. En cuarto lugar, detallamos el desarrollo del proyecto. Por último, se mencionan las conclusiones del proyecto.

# Objetivos

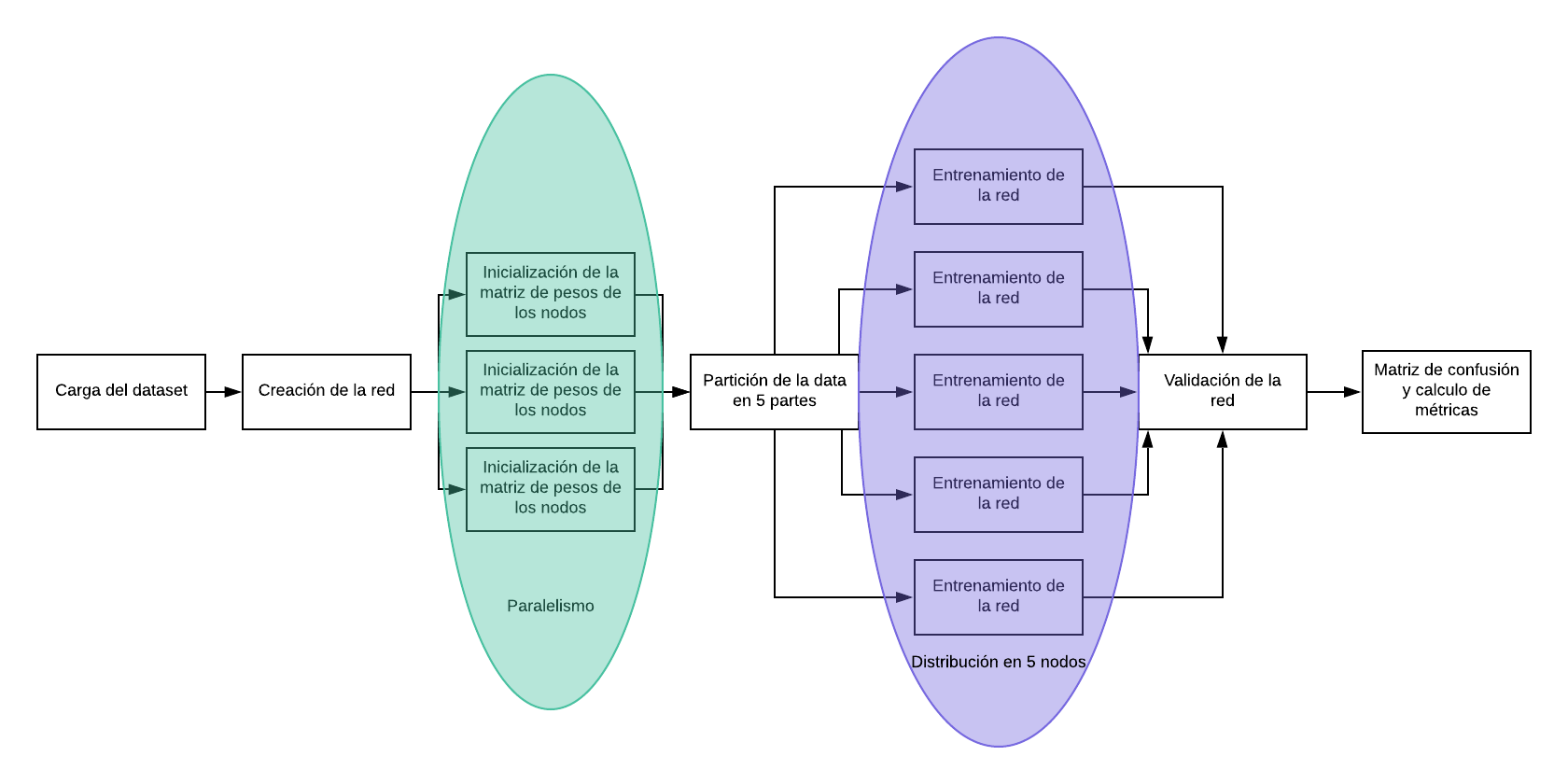
* Implementar un algoritmo de Machine Learning supervisado que aplique paralelismo.
* Desarrollar un API distribuido que pueda acceder al modelo de predicción.
* Implementar un backend para el API que funcione en al menos 5 nodos.
* Implementar un módulo de consultar JSON para el API.

# Descripción del dataset

El dataset utilizada en este trabajo consta de las siguientes categorías:

* Número de embarazos
* Concentración de glucosa (a través de un test de glucosa)
* Presión en la sangre
* Espesor del pliegue cutáneo del tríceps
* Nivel de insulina
* Número de BMI
* Función del pedigrí de la diabetes
* Edad.
* Clase (diabetes o no diabetes)

# Diseño



Librerías utilizadas para la implementación de la red:

* “math”, “math/rand”, “log”, “time”

Librerías utilizadas para las consultas JSON:

* “net/http”, “strconv” y “github.com/labstack/echo”.

# Desarrollo

El proyecto se dividió en dos hitos o entregables:

* Hito 1, un algoritmo de Machine Learning que emplee paralelismo. Para este entregable se implementó una red neuronal en el lenguaje de programación Go que inicializa la matriz de pesos de las neuronas de forma paralela.
* Hito 2, un API distribuido para acceder al modelo del hito anterior. Para este entregable se decidió realizar el entrenamiento de la red de manera distribuida en 5 nodos distintos. También, se desarrolló un módulo que maneje la predicción de nuevas instancias mediante archivos JSON.

El proceso el entrenamiento y validación de la red neuronal consiste en los siguiente pasos:

1. Se carga el dataset desde una url pública.
2. Se crea la red neuronal.
3. Se inicializa la matriz de pesos de las neuronas de forma paralela.
4. Se divide en dataset en 5 partes.
5. Se envía una partición del dataset y la red existente a cada uno de los 5 nodos.
6. En cada nodo se realiza el entrenamiento de la red.
7. En un solo nodo se realiza la validación de la red.
8. Finalmente se imprime la matriz de confusión y las métricas de validación.

Al finalizar este proceso, es posible realizar la predicción de nuevas instancias mediante una consultas usando un archivo JSON.

# Conclusiones

* Aplicar estrategias de paralelización y distribución al momento de implementar el modelo de predicción, redujo sustancialmente el tiempo requerido para el entrenamiento de la red neuronal.
* Si bien la programación concurrente es muy útil, emplearla ocasiona un mayor consumo de recursos y aumenta la dificultad de escribir el programa debido a que existe una gran dificultad para lograr una buena sincronización y comunicación entre las tareas.
* Utilizar la implementación realizada como API permite escalar ampliamente la propuesta ya que se puede enviar a un servidor externo y utilizar en clouding.

# Referencias

* Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (2018). *Boletín Epidemiológico del Perú*. Ministerio de Salud (MINSA). Volumen 27 - se 36.
* Nicolaisen, I. (2006). *Ignorados y en peligro: pueblos indígenas*. Diabetes y sociedad. Volumen 51 Número 2.

1. Kaggle. Pima Indians Diabetes Database. (recuperado de https://www.kaggle.com/uciml/pima-indians-diabetes-database) [↑](#footnote-ref-0)