Sistema completo de reconocimiento de patrones para un conjunto de datos referente al diagnóstico de COVID-19.

Luis Eduardo De Lira Hernández
Martín Montes Rivera
Alberto Ochoa Zezzatti

[1] Universidad Politécnica de Aguascalientes, Dirección de Posgrado e Investigación, Calle Paseo San Gerardo No. 207, Fracc. San Gerardo C.P.20342 Aguascalientes, Ags., México, mc200017@alumnos.upa.edu.mx

Resumen

La pandemia causada por el reciente hallazgo del coronavirus COVID-19 ha traído consigo un sinfín de inconvenientes para la humanidad. La interacción entre la tecnología y la pandemia ha arrojado un número muy alto de posibilidades para ayudar al sector a saludo a poder dar veredictos acerca de si una persona está infectada o no de COVID-19. Para poder alcanzar estos fines, es necesario el uso de Inteligencias Artificiales que ayuden a la humanidad. En el trabajo que nos compete en esta ocasión, se utilizará un conjunto de datos de pacientes que fueron diagnosticados, o no, de COVID-19, teniendo como características el sexo, la edad y la fecha. El uso de reconocimiento de patrones ayuda a que los alcances de las tecnologías sean mejores y en mayor proporción. Existen algunos métodos en el reconocimiento de patrones que, al trabajar en conjunto, se logran excelentes resultados. Estos métodos mencionados van desde el preparar el conjunto de datos para su mejor entendimiento y mejor uso en el sistema, hasta encontrar la característica más importante y por último, poder clasificarla.

***Palabras clave*— COVID-19, Reconocimiento de Patrones, Conjunto de Datos, Clasificador.**

I. INTRODUCCIÓN

El mayor problema de la pandemia de covid-19 que ha afectado a la humanidad es la falta de recursos adecuados para determinar si una persona está enferma o no. el uso de funciones de reconocimiento de patrones tiene una influencia inmediata en el procesamiento del conjunto de datos y sus valores, por lo que el uso de clasificadores lineales y no lineales como MSE Linear y MSE Polynomial nos proporcionaría un rendimiento útil.

El objetivo clave de este documento es determinar cuál de las características es la mejor para tratar. el conjunto de datos incluye valores de sexo, fecha, identificación del paciente y si el paciente está infectado o no. en mi opinión, el rasgo de infección covid-19 es el factor más importante para lograr mejores resultados. Desarrollar un sistema de reconocimiento de patrones completo no es un trabajo fácil, se necesita mucho conocimiento sobre diferentes métodos para seleccionar y clasificar características de un conjunto de datos de confianza.

II. MARCO TEÓRICO.

El reconocimiento de patrones ocurre en dos etapas.

1. Primero va la parte exploratoria. El algoritmo busca patrones en general.
2. A continuación, está la parte descriptiva, donde el algoritmo comienza a categorizar los patrones encontrados.

La combinación de los dos se utiliza para extraer conocimientos. El proceso en sí tiene este aspecto:

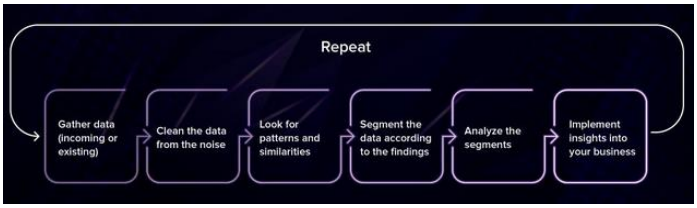

Fig. 1. Proceso de Reconocimiento de Patrones.

- Primero, necesita recopilar datos.
- Luego, lo procesa previamente y lo limpia del ruido.
- El algoritmo examina los datos y busca características relevantes o elementos comunes.
- Luego, estos elementos se clasifican o agrupan;
- Cada segmento se analiza para obtener información;
- Finalmente, los conocimientos extraídos se implementan en la práctica.

A. Etapa de selección de características.

La selección de características para la categorización de texto es un problema bien estudiado y su objetivo es mejorar la efectividad de la categorización, o la eficiencia del cálculo, o ambas.

El sistema de categorización de texto basado en la coincidencia de términos tradicional se utiliza para representar el modelo de espacio vectorial como un documento; sin embargo, necesita un gran espacio dimensional para representar el documento y no tiene en cuenta la relación semántica entre términos, lo que conduce a una pobre precisión de categorización. [1]

Las técnicas utilizadas para este trabajo son las siguientes:

- Ganancia de información. - Dado un modelo paramétrico de dependencia entre dos cantidades aleatorias, X e Y, la noción de ganancia de

información se puede utilizar para definir una medida de correlación. [2]

- Chi al cuadrado. - El estadístico Chi-cuadrado es una herramienta no paramétrica (sin distribución) diseñada para analizar diferencias de grupo cuando la variable dependiente se mide a un nivel nominal. [3]
- Puntaje de Fisher. - Se da un análisis de las propiedades computacionales del método de puntuación de Fisher para maximizar las probabilidades y resolver ecuaciones de estimación basadas en cuasi-verosimilitudes. Se ha demostrado que la estimación coherente del vector de parámetros verdadero es importante si se quiere lograr una tasa rápida de convergencia, pero si se cumple esta condición, el algoritmo es muy atractivo. [4]
- Coeficiente de correlación de Pearson. - El coeficiente de correlación de Pearson (PCC) y el coeficiente de superposición de Mander (MOC) se utilizan para cuantificar el grado de colocación entre fluoró foros. El MOC se introdujo para superar los problemas percibidos con el PCC. Los dos coeficientes son matemáticamente similares y difieren en el uso de las intensidades absolutas (MOC) o de la desviación de la media (PCC). [5]
- Umbral de varianza. - El método propuesto se compara con dos métodos basados en SVM, obteniendo un conjunto más pequeño de características con una precisión similar. [6]
- Métodos de envoltura. - Se presentan métodos de búsqueda secuencial caracterizados por un número cambiante dinámicamente de características incluidas o eliminadas en cada paso, en adelante métodos “flotantes”. Se ha demostrado que dan muy buenos resultados y son computacionalmente más efectivos que el método de bifurcación y enlace. [7]

B. Etapa de preprocesamiento

Como sabemos, la normalización es una etapa de preprocesamiento de cualquier tipo de problema de enunciado. Especialmente la normalización tiene un papel importante en el campo de la computación en software, la computación en la nube, etc. para la manipulación de datos como reducir o escalar el rango de datos antes de que se utilicen para etapas posteriores.

- Normalización de datos. - es el proceso de estructurar una base de datos relacional de acuerdo con una serie de los llamados formularios normales con el fin de reducir la redundancia de datos y mejorar la integridad de los datos.
- Eliminación de valores atípicos. - Un valor atípico es una observación que es diferente a las otras observaciones. Entonces, por esta razón, podríamos eliminarlo y, mediante esta acción, reducir el conjunto de datos y mejorar la integridad de los datos.

C. Etapa de clasificación.

Clasificadores Bayesianos.

Clasificador bayesiano ingenuo. - Se utiliza un algoritmo probabilístico de aprendizaje automático llamado clasificador Naive Bayes para ejecutar tareas de clasificación. El teorema de Bayes está en el corazón del clasificador. Podemos estimar la probabilidad de que ocurra A si B efectivamente ocurrió usando el teorema de Bayes. La prueba es B y la teoría es A. Se considera que los predictores / características están distribuidos en esta situación. Es decir, la presencia de una característica no afecta a la otra. Como resultado, se lo conoce como ingenuo.

Modelo de Markov oculto. - Un modelo estadístico de Markov en el que se supone que el sistema que se está modelando es un proceso de Markov con estados no observados (es decir, ocultos) se clasifica como un modelo de Markov oculto (HMM). Los modelos ocultos de Markov son bien conocidos por su uso en el aprendizaje por refuerzo y en implementaciones de reconocimiento de patrones temporales como el habla, la escritura a mano, el reconocimiento de gestos, el etiquetado de parte del discurso, el seguimiento de partituras musicales y la descarga parcial.

Clasificadores Lineales y No Lineales.

Descenso de gradiente. - Para encontrar un mínimo local de una función diferenciable, el descenso de gradiente es un algoritmo de optimización iterativo de primer orden. Como este es el descenso más empinado, la idea es realizar pasos repetidos en sentido inverso al gradiente de la función (o gradiente aproximado) en el punto actual.

Perceptrón. - El algoritmo Perceptron es un algoritmo de aprendizaje automático de 2 clasificaciones (clasificación binaria). Es un tipo de modelo de red neuronal, pero probablemente sea el tipo más simple de modelo de red neuronal. Está formado por un solo nodo o neurona que acepta una fila de datos y predice una etiqueta de clase.

MSE Lineal. - El error cuadrático medio de una línea de regresión significa qué tan cerca están de un conjunto de puntos. Lo consigue elevando al cuadrado las distancias entre los puntos y la línea de regresión (estas distancias son los "errores"). Debido a que está midiendo el promedio de un conjunto de errores, se llama error cuadrático medio.

Máquinas de vectores soporte. - Las máquinas de vectores de soporte, también conocidas como SVM, son un conjunto de algoritmos de aprendizaje supervisados desarrollados por Vladimir Vapnik de AT&T Labs y su equipo. Estos métodos son muy buenos para problemas de clasificación y regresión.

Vecino más cercano. - El problema de optimización de encontrar el punto en un conjunto dado que está más cerca (o más similar) a un área dada se llama búsqueda de vecino más cercano (NNS). Las métricas de disimilitud también se utilizan para expresar cercanía: cuanto mayores son los valores de la función, menos similares se vuelven los objetos.

En la clasificación, el objetivo es encontrar un mapeo de entradas a salidas dado un conjunto etiquetado de pares de entrada-salida (conjunto de entrenamiento).

III. MÉTODO

El uso de todas esas técnicas / procesos para obtener un conjunto de datos ideal fue una de las mejores decisiones que un analista pudo tomar debido a todos los beneficios que podría aportar a la gestión de un conjunto de datos para resolver un problema.

En particular, el conjunto de datos utilizado contenía alguna información que por sí sola y entre otras características no nos ayuda, por ejemplo, la edad, es una característica inútil.
Hablando de los valores atípicos, en particular para la función Edad, contiene información sobre las edades de los pacientes que no es útil, por ejemplo, el código devolvió que la edad media en la que se infectaron más pacientes era de 65 años, por lo que, es mejor eliminar todos los valores antes de esa edad media.

Además, el uso de un conjunto de datos normalizado trajo más datos integrados para administrar.

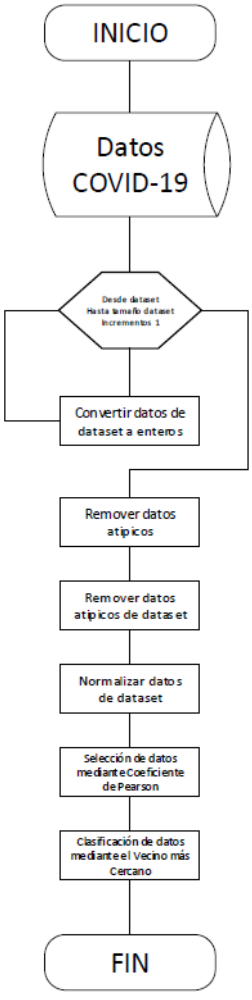

Fig. 2. Dataset primitivo.

Se utilizó el lenguaje de programación Python para realizar todos los métodos antes mencionados. Para este trabajo, los métodos que se usaron son:

- Convertir el conjunto de datos.
- Remover valores atípicos.
- Normalizar el conjunto de datos.
- Seleccionar la mejor característica.
- Clasificar la mejor característica.

IV. RESULTADOS Y DISCUSIONES.

Los siguientes resultados son de los procesos mencionados que se utilizaron para el desarrollo del sistema.

A. *Convertir valores de dataset a tipo float.*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 377 entries, 0 to 376
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Patient_ID         377 non-null    object
 1   SEX                377 non-null    object
 2   AGE                377 non-null    object
 3   Date               377 non-null    int64
 4   COVID19_Infection  377 non-null    object
dtypes: int64(1), object(4)
memory usage: 14.9+ KB
```

Fig. 3. Dataset primitivo.

Se convierten los valores tipo object a tipo float (Ver Fig.2) debido a que se necesitan para los futuros cálculos y procesos (Ver Fig. 3)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 377 entries, 0 to 376
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Patient_ID         377 non-null    float64
 1   SEX                377 non-null    float64
 2   AGE                377 non-null    float64
 3   Date               377 non-null    float64
 4   COVID19_Infection  377 non-null    float64
dtypes: float64(5)
memory usage: 17.7 KB
```

Fig. 4. Dataset convertido.

B. *Remover datos atípicos.*

Es recomendable remover los datos que se encuentra muy por encima o muy por debajo de la media para así no obtener contratiempos durante el entrenamiento de nuestro sistema.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 377 entries, 0 to 376
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Patient_ID        377 non-null    float64
 1   SEX               377 non-null    float64
 2   AGE               377 non-null    float64
 3   Date              377 non-null    float64
 4   COVID19_Infection 377 non-null    float64
dtypes: float64(5)
memory usage: 17.7 KB
```

Fig. 5. Dataset con todos los datos.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 70 entries, 0 to 372
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Patient_ID        70 non-null     float64
 1   SEX               70 non-null     float64
 2   AGE               70 non-null     float64
 3   Date              70 non-null     float64
 4   COVID19_Infection 70 non-null     float64
dtypes: float64(5)
memory usage: 3.3 KB
```

Fig. 6. Dataset con todos los datos.

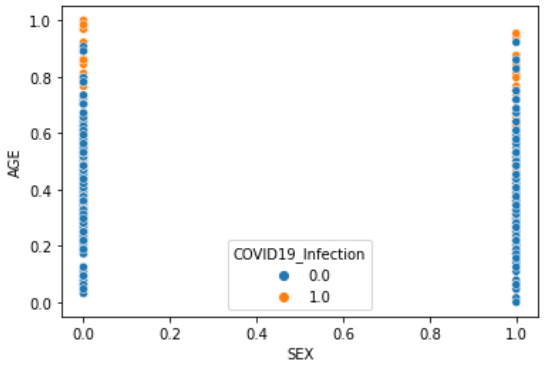

Fig. 7. Datos atípicos sin remover.

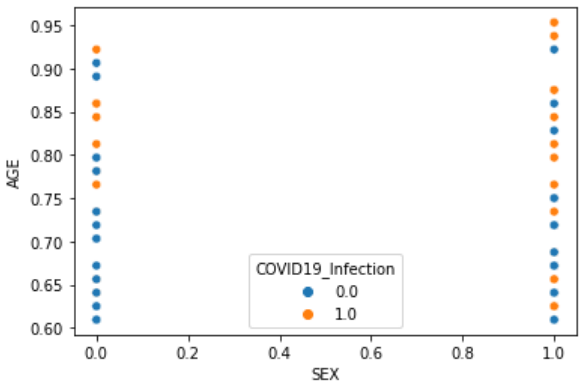

Fig. 8. Datos atípicos removidos.

C. *Normalizado de dataset.*

La normalización de un dataset es el método para asemejar los valores de los datos obtenidos en orden para evitar problemas.

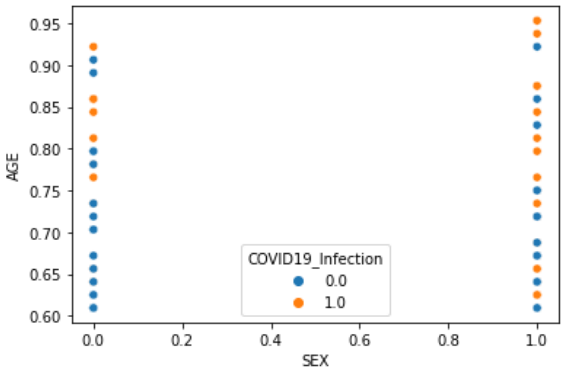

Fig. 8. Dataset sin normalizar.

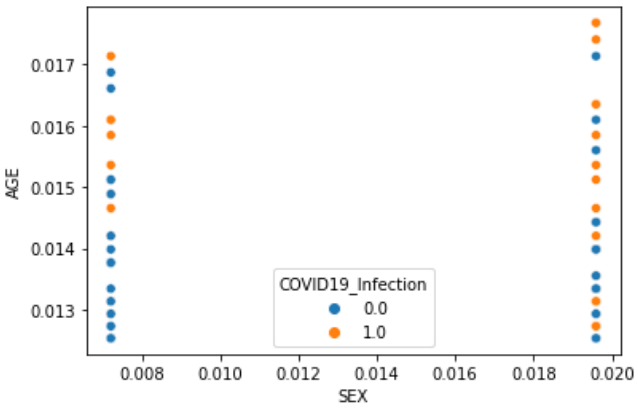

Fig. 9. Dataset normalizada.

D. *Selección de características.*

Se selecciona la característica más acorde con el fin del estudio. En este caso se utilizó la Correlación de Pearson. (Ver Fig. 9)

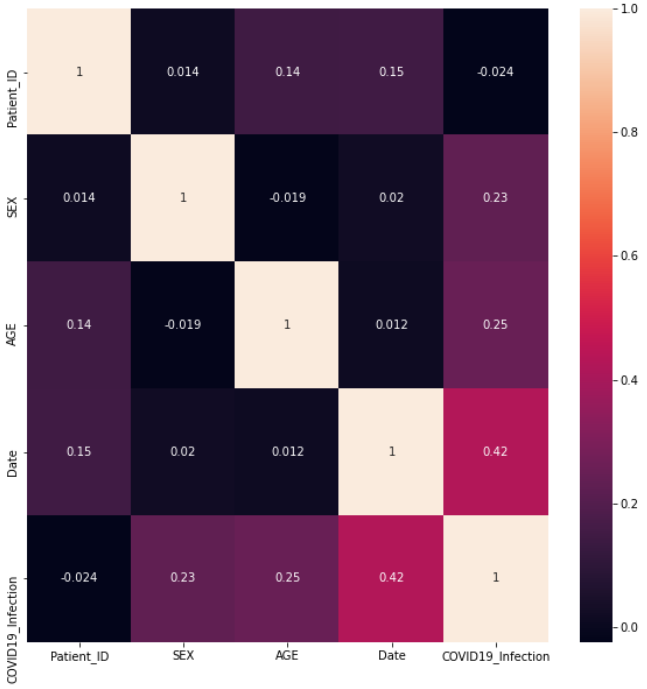

Fig. 10. Mapa de calor de mejor característica.

E. Clasificación de características.

Se utiliza el método del vecino más cercano, ya que la forma en la que este clasifica arroja resultados esperados. (Ver Fig. 10)

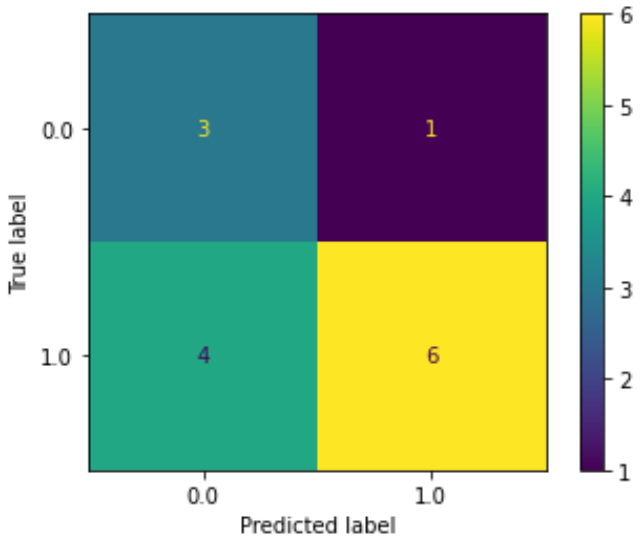

Fig. 11. Matriz de confusión.

```
COVID19_Infection    1.000000
Date                 0.424499
AGE                  0.247759
SEX                  0.230940
Patient_ID           0.024346
Name: COVID19_Infection, dtype: float64
['Date']
```

Fig. 12. La fecha, es la mejor característica seleccionada y que tiene una mejor relación con la salida.

El escenario ideal sería que el sistema arrojará 0 falsos positivos y 0 falsos negativos en la matriz de confusión.

V. CONCLUSIONES.

El objetivo inicial del trabajo fue trabajar con los distintos métodos de reconocimiento de patrones, se trabajó con ellos, se encontraron resultado hasta cierto punto razonables y correctos, pero, a decir verdad, se necesita un trabajo exhaustivo con otros diferentes métodos, ya sea para clasificar las características, entrenar, seleccionar otras mejores, etc.

VI. REFERENCIAS.

[1] Jiana, M., Hongfei, L., Yuhai, Y.: A two-stage feature selection method for text categorization, Computers & Mathematics with Applications, Volume 62, Issue 7. (2011,)

[2] Kent, John T.: Information gain and a general measure of correlation, Biometrika, Volume 70, Issue 1. (1983).

[3] McHugh, M. L.: The Chi-square test of independence. Biochemia Medica, vol. 23. (2013).

[4] Osborne, M. R.: Fisher's Method of Scoring. International Statistical Review / Revue Internationale De Statistique, vol. 60, no. 1. (1992).

[5] Adler, J., Parmryd, I.: Quantifing colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. (2010)

[6] Sánches-Maroño, N., Caamaño-Fernandez, M., Castillo, E., Alonso-Betanzos, A.: Functional Networks and Analysis of Variance for Feature Selection. (2003)

[7] Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection, Pattern Recognition Letters. (1994)