# Análisis de bioseñales y factores de riesgo en estudiantes universitarios*

Cervantes Rubí Brandon

Universidad Nacional Autónoma de México, CDMX, México
brandon.ceru@gmail.com

**Abstract.** Nowadays, there it's been collected a huge amount of data from different kinds of devices and all that data is stored in repositories around the world by companies and governments, among others. Smart watches are a device that generates a large amount of this data which, depending on the manufacturing, measures different fiscological variables known as biosignals in medicine fields (Heart Rate, Stress Levels, SpO2, BMI, etc.). The analysis of these biosignals has become a new world of study for companies because, with continuous monitoring and the creation of great models based on artificial intelligence, they can predict different kinds of user behaviour under different environments like sport, sleeping and short breaks with the objective of predicting catastrific events such as arrythms, a heart attack or hints about some disease. This paper is focused on analysing this kind of data but applied to student behaviour under academic environments like a class or speaking in public.

**Keywords:** Heart Rate · Time Series · Biosignals · Artificial Inteligence.

## 1 Introducción

Durante los últimos años la inteligencia artificial ha sido un campo en continuo crecimiento debido principalmente a múltiples factores tales como el poder de cómputo al que tenemos acceso, la cantidad de datos que se recolectan en las diferentes industrias y lo flexible que puede ser al presentar sus aplicaciones en diferentes áreas como la educación, el cine, el hogar, la salud entre muchas otras áreas de oportunidad.

El trabajo de investigación abordará el tema de la salud centrándose en los estudiantes de la Facultad de Ingeniería pertenecientes a la Universidad Nacional Autónoma de México (UNAM) como sujetos de prueba. Se recabaron y analizaron datos de los alumnos mientras realizan sus deberes escolares utilizando para esto relojes inteligentes los cuales se usan cada día más entre la comunidad universitaria con el objetivo de identificar posibles patrones y tendencias destacando las anomalías dentro del conjunto de datos e interpretándolas

* Universidad Nacional Autónoma de México, Facultad de Ingeniería.

tomando como referencia un marco teórico que define una serie de conceptos y limites los cuales son aceptados dentro de lo normal para determinar si pudieran ser perjudiciales en su salud.

Para lograr comenzar a desarrollar proyectos de este estilo es necesario seguir una metodología que siga toda una serie de pasos centrándose en la parte más básica y fundamental, los datos, es así como entra en juego el análisis exploratorio de datos o EDA (Exploratory Data Analisis) el cual sirve como un primer acercamiento a los datos; establecer patrones, anomalías y diferentes pruebas estadísticas para poder ver más de cerca cuáles son las propiedades que componen el conjunto de datos y de esta manera saber si vamos a poder contestar la pregunta inicial que vio nacer a nuestro proyecto.[10] Una vez se haya tenido este primer acercamiento se tienen que preparar los datos para después aplicar técnicas para la extracción de conocimiento o tendencias con el objetivo de interpretarlas.

## 2 Antecedentes

En la actualidad el uso de relojes inteligentes se está volviendo cada vez más popular dentro de la comunidad universitaria debido a que es un dispositivo que sirve principalmente como un extensible para nuestros teléfonos inteligentes haciendo nuestra vida más cómoda al no tener que buscar nuestro teléfono cuando se necesite enviar un mensaje, contestar una llamada o cambiar el contenido a reproducir, entre muchas otras funciones.

Sin embargo, una de las características más llamativas que ha llevado a los relojes inteligentes a presentar innovaciones importantes es incorporar la lectura de los bioseñales del usuario tales como el ritmo cardiaco, la oxigenación en la sangre o los niveles de estrés, entre otras, para llevar un monitoreo de como es que reacciona nuestro cuerpo a las actividades que realizamos en nuestra vida cotidiana. Lo cual representa un gran banco de información a nuestro alcance para realizar todo tipo de análisis y aplicaciones basadas en datos.

### 2.1 Bioseñales

Una señal por definición es una carga con información la cual viaja por algún medio (E.g. Aire, agua o cobre) dicha información pertenece a la fuente donde se originó en un principio, cuando los seres humanos somos la fuente de la que se origina la señal ésta se clasifica como bioseñal. La información que contienen estas señales se utiliza para explicar los mecanismos fisiológicos[2] del cuerpo humano.

Actualmente son de suma importancia en el área de la medicina y las ciencias biomédicas ya que la información que se transmite esta relacionada con el funcionamiento de los principales órganos del cuerpo ejemplos de esto son; el corazón con electrocardiogramas (ECG), el cerebro con electroencefalograma (EEG), los músculos con el electromiograma (EMG)[2] entre muchas otras más. La correcta adquisición, procesamiento y análisis de este tipo de señales tiene como objetivo dar un diagnóstico al paciente con base en el monitoreo de las bioseñales y, de ser necesario, un seguimiento al caso del paciente con algún tratamiento para evitar complicaciones.

### 2.2 Ritmo Cardiaco (HR)

El ritmo cardiaco o HR por sus siglas en inglés (Heart Rate) es una de los bioseñales más importantes del cuerpo humano ya que indica cuál es el número de veces que late el corazón a cada minuto que pasa, actualmente la gran mayoría de relojes inteligentes, incluyendo los que se utilizaran para esta investigación, están equipados con un sensor que mide el ritmo cardiaco de los usuarios con una técnica llamada fotopletismografía.

Para comprender como funciona esta técnica es importante entender cuál es la función del corazón en el cuerpo humano. Se trata del órgano el cual bombea sangre a todas las partes del cuerpo provocando una momentánea variación del volumen sanguíneo en nuestras venas y arterias, a esto le llamamos pulso sanguíneo lo podemos sentir en algunas partes del cuerpo como nuestras muñecas y cuello. La fotopletismografía utiliza un sistema óptico compuesto por un emisor y un receptor. Se tiene una luz verde que parpadea rápidamente (Emisor) apuntando a nuestra muñeca traspasando la piel y llegando a la sangre que es de color rojo, color que se obtiene al absorber la luz verde del emisor y reflejarla dándole un tono diferente dependiendo de la cantidad de sangre. Por último, interviene un fototransistor (receptor) cuya función es medir la cantidad de luz que está rebotando, logrando así detectar los cambios en los volúmenes sanguíneos pues a mayor cantidad de sangre en los vasos sanguíneos, mayor será la cantidad de luz que se refleje. [9] [2]

Los niveles normales de frecuencia cardiaca varían, pues no es lo mismo bombear sangre para oxigenar cada una de las células de un hombre que mide 1.9 m y pesa 85 kg. en comparación a uno de 1.6 m que pesa 58 Kg. [13] otras variables fisiológicas que afectan el ritmo cardiaco son el género y edad del usuario, pues al nacer nuestro corazón late mucho más rápido que cuando somos o personas de la tercera edad.

Está demostrado que los latidos de un corazón sano tienen un comportamiento no lineal y no periódico pues este tiene que estar en constante cambio dependiendo de los factores a los que se enfrente, envejecer puede ser una causa de que este comportamiento no lineal se vaya perdiendo con el tiempo incrementando el riesgo de arritmias o fibrilaciones lo cual puede llevar a fallos en el corazón.[13]

Considerando que el público objetivo del proyecto se encuentra al rededor de los 20 años de edad y un estudio enfocado a los pulsómetros realizado por el gobierno de México [8] un ritmo cardiaco normal se encontrará entre 60 y 100 pulsaciones por minuto en condiciones de reposo la cual puede incrementar dependiendo del género o si es que se realiza una actividad física elevándose a un determinado número máximo de pulsaciones por minuto.

Hombres:

$$HR_{max} = 220 - edad \tag{1}$$

Mujeres:

$$HR_{max} = 226 - edad \tag{2}$$

El sistema nervioso autónomo es el responsable de controlar, entre otras cosas, el ritmo cardiaco y la presión arterial destacando 2 componentes para la regulación de la respuesta cardiaca ante diferentes situaciones. El primero es el sistema simpático el cual es el responsable del gasto de energía en situaciones estresantes elevando el ritmo cardiaco y el segundo tiene por nombre sistema parasimpático que es el responsable de hacer todo lo contrario, es decir, almacenar y preservar la energía intentando mantener el ritmo cardiaco cuando el peligro o el estrés ha pasado. [13]

### 2.3 Estrés

El estrés es un fenómeno que se manifiesta en los seres humanos como consecuencia a la reacción que llega a tener una persona con su entorno. Durante su estudio al estrés se le ha relacionado con múltiples variables, entre ellas el género pues existen ciertos estereotipos establecidos en cada sociedad los cuales crean expectativas a cumplir demostrándose que las mujeres tienden a presentar mayores niveles de estrés que los hombres, las actividades realizadas que principalmente se dividen en actividades laborales o académicas pues son las que más tienden a presentar altos niveles de estrés, la edad es otra variable a tomar en cuenta ya que al crecer normalmente se incrementan las responsabilidades.

Las mediciones de estrés en la mayoría de las ocasiones se representan con un rango del 0 al 100, sin embargo, para este estudio se tomará el rango establecido por Huawei el cual abarca índices del 1 al 99 dividiéndose a su vez en 4 categorías; de 1 a 29 se clasifica como bajo, de 30 a 59 como normal, de 60 a 79 como medio y de 80 a 99 como alto. Los relojes utilizados realizan estas estimaciones de estrés con un algoritmo el cual toma la variabilidad del ritmo cardiaco como base.

Cuando estamos estresados ponemos a trabajar innecesariamente al sistema nervioso pues se activan mecanismos de defensa preparando al organismo para luchar o huir de ser necesario lo cual es útil cuando estamos en un peligro real, sin embargo, en la mayoría de las ocasiones esto no es así.

En el ámbito académico los estudiantes se enfrentan cada día a nuevos retos y exigencias para resolverlos lo cual demanda un gran desempeño cognitivo, físico y psicológico generando altos niveles de estrés cuando estos retos no traen consigo los resultados esperados. En consecuencia, los estudiantes pueden presentar síntomas de agotamiento, nerviosismo, evasión de sus responsabilidades, problemas relacionados con la perdida de cabello además de enrojecimiento y picazón en la piel. A nivel hormonal, internamente se liberan glucocorticoides en la sangre, entre ellas la adrenalina y la noradrenalina la cual repercute en los niveles de presión de las personas, por otra parte, se libera cortisol y los niveles de glucosa aumentan. [7] Una respiración lenta y profunda pudiera ayudar a contrarrestar estos fenómenos pues el diafragma se despliega lo cual ayuda a revertir procesos químicos y fisiológicos que son las consecuencias del estrés.

Todas las problemáticas mencionadas anteriormente repercuten de manera negativa tanto en su desempeño personal como profesional siendo el objetivo al analizar este parámetro el encontrar anomalías en los niveles de estrés tomando en consideración el contexto para evitar estos problemas.

### 2.4 SpO2

La saturación de oxígeno se refiere a la cantidad de oxígeno que circula por la sangre, enfocándose en mayor medida en los glóbulos rojos los cuales son los que se encargan de transportar la mayor cantidad de oxígeno desde los pulmones a todas partes del cuerpo de modo que se suministre suficiente a cada célula del cuerpo.

Los niveles de SpO2 se miden en porcentajes, donde un rango normal de oxigenación en sangre puede ir desde el 95% hasta el 100%. Huawei tiene un sistema en el que clasifican los niveles de SpO2 como buenos cuando este índice se encuentra por arriba del 90%, de riesgo moderado cuando se encuentra entre 70% y 89%, por último clasifica como niveles de alto riesgo índices por debajo del 70%. Un porcentaje por debajo de este rango puede indicar 2 cosas; una mala medición o principios de Hipoxemia la cual es una enfermedad que se presenta al momento de que los niveles de SpO2 bajan repercutiendo en la respiración y circulación.

Una medición errónea puede ser consecuencia de muchos factores, la condición física del paciente, alguna enfermedad, vellos en el brazo, tatuajes, la temperatura (Niveles óptimos de 15°C a 45°C) o la altitud a la que se esté realizando la medición ya que a mayor altitud la presión atmosférica es menor lo cual dificulta el intercambio de sangre entre los alveolos en los pulmones y la sangre disminuyendo en mayor medida los niveles de SpO2 sin que necesariamente sea un indicativo de algún trastorno de salud, sin embargo, estos cambios se comienzan a notar cuando se supera la altitud de los 2500 metros sobre el nivel del mar. Tomando en consideración que la Ciudad de México se encuentra a 2240 metros, estas condiciones no afectaran en las pruebas.

### 2.5 Trabajos relacionados

El continuo análisis de la variabilidad del ritmo cardiaco y sus variables relacionadas tiene diversas aplicaciones tales como; la predicción de arritmias, predicción de muertes espontáneas, monitoreo de hipertensión, pronostico de pacientes que necesitarán un transplante de corazón, etc.[13]. A continuación se muestra una tabla con los trabajos relacionados al tratamiento de bioseñales en diferentes ámbitos académicos y profesionales.

| Autor | Medios | Resultado | Bioseñales |
|---|---|---|---|
| Pyoung Won Kim (2018) | Sensores de actividad electrodermica y respuesta galvánica de la piel | Se desarrolló una aplicación con un semáforo de 5 colores que indica el nivel de compromiso o participación en clase | EDA (Electrodermal Activity) y GSR (Galvanic Skin Response). [6] |
| Heber Avalos Viveros (2020) | Dispositivos corporales inteligentes; relojes y teléfonos celulares | Se creó un marco de trabajo orientado al tratamiento de la bioseñal HR integrando diferentes fuentes para su análisis. [4] | HR (Hearth Rate) |
| Efraín Villegas Sánchez y Alfredo Armenta Espinosa (2022) | Arduino, microfibra, sensores de frecuencia cardiaca y temperatura | Se desarrolló el prototipo de una pulsera inteligente capaz de medir los niveles de estrés en los estudiantes. [5] | HR (Hearth Rate) y Temperatura |
| Santiago Adolfo Gómez Lagua (2022) | Raspberry Pi Zero, sensor de temperatura corporal, sensor de pulsioximetria | Se analizaron sistemas electrónicos para la adquisición de bioseñales y se utilizó SVM para clasificar casos de Covid [3] | SpO2 (Oxigenacion en la sangre) y temperatura. |
| Brian Buendia Sosa(2016) | Arduino, sistema de fotopletismografia, electromiografía y respuesta electrodermica | Se logró desarrollar un sistema que identificara la tendencia del estado de ánimo que tenía una persona con ayuda de sus bioseñales [1] | EMG (Actividad muscular), EDA (Electrodermal Activity) y HR (Hearth Rate) |

**Table 1.** Trabajos relacionados

### 2.6 Justificación

Aportar las bases que puedan llevar a elevar la calidad de vida de la población universitaria es el principal objetivo de esta investigación y se pretende lograrlo prestando atención a problemas internos que afectan el desempeño de los estudiantes, en este sentido el análisis de bioseñales nos puede ayudar a demostrar diferentes patrones y tendencias para identificar diferentes reacciones o cambios en el comportamiento que sirvan como indicativo para prevenir enfermedades físicas y psicológicas.

Es bien conocido que el ambiente universitario y las responsabilidades que éste conlleva pueden llevar a los estudiantes a someterse a situaciones complejas, tales como pocas horas de sueño, una buena cantidad de horas de estudio, hablar en público, enfrentarse a temas desconocidos, presentar proyectos o realizar exámenes por poner algunos ejemplos. Este tipo de condiciones enseña a los estudiantes a lidiar con situaciones bajo estrés ya que en un futuro tendrán que tomar decisiones en estas circunstancias, sin embargo, hay veces en las que se vuelve demasiado y no somos capaces de establecer un límite ya que es un fenómeno que no tiene un marco de referencia bien establecido [5], en consecuencia, todos estos factores pueden repercutir irremediablemente en nuestro cuerpo.

La característica de la adquisición de bioseñales pudiera ayudarnos a establecer precisamente este marco de referencia. En este trabajo se abordan estas lecturas con relojes inteligentes monitoreando el día a día de los estudiantes llevando a estos dispositivos a ser una potencial fuente datos crudos los cuales pueden ser tratados bajo ciertos marcos de adquisición y análisis identificando de esta manera actividades o hábitos personales que pueden llevar a nuestro cuerpo a responder negativamente llegando a repercutir en nuestra salud pues gracias a las bioseñales podemos identificar rápidamente si un individuo se encuentra saludable o por el contrario está enfrentando una situación complicada que podría afectar su condición de salud para poder canalizar a este estudiante con las atenciones y cuidados pertinentes.

## 3 Metodo

Se utilizará la ideología planteada en KRISP-MD la cual se ha vuelto un estándar dentro de los proyectos relacionados con la minería de datos resumiendo toda la metodología en 6 principales pasos. El primero de ellos es el entendimiento del problema donde se plantea bien que es lo que se quiere llegar a resolver con este trabajo, en este caso, es identificar cuáles son los posibles factores fisiológicos que pueden afectar el rendimiento de los estudiantes al desempeñar sus actividades académicas y las repercusiones que pudieran tener en su salud.

Una vez se tiene claro cuáles son las necesidades e implicaciones que plantea el problema podemos proceder al segundo paso el cual consiste en comprender los datos que son necesarios para resolver el problema implicando una sección en la recolección de los datos, la siguiente etapa es la preparación de los datos que consiste en aplicar un preprocesamiento a los mismos reduciendo posibles problemas de dimensionalidad, escala, tipos de dato, etc. de igual manera se entrara más a detalle en la sección de Limpieza y análisis de datos.

Como cuarto paso, una vez que se tienen los datos adecuados con la estructura adecuada se plantea pasar a la parte del modelado que consiste en seleccionar uno o más algoritmos que sean capaces de solucionar el problema planteado con los datos disponibles, cabe mencionar que si los datos por alguna circunstancia requieren de algún ajuste adicional es posible volver a la etapa anterior para corregir este detalle, la sección que tratara este punto tiene por nombre aplicación del algoritmo. Una vez que el modelo entrega un resultado podemos proceder a evaluarlo y determinar si cumple con los estándares establecidos para el problema plantado en la primera etapa, de no ser así será necesario volver a esta primera etapa para volver a iterar, pero con un concepto más claro de lo que se necesita para llegar a resolver el problema, si satisface los estándares de respuesta esperados podemos llegar a trabajar en alguna aplicación de uso común en el último paso.

### 3.1 Recolección de los datos

Para obtener los datos que corresponden a los bioseñales se ha pedido apoyo un grupo mixto de estudiantes de la facultad de ingeniería de la UNAM solicitándoles que porten una pulsera inteligente (Huawei Smart Band 7) durante ambientes académicos que pudieran someterlos a diferentes niveles de presión académica tales como clases teóricas, exámenes, prácticas, participaciones personales y al hablar en público. Al momento de realizar las pruebas los estudiantes cursaban diferentes cursos entre los cuales se encontraban; Bases de Datos, Minería de Datos, Sistemas Distribuidos, Sistemas Embebidos, etc.

El tiempo de cada una de las muestras varía de los 15 a los 120 minutos aproximadamente, durante este periodo el reloj registro mediciones de ritmo cardiaco

HR, niveles de estrés en una escala porcentual y niveles de oxigenación en la sangre de igual manera en una escala porcentual los cuales se sincronizan con la nube y posteriormente fueron descargados para su análisis.

Al final del periodo de pruebas se reunieron 59 muestras, cada una correspondiente a un día de actividad, estas muestras conforman la estructura de series de tiempo debido a que para cada registro se tiene la hora de inicio, la hora de fin y el valor de la bioseñal el cual varía entre las 3 variables mencionadas en el párrafo anterior.

### 3.2 Limpieza y análisis de los datos

Para trabajar con los datos recolectados se hizo un análisis exploratorio de datos con el objetivo de tener un primer acercamiento a como es que se comporta el conjunto de datos. En primera instancia cuál es su dimensionalidad la cual varía con cada muestra y sus tipos de datos para después obtener estadísticas tanto para las variables numéricas (Medias, Desviaciones, Cuartiles, etc.) como para las variables categóricas (Frecuencias de aparición, número de categorías, etc.).

El primer problema con estas series de tiempo era transformar o limpiar el conjunto de datos. Se partió de una serie de tiempo unidimensional la cual tenía valores para los 3 tipos de variables y se separó en una matriz de 4 dimensiones donde se pretende conocer el periodo y el valor de cada uno de los registros en la muestra de datos.

$$\begin{vmatrix} startTime & endTime & HR & Stress & SpO2 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{vmatrix} \tag{3}$$

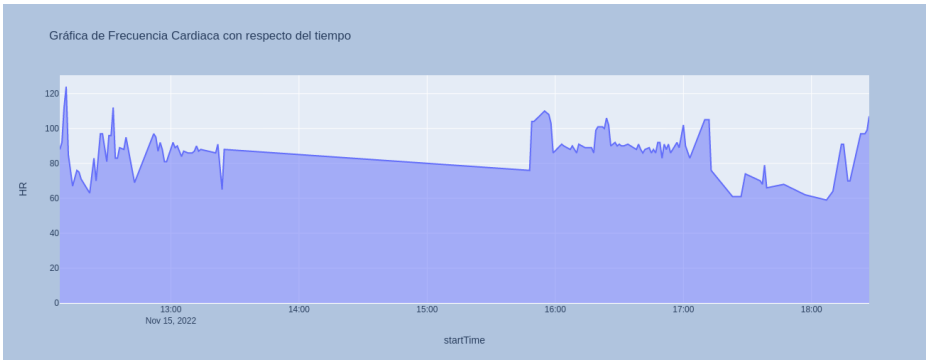


**Fig. 1.** Visualizacion del ritmo cardiaco para el dia 15 de Noviembre del 2022

Recordando que los latidos de un corazón sano tienen un comportamiento no periódico es necesario interpolar los valores correspondientes a las muestras debido a que parte de la estructura de las series de tiempo es que estas cuenten con un periodo constante entre sus elementos para obtener mejores resultados. Con esto en mente se partió del periodo mínimo observado entre el tiempo final y el tiempo inicial obtenido con los relojes el cual es de un minuto extendiendo el valor de la última muestra conocida hasta la subsecuente agregando con esto valores al conjunto de datos.

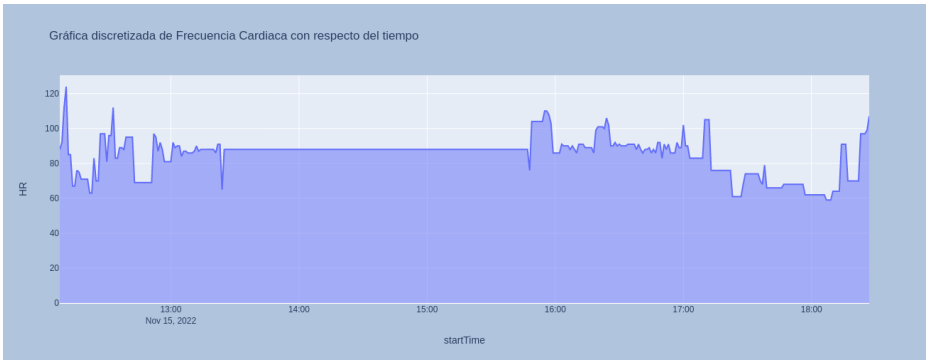


**Fig. 2.** Visualizacion de la figura 1 interpolada

Con nuestras series de tiempo interpoladas es necesario comenzar a separarlas por actividad, para este análisis se tomaron 4 tipos de actividad como punto de partida; clases normales, prácticas, exámenes y presentaciones en público. En la figura 2 podemos observar dos periodos de actividad los cuales podemos separar al dividirla por la mitad (El primero desde el inicio hasta las 14 hrs y el segundo desde las 15 hrs hasta el término), si hacemos este proceso iterativamente podemos ampliar que el conjunto de datos hasta las 84 tomas y tener la base para un modelo de clasificación supervisado.

A continuación se muestran gráficas de autocorrelación [14] para algunas de las series de tiempo recolectadas correspondientes a cada una de las actividades (Se utilizó la librería statsmodels para python), sin embargo todas comparten la misma estructura pues es común que para el primer retraso el nivel de correlación sea alto y positivo esto significa que si en un tiempo t se tiene un valor de ritmo cardiaco es muy probable que este valor afecte a los primeros ritmos subsecuentes positivamente (seguirán incrementando o decrementando). Sin embargo, esta característica se pierde o se vuelve despreciable cuando avanzamos en el tiempo,

por ejemplo, el ritmo cardiaco del minuto 10 nos representa una fuerte correlación con el ritmo cardiaco del minuto 11 y 12 pero esta correlación va disminuyendo a medida de que avanzamos en el tiempo y para el minuto 15 o 20 el ritmo cardiaco del minuto 10 ya no tendrá prácticamente ninguna influencia en estas muestras. En resumen, las muestras más cercanas temporalmente son aquellas que tendrán más influencia entre sí.

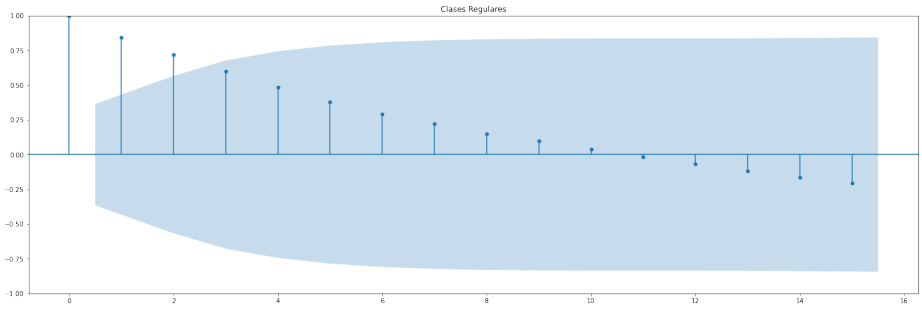


**Fig. 3.** Grafica de autocorrelacion para serie temporal en clases

Cuando tenemos este tipo de comportamiento en nuestras series de tiempo podemos decir que son estacionarias, esto puede deberse al periodo con el cual están siendo tomadas las muestras pues con un periodo más corto (Digamos 1 seg. en lugar de 1 min.) este comportamiento podrá cambiar. La característica más importante de una serie de tiempo estacionaria es que estas son estables a lo largo del tiempo, es decir la media y la varianza son estables también lo cual ayuda a construir modelos de regresión.[15]

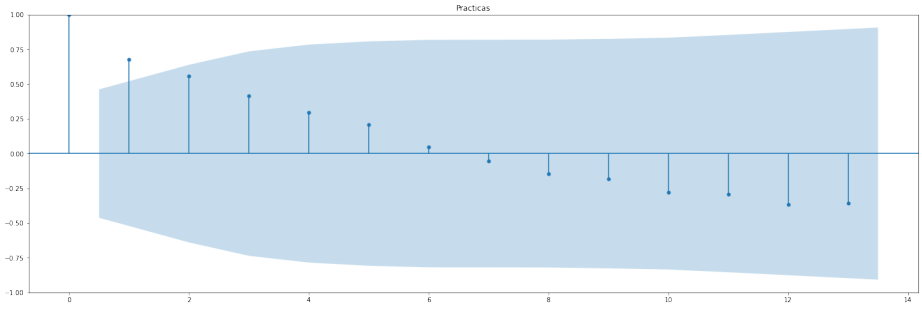


**Fig. 4.** Grafica de autocorrelacion para serie temporal en presentaciones

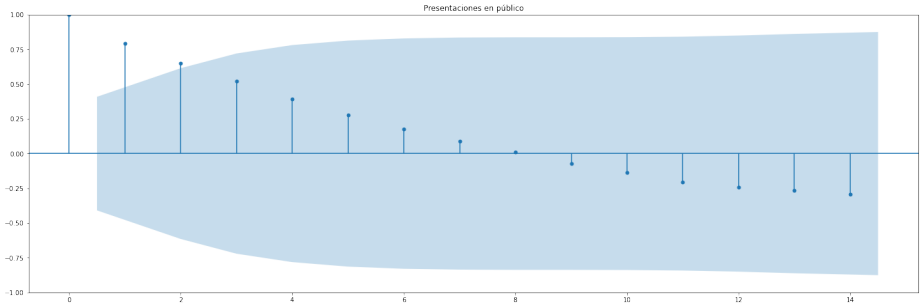


**Fig. 5.** Grafica de autocorrelacion para serie temporal en examenes

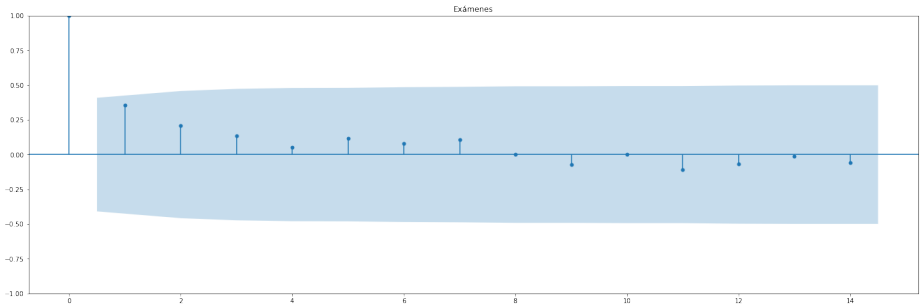


**Fig. 6.** Grafica de autocorrelacion para serie temporal en clases

Con estas gráficas también podemos dar un aproximado del periodo de cambio en el patrón del ritmo cardíaco en los estudiantes el cual varía dependiendo de la actividad que estén realizando. Se encontró que las clases normales tienen el periodo de tiempo más largo, de 10 minutos, después de este tiempo es muy poco probable que el ritmo con el que se inició la clase se mantenga, pero posterior a este tiempo el patrón será muy similar por un nuevo periodo de tiempo, de ahí que sea una serie estacionaria. En las prácticas el tiempo de acoplamiento a la clase se disminuye a un aproximado de 7 minutos, para el caso de las exposiciones este tiempo varía entre los 6 y los 10 minutos según las muestras y por último en el caso de los exámenes no podemos decir con certeza un tiempo

el cual sea determinante debido a que el número de muestras recolectadas no puede considerarse significativo, sin embargo, para las muestras recolectadas el tiempo de cambio en la correlación es el más bajo teniendo un promedio de 5 minutos.

Por último para visualizar las estadísticas de una mejor manera es factible utilizar gráficas como histogramas, gráficos de caja o mapas de calor. En el caso de las variables numéricas es importante detectar valores atípicos ya que pueden indicar anomalías dentro del conjunto de datos y de ser posible corregirlos para prevenir eliminarlos u omitirlos. A continuación se muestra un gráfico de cajas enfocado a analizar los valores de la frecuencia cardiaca (HR) separándolos por el nivel de estrés observado en la muestra. En él podemos observar que el número de pulsaciones por minuto en el usuario repercute directamente en el nivel de estrés calculado pues la media en cada uno de los diagramas de cajas se va incrementando conforme lo hace su nivel de estrés.


**Fig. 7.** Valores de frecuencia cardiaca separados por niveles de estres

Aunque es posible que éste no sea el único factor ni el más determinante para calcular los niveles de estrés si es una variable correlacionada en cierta manera con la cantidad de pulsaciones por minuto en el usuario, parámetro que al crecer continuamente representa un patrón de riesgo que deberá de ser atendido.

### 3.3 Aplicación del algoritmo

En esta sección se plantearán 2 tipos de modelos, uno de clasificación y uno de regresión. Dichos modelos se entrenarán con base en las series de tiempo resultantes del punto anterior y con ayuda de la librería sktime disponible en python la cual nos brinda de algoritmos de clasificación y regresión los cuales

toman a la entrada no solo una serie de tiempo, sino un conjunto de series de tiempo de múltiples las cuales pueden estar compuestas por múltiples variables.

Los bosques aleatorios son algoritmos que buscan generalizar más las soluciones que nos pueden entregar los árboles aleatorios mediante la combinación de varios de estos árboles en la misma estructura de tal forma que cada uno de ellos aporte una solución similar para después llegar a un consenso. El árbol de decisión es un algoritmo el cual toma una serie de variables independientes como entrada y son analizadas una por una en cada uno de los diferentes niveles del árbol partiendo de una raíz para obtener el determinado valor de una variable dependiente. En este caso las variables independientes del modelo son los valores de las bioseñales (HR, SpO2 y estrés) mientras que la variable dependiente para la clasificación corresponde a la etiqueta la cual se asigna a la serie de tiempo que se proporcione a la entrada o en el caso de la regresión corresponde al próximo valor del ritmo cardiaco.

K-Nearest Neighbours es un algoritmo de clasificación cuyo funcionamiento se basa en asignar una etiqueta de acuerdo con la muestra o el conjunto de muestras más cercanas consideradas en el entrenamiento, para ello será necesario calcular la distancia de todos los puntos de entrenamiento con la de las muestras a pronosticar con el objetivo de determinar cuál de ellas es la menor y de esta manera asignar una clasificación. El valor de K índica el número de vecinos a considerar, de esta manera si k = 1 entonces la etiqueta del vecino más cercano es la que será asignada [12], si k 1 entonces se entra en un problema de consenso ya que se consideraran a los K-Vecinos más cercanos los cuales asignaran la etiqueta por mayoría. En este caso las variables independientes del modelo vuelven a ser el ritmo cardiaco, la oxigenación en la sangre y los niveles de estrés para asignar una etiqueta correspondiente a la actividad como variable dependiente.

**Clasificación** Una vez que los datos se encuentran interpolados y etiquetados podemos comenzar en primera instancia con un modelo de clasificación el cual pretende generar etiquetas para futuras muestras. En particular la etiqueta que se mostrará a la salida corresponde a la actividad que se está realizando; clases normales, prácticas, exámenes o presentaciones en público. Para generar el modelo se considerarán 2 algoritmos; Bosques aleatorios y K-Nearest Neighbours.

Ambos modelos proponen una variacion ya que la entrada en esta ocasión no es una matriz convencional de datos donde cada una de las celdas solamente contiene un valor sino que para la estructura del problema todas las celdas contienen una serie de tiempo, es decir, que para el primer registro de entrenamiento se cuenta con 3 series de tiempo, una para la frecuencia cardiaca, una para la oxigenación en la sangre y una para los niveles de estrés donde todas y cada una de estas 3 series de tiempo tienen la misma longitud, sin embargo, esto puede último puede variar con respecto del conjunto de 3 series de tiempo del segundo registro y así sucesivamente. Una consideración importante a recalcar es que para trabajar con los modelos es necesario que todas y cada una de las series

de tiempo sean del mismo tamaño, tanto para las series de tiempo que servirán para entrenar al modelo como las series de tiempo que harán uso del mismo pues lo podemos ver como un hiperparametro más el cual afecta a ambos modelos pues a mayor tamaño en las series de tiempo el conjunto de datos se hace más pequeño, ya que tenemos que ir descartando las series de tiempo que no cumplen con este tamaño establecido.

Otra de las consideraciones importantes a tomar en cuenta antes de entrenar al modelo es el ajuste de hiperparametros, para esta investigación se realizó una aplicación web la cual ayuda al usuario con la carga de datos tanto para el análisis exploratorio de datos como para el entrenamiento de algoritmos de clasificación tomando como base el formato que nos proporciona la plataforma de Huawei (Json). Como parte del entrenamiento del modelo el usuario puede modificar los hiperparametros de manera más intuitiva para probar con varias configuraciones, en el caso de los bosques aleatorios se permite establecer el número de árboles a considerar en el consenso mientras que por parte del algoritmo de K-Nearest Neighbours se permite modificar el número K de vecinos y el tipo de distancia que se utilizará. Como salida de este conjunto de configuraciones tenemos el modelo cuya precisión puede variar, sin embargo, la más acertada fue de 80.0% con el modelo de bosques aleatorios y 93.3% con el modelo de K-Nearest Neighbours.

| Modelos de Clasificación | | | |
|---|---|---|---|
| No. de Árboles | Score | Tipo de Distancia | Score |
| 50 | 0.733 | euclidean | 0.933 |
| 100 | 0.8 | squared | 0.933 |
| 200 | 0.8 | dtw | 0.8 |
| 250 | 0.733 | ddtw | 0.933 |
| 300 | 0.733 | wdtw | 0.866 |
| 350 | 0.733 | wddtw | 0.933 |
| 400 | 0.8 | lcss | 0.933 |
| 450 | 0.8 | edr | 0.866 |
| 500 | 0.8 | erp | 0.933 |
| 800 | 0.8 | msm | 0.933 |
| 1000 | 0.8 | twe | 0.933 |

**Table 2.** Tabla de indices de presicion para los modelos de clasificación

Por último es importante recalcar que ambos modelos pueden trabajar con una o múltiples variables, es decir, se parte de un escenario en el que entrenamos al modelo con un conjunto de series de tiempo de frecuencia cardiaca siendo esta nuestra única variable para pronosticar las etiquetas. Después podemos probar con diferentes variables para pronosticar la etiqueta considerando todas las posibles combinaciones entre frecuencia cardiaca, estrés y oxigenación en la sangre.

**Regresion** En el modelo de regresión se consideró un bosque aleatorio para generar una solución que fuera capaz de predecir el próximo valor del ritmo cardiaco lo cual puede tener diversas aplicaciones ya que con este monitoreo se pueden detectar tendencias a futuro considerando el historial de muestras en la sesión que se esté llevando acabo y de esta manera detectar posibles anomalías que representen algún riesgo en la salud del usuario, para ello tomara a la entrada una serie temporal de una longitud fija la cual representa el histórico del ritmo cardiaco en minutos anteriores y construir aun modelo para establecer un mejor ajuste en la salida. Inicialmente también se había considerado un modelo de KNN para regresión, sin embargo, los resultados obtenidos fueron muy decepcionantes entregando índices de precisión por debajo del 10%.

En este modelo los posibles hiperparametros que se tienen que revisar son la semilla de aleatoriedad que se utilizara para generar un modelo replicable en el tiempo y el número de estimadores o de árboles a utilizar, tanto para este modelo como para el modelo de clasificación se fueron variando ambos parámetros con los valores mostrados en la tabla para llegar al mejor modelo posible con los datos disponibles. El mejor modelo para el pronóstico de ritmo cardiaco se obtuvo utilizando 200 estimadores llegando a un score de 80%, por otra parte el mejor modelo para pronosticar los niveles de estrés se obtuvo de igual manera con 200 estimadores llegando a un índice de precisión del 96.9% y por último el mejor modelo para pronosticar los niveles de SpO2 se obtuvo al utilizar cualquier cantidad de estimadores con un índice de precisión del 99.7% el cual podría considerarse muy alto llegando a caer en un posible sobreajueste, esto se debe a que los niveles de SpO2 cambian más lentamente en comparación con el ritmo cardiaco o el estrés lo cual indica una menor variabilidad en el SpO2, por ende el rango de posibles valores se ve recortado y con ello la posibilidad de error por parte del modelo.

Cabe resaltar que a diferencia de los modelos de clasificación, la implementación de un modelo de regresión con bosques aleatorios que proporciona sktime solo nos permite construir modelos con una sola variable para las series de tiempo, es por ello que se construyeron 3 modelos, uno para pronosticar cada una de las bioseñales analizadas. Como ultima consideración se observó que, en ambos tipos de modelo, el valor de la semilla de aleatoriedad tanto para dividir los datos como para entrenar al modelo es determinante en el índice de precisión por lo cual se iteró sobre las 50 primeras semillas para obtener el mejor indice.

| Modelos de Regresión | | | | | |
|---|---|---|---|---|---|
| No. de Árboles HR | Score | No. de Árboles Éstres | Score | No. de Árboles SpO2 | Score |
| 50 | 0.774 | 50 | 0.966 | 50 | |
| 100 | 0.799 | 100 | 0.962 | 100 | |
| 200 | 0.809 | 200 | 0.969 | 200 | |
| 250 | 0.808 | 250 | 0.964 | 250 | |
| 300 | 0.803 | 300 | 0.962 | 300 | |
| 350 | 0.807 | 350 | 0.965 | 350 | 0.997 |
| 400 | 0.794 | 400 | 0.964 | 400 | |
| 450 | 0.798 | 450 | 0.965 | 450 | |
| 500 | 0.803 | 500 | 0.964 | 500 | |
| 800 | 0.797 | 800 | 0.964 | 800 | |
| 1000 | 0.793 | 1000 | 0.963 | 1000 | |

**Table 3.** Tabla de indices de presicion para los modelos de regresión

## 4 Resultados

A lo largo de la investigación se encontraron diferentes indicativos del comportamiento de las diferentes bioseñales en los estudiantes. La primera de ellas tiene que ver con la correlación que existe entre el aumento del ritmo cardiaco en paralelo con en el nivel de estrés que se puede observarse en la figura 7, en este gráfico cada una de las cajas representa una medición de estrés diferente y el eje Y representa el ritmo cardiaco de tal forma que se puede dibujar una recta creciente que cruce por las cajas indicando una relación positiva entre estas dos variables.

Se logró aproximar un posible tiempo de acoplamiento que tienen los usuarios ante alguna actividad académica, esto con ayuda de las gráficas de autocorrelación en las series de tiempo. Es decir, el tiempo que le podría tomar a un usuario cambiar por completo el patrón de comportamiento en sus bioseñales básicas lo cual podrá llegar a representar un cierto incremento en el nivel de atención a la actividad que se esté realizando. El análisis se hizo con respecto al ritmo cardiaco, ya que las series de tiempo recolectadas para esta bioseñal son aquellas con la mayor variabilidad en sus datos, la siguiente bioseñal que más varía son los niveles de estrés y por último los niveles de SpO2 cambian en un mayor tiempo comparado con los dos anteriores.

Se construyeron 2 tipos de modelos; clasificación y represión, para cada uno de ellos se variaron diferentes hiperparametros como el número de árboles en el caso de los bosques aleatorios y el tipo de distancia utilizada para KNN hasta llegar al mejor modelo. Comenzando con los modelos de clasificación se tiene la figura 8.

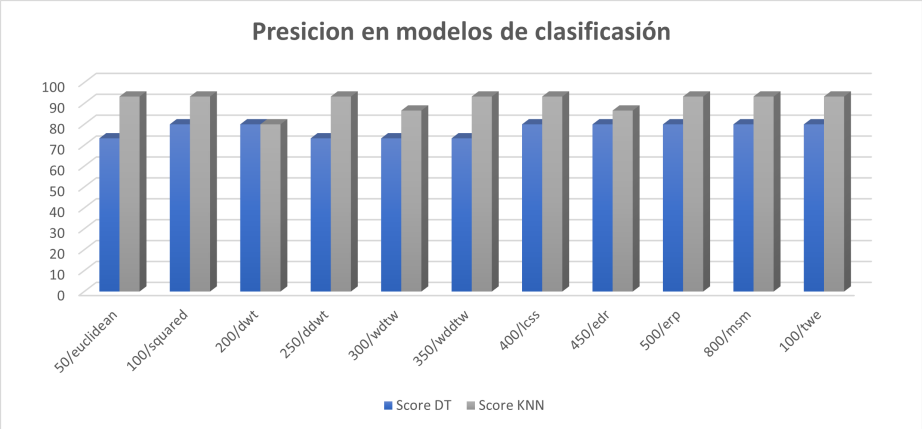


**Fig. 8.** Indices de precision de los modelos de clasificación

En la figura 8 podemos ver los índices de precisión para los modelos de bosques aleatorios (Azul) y KNN (Gris). Si los comparamos es evidente que KNN además de que se comporta mejor en términos de rendimiento comparándolo con los bosques aleatorios también ofrece un buen desempeño con la mayoría de las distancias disponibles siendo una de las más destacadas la distancia euclidiana. Por otro lado los modelos de bosques aleatorios ofrecen un buen rendimiento desde los 100 y 200 estimadores, indice que no incrementa en comparación con 1000 estimadores. Por estas razones en el caso de la clasificación de las series de tiempo es mejor utilizar un modelo de K-Nearets Neighbours.

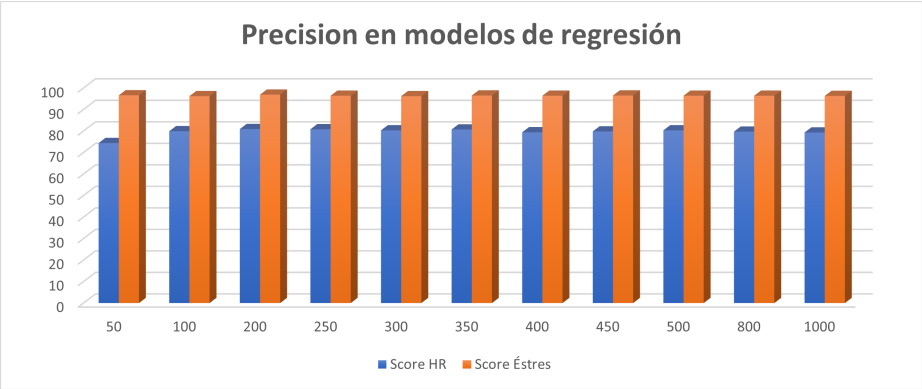


**Fig. 9.** Indices de precision de los modelos de regresión

Podemos observar una gráfica similar pero para los modelos de regresión en la figura 9, recordando que en la sección anterior se crearon modelos de bosques aleatorios para pronosticar cada una de las bioseñales en el análisis. Observamos que los modelos relacionados con el pronóstico del estrés(Naranja) tienden a ser más precisos que los modelos destinados a pronosticar el ritmo cardiaco (Azul), además de ser más precisos (Los modelos de estrés) también se estabilizan con pocos estimadores necesitando solamente de 50 para entregar resultados satisfactorios mientras que los modelos para pronosticar el ritmo cardiaco necesitan de más estimadores, alrededor de 100, lo cual repercute directamente en el tiempo y el poder de cómputo necesario para ejecutar los modelos.

Con esto en consideración podemos decir que los modelos de ritmo cardiaco y niveles de estrés son tienen un buen desempeño al momento de detectar posibles cambios en el comportamiento de los usuarios, cambios que pudieran llegar a una complicación de salud. Sin embargo, para el modelo de SpO2 detectar estos saltos en sus valores se vuelve más complicado, ya que es la bioseñal con un menor índice de variabilidad. Para corregir este detalle podríamos utilizar un sensor dedicado o bien ampliar el periodo de muestra por encima de los 120 minutos.

## 5 Conclusiones y Trabajo Futuro

El análisis de bioseñales es un campo de estudio el cual se puede extender tanto como nuestros sensores y voluntarios lo permitan, es claro que a mayor número de datos los modelos tendrán un mejor desempeño, sin embargo, en este trabajo se plantean algunas bases de análisis y uso de las bioseñales para la generación de diferentes modelos mostrando que son un tipo de dato el cual se puede adaptar para diferentes aplicaciones, en este artículo se revisaron clasificación y pronostico como punto de partida. No obstante para obtener un buen resultado en estos modelos es fundamental pasar la mayor parte del tiempo limpiando los datos, es decir, dándole diferente forma a sus estructuras dependiendo del análisis a realizar, eliminando o interpolando valores según sea necesario, entre otros procedimientos.

Los modelos de clasificación resultaron ser más complicados de implementar por toda la preparación de los datos que implica, sin embargo, los resultados fueron los más satisfactorios llegando a obtener modelos con un buen desempeño en la práctica el cual nos ayudará a predecir las etiquetas de nuevas muestras a agregar en el conjunto de datos. Un detalle importante a tomar en cuenta al momento de realizar el modelado es la variación de los hiperparametros los cuales son completamente dependientes del modelo y repercuten en el resultado final.

Analizando el desempeño de los modelos de regresión podemos decir que para entregar predicciones relacionadas con nuevos valores de estrés y ritmo cardiaco los modelos construidos con bosques aleatorios tienen un buen desempeño considerando el contexto bajo el cual se están aplicando, sin embargo, no podemos decir lo mismo para el modelo destinado a predecir niveles de SpO2 ya que para llegar a un mejor resultado tendremos que ampliar el número de series de tiempo a tomar en cuenta lo cual nos brinda la posibilidad de implementar nuevos algoritmos relacionados con el campo del Deep Learning.

Como trabajo a futuro podemos destacar varios detalles los cuales se fueron encontrando a lo largo de la elaboración de este artículo, el primero de ellos tiene que ver con la aplicación, comparación y análisis de nuevos algoritmos de Inteligencia Artificial, por ejemplo, uno de los que se tuvieron que dejar fuera por cuestiones del número de muestras son los relacionados con las Redes Neuronales Recurrentes (RNN), en particular las Reded Neuronales Long Short-Term Memory (LSTM) las cuales se suelen usar cuando se trabaja con series de tiempo debido a que tienen buenos resultados.

Por otro lado, si el número de muestras se incrementa en consecuencia las estadísticas serán más detalladas y describirán mejor los patrones de comportamiento. Sin embargo este aumento en el número de muestras también tendrá implicaciones en el poder de cómputo necesario para entrenar los algoritmos.

### 5.1 Consideraciones finales

La identidad de todos los estudiantes que participaron en el proyecto será confidencial a menos que se indique lo contrario,se les solicitara el consentimiento para el uso de sus datos además de que tienen la opción de solicitar sus datos en cualquier momento que deseen. Las pruebas se harán a estudiantes de la Facultad de Ingeniería de la UNAM en diferentes materias (e. g. Bases de Datos, Minería de Datos, Inteligencia Artificial).

En cuanto al tratamiento de los datos, estos en un principio vienen en un formato JSON los cuales son procesados para convertirlos a csv y que estén disponibles para cargarlos a la aplicación web donde se mostraran los resultados del análisis.

## References

1. Buendia, B.: Desarrollo de un sistema para caracterizar el estado de la alegría mediante bioseñales. UNAM –Dirección General de Biblioteca, México, 33–89 (2016).
2. Ortega E. A, Gonzáles T. Y. y Mendoza R. M. A: Red de sensores inalámbricos para el monitoreo de bioseñales. Revista Cubana de Transformación Digital, Cuba, 3–6 (2016).
3. Gómez, S.: Sistema electrónico de monitoreo de bioseñales para el diagnóstico médico de covid-19 en personas mediante inteligencia artificial. Universidad Técnica de Ambato, Ecuador, 35–104 (2022).

4. Avalos, H.: Framework para el tratamiento de bioseñales de tipo electrocardiografía obtenidas a través de dispositivos corporales inteligentes. Universidad Veracruzana, México, 11–74 (2020).
5. IMER, Estudiantes del IPN desarrollan pulsera para medir el estrés escolar https://noticias.imer.mx/blog/estudiantes-del-ipn-desarrollan-pulsera-para-medir-el-estres-escolar/. Last accessed 2 Nov 2022
6. Brainsigns, Respuesta galvánica de la piel (GSR) https://www.brainsigns.com/es/science/s2/technologies/gsr. Last accessed 2 Nov 2022
7. Berrío, N.: Estrés Académico. Revista de Psicología Universidad de Antioqia, Colombia, 1–18 (2011).
8. Profeco: Pulsómetros. Revista del Consumidor, México, 1–18 (2011).
9. ThinkBig, Fotopletismografía, la técnica detrás del éxito de los relojes inteligentes https://blogthinkbig.com/fotopletismografia-telefonicar. Last accessed 12 Oct 2022
10. Towards Data Science, What is Exploratory Data Analysis? https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. Last accessed 06 sep 2022
11. Kramer, O.: Dimensionality Reduction with Unsupervised Nearest Neighbors, Alemania, 13–23 (2013).
12. Breiman, L.: Random Forests, California, 1–10 (2001).
13. García, C.: Heart Rate Variability Analysis with the R package RHRV, Estados Unidos, 1–27 (2017).
14. Funciones de autocorrelación y autocorrelación parcial https://www.ibm.com/docs/es/spss-modeler/saas?topic=data-autocorrelation-partial-autocorrelation-functions. Last accessed 04 dic 2023
15. Introducción a Series de Tiempo. Estadisticas PR, Puerto Rico, 2–6 (2011).