**MACHINE LEARNING PARA LA PREDICCIÓN DE LA SUPERVIVENCIA DE LOS PACIENTES CON INFECCIÓN GRAVE POR EL VIRUS COVID – 19**

**PRESENTADO POR:**

**CARLOS ALBERTO GUERRERO NAVARRO**

**DOCENTE:**

**ÁLVARO OÑATE BOWEN**

**UNIVERSIDAD POPULAR DEL CESAR**

**FACULTAD DE INGENIERIAS Y TECNOLOGIA**

**INGENIERÍA DE SISTEMAS**

**VALLEDUPAR – CESAR**

**2020 – 2**

**INTRODUCCIÓN**

Actualmente el Machine Learning (ML) se ha vuelto indispensable en todos los ámbitos, la necesidad de que las maquinas aprenda y actúen de forma inmediata por sí solas se ha convertido en un completo reto para los que trabajan en esta área de los sistemas computacionales. Una definición formal de ML viene siendo básicamente la capacidad que tienen las máquinas (los ordenadores) para aprender y llegar a conclusiones o lo que es lo mismo, aportar soluciones a problemas concretos y generar conocimiento a partir de la información por el Big Data. Concretamente en el ámbito de la salud, el objetivo del ML es dotar a las herramientas informáticas de un sistema que les permita procesar los datos de información sanitaría (procedentes todos de la historia clínica) de tal forma que puedan realizar una serie de acciones relacionadas con la atención y gestión de la salud [1].

Desde que se conoció el primer caso del virus COVID – 19, se ha iniciado una ardua labor por la disminución de los contagios y contraer un poco la enfermedad; conforme se han ido experimentado las medidas de confinamiento a lo largo y ancho del planeta, el ritmo de trabajo de la comunidad científica para entender la naturaleza y los mecanismos de propagación del virus ha crecido hasta niveles frenéticos, con el objetivo de encontrar una vacuna cuanto antes. Sin embargo, un hecho que está pasando bastante desapercibido es la contribución que la tecnología está haciendo a todo ese proceso, lleva trabajando sin descanso entre bastidores para ofrecer soluciones que permitan superar las limitaciones intrínsecas del conocimiento humano en esta abrumadora tarea. [2].

El ML posee múltiples modelos que permiten aplicarse para poder obtener análisis e información que es de gran importancia, uno de ellos son los modelos de predicción que han servido de herramienta de apoyo a los profesionales para evaluar a los pacientes con COVID – 19 y asignar recursos de salud. Estos modelos de predicción combinan diversas variables o características para estimar el riesgo de que las personas se infecten o experimente un mal resultado de la infección (en definitiva, que tengan un mal pronóstico) [3]. Este proyecto busca aplicar las técnicas de ML a los datos de las historias clínicas registradas de 485 pacientes en la facultad de medicina del Hospital Tongji recogida entre el 10 de enero y el 18 de febrero de 2020 de Wuhan, China.

**ESTADO DEL ARTE**

La aplicación de ML en el virus COVID – 19 es un área de investigación completamente reciente, que busca servir de ayuda a los trabajadores de la salud para poder brindar una mejor calidad de atención y de vida a los pacientes que se encuentran infectados gravemente, a continuación se presentan unos estudios de investigación en donde trabajaron el ML.

Los autores Yao, Zhang, Duan, etc. [4], realizaron un estudio sobre la detección de gravedad para pacientes con enfermedad de coronavirus 2019 en donde utilizan un modelo de aprendizaje automático basado en análisis de sangre y orina de aquellos con síntomas leves, utilizaron varios algoritmos de clasificación para construir los modelos de predicción de pacientes gravemente enfermos; modelos como regresión logística predictiva, máquina de vectores de soporte, Random Forest, K Nearest Neirgbor (vecino más cercano) y por último el algoritmo AdaBoost. Los datos experimentales que utilizaron demostraron fuertes correlaciones con la severidad de COVID – 19. Y el modelo final de detección de gravedad logró una precisión de 0.8148 en el conjunto de datos de pruebas independientes.

En la Universidad Huazhong de Ciencia y tecnología los autores Yan, Zhang, Gonçalves, desarrollaron un modelo basado en aprendizaje automático para la predicción de supervivencia en pacientes con infección grave por COVID – 19 [5]. Este estudio aprovecha una base de datos de muestras de sangre de 404 pacientes infectados en la región de Wuhan, China, para identificar biomarcadores predictivos cruciales de la gravedad de la enfermedad. Utilizaron el algoritmo de clasificación XGBoost como modelo predictivo, este algoritmo se basa en utilizar arboles de decisión. En general este documento sugiere una fórmula simple y operable para predecir rápidamente a los pacientes con el riesgo más alto, lo que les permite priorizar y potencialmente reducir la tasa de mortalidad.

En la Universidad de Qatar, Malasia y Bangladesh los autores Chowdhury, Rahman, Khandakar, etc., desarrollaron una herramienta de alerta temprana para predecir el riesgo de mortalidad de los pacientes con COVID – 19 mediante aprendizaje automático [6]. En este estudio se realizo un análisis retrospectivo en 375 pacientes con COVID – 19 ingresados al Hospital Tongji (China), utilizaron modelos predictivos de clasificación y regresión en donde analizaron el AUC (área bajo la curva), esto lo trabajaron con el algoritmo XGBoost que mostro la relación entre la enfermedad gravedad y los datos clínicos. Se identificaron diez predictores por el algoritmo Multi-Tree XGBoost como probabilidad de muerte que son basados en los datos al momento de ingreso.

En India, los autores A. Karthikeyan, A. Garg, P. K. Vinod, U. Priyakumar realizaron un sistema de soporte de decisiones clínicas basado en aprendizaje automático para la predicción temprana de la mortalidad por COVID – 19 [7]*.* En donde se busca identificar los factores claves de mortalidad que proporcionan alta precisión y consistencia para optimizar la estrategia de tratamiento del paciente. El estudio utiliza métodos de aprendizaje automático para identificar una poderosa combinación de cinco características que ayudan a predecir mortalidad con 96% de precisión: neutrófilos, linfocitos, lactato deshidrogenasa (LDH), alta sensibilidad Proteína C reactiva (hs-CRP) y edad. Se compararon varios algoritmos de aprendizaje automático para lograr una alta precisión constante a lo largo de los días que abarcan la enfermedad. Las pruebas sólidas con tres casos confirman el fuerte rendimiento predictivo del modelo propuesto. El modelo predice con una precisión del 90% como tan pronto como 16 días antes del resultado. Este estudio ayudaría a acelerar el proceso de toma de decisiones en sistemas de salud para tratamientos médicos enfocados de manera temprana y precisa

Otro trabajo de investigación desarrollado por los autores Yan, L., Zhang, H., Goncalves, J. fue un modelo de predicción de mortalidad interpretable para pacientes con COVID-19, [8] Este estudio ha identificado tres indicadores (LDH, hs-CRP y linfocitos), junto con una ruta clínica, para la predicción del pronóstico de COVID-19. Hemos desarrollado un modelo basado en el aprendizaje automático XGBoost que puede predecir las tasas de mortalidad de los pacientes con más de 10 días de anticipación con más del 90% de precisión, lo que permite la detección, la intervención temprana y potencialmente una reducción de la mortalidad en pacientes con COVID-19.

**METODOLOGÍA CRISP – DM**

CRISP – DM es un método probado para orientar trabajos de minería de datos, sus siglas en ingles significan Cross-Industry Standard Process for Data Mining.

* Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
* Como modelo de proceso, CRISP – DM ofrece un resumen del ciclo vital de minería de datos.

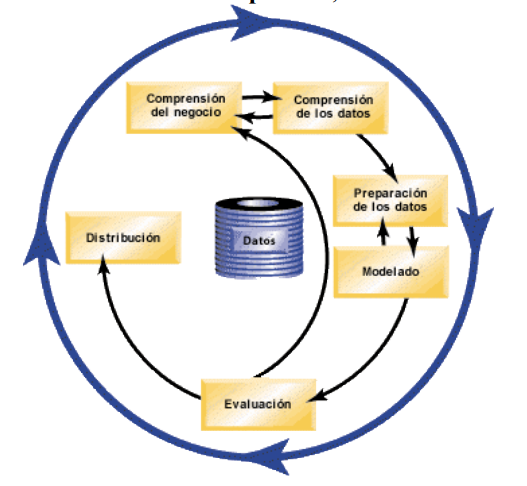


Figura . Metodología CRISP – DM

La figura 1, muestra el ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de fases no es escrita. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

**IMPLEMENTACIÓN DE LA METODOLOGÍA**

1. **FASE 1: COMPRENSIÓN DEL NEGOCIO**

Es la más importante porque muestra las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.

* 1. **ENFERMEDAD DEL CORONAVIRUS (COVID – 19) EN EL MUNDO**

Desde que se detecto el primer caso de COVID – 19, el mundo entro en un estado de alerta frente a lo que se avecinaba, múltiples enfermedades, muertes y contagios venían cargados para arrasar sin piedad con todo. La pandemia actual está afectando seriamente a la desigualdad social, la distribución de recursos en largos plazos, el deterioro potencial de la salud, sobre todo en esta última que ha sido la más afectada. Según la Organización Mundial de la Salud el COVID -19 ha sido una pandemia que ha puesto a prueba los sistemas de salud de países en todo el mundo, las reacciones inmediatas y decisiones basados en salud pública, así como conocer los determinantes sociales han generado nuevos conceptos en la cotidianidad de las sociedades y, además, el reforzamiento del autocuidado como medida de protección colectiva.

Mirándolo desde la tasa de mortalidad especifica por la enfermedad producida por este virus, que por ahora es el indicador más confiable para comparar entre países, analizó que los sistemas de salud de los países de Europa y Norteamérica no salen bien librados en términos de capacidad para atender la crisis. "Bélgica que es uno de los pocos países de Europa donde los pacientes tienen que cubrir un porcentaje importante de sus gastos en salud, es el país con la mayor tasa de mortalidad superando a Italia y España pese a que su población es más joven, en promedio, que la de esos países" [9].

* 1. **Determinación de los objetivos comerciales**
     1. **Compilación de la información de la empresa**

El caso de estudio esta centrado en los datos que ha suministrado el Hospital de Tongji en Wuhan, China, para que se realicen diferentes casos de investigación frente al COVID – 19, quien a su vez ha sido uno de los hospitales que ha trabajado fuertemente evitando que este virus se siga propagando.

* + 1. **Describir el área problemática**

Según los datos suministrados por el Hospital de Tongji desde el 10 de enero se han tomado todos los casos de estudios pertinentes en Wuhan, China que hasta la actualidad va con un total de casos de 50.340 confirmados, personas recuperadas en un margen de 46.471 y de muertos 3869. La Organización Mundial de la Salud está colaborando estrechamente con expertos mundiales, gobiernos y asociados para ampliar rápidamente los conocimientos científicos sobre este nuevo virus, rastrear su propagación y virulencia y asesorar a los países sobre las medidas para proteger la salud y prevenir la propagación del brote.

* + 1. **Definición de los objetivos del negocio**

Para el desarrollo de este proyecto estableceremos los siguientes objetivos de negocio:

* Determinar cuales son los biomarcadores que más causan impacto en la infección del COVID – 19.
* Identificar cual es el margen de edad que más se encuentra vulnerable frente a la gravedad del virus.
* Establecer cuál es el género que más se ve afectado.
  + 1. **Valoración de la situación**

Para este proyecto se obtuvo la información brindada por el Hospital de Tongji ubicado en Wuhan, China.

* **Datos:** La información medica recopilada es de todos los pacientes que ingresaron entre el 10 de enero y el 18 de febrero del 2020, con un total de 375 pacientes. Básicamente, han sido dos conjuntos de datos en el primero se presentan 74 variables y en el segundo recopila solo tres biomarcadores que son importantes para el desarrollo del proyecto.

Los datos se encuentran disponible en la información suplementaria del siguiente link: https://www.nature.com/articles/s42256-020-0180-7#MOESM3

* **Factores de riesgo:** El proyecto puede tener falencias por el tiempo de desarrollo, puesto que no hay posibilidad de volver a realizar todo de nuevo una vez realizada la evaluación, entonces debe establecerse los pasos de la mejor manera posible para tener éxito.
  + 1. **Inventario de recursos**
       1. **Recursos del hardware**

Para este proyecto se cuenta con un equipo de cómputo con pocas características, puesto que se realizará un tratamiento de datos con baja necesidad de requerimientos, sin embargo, posee una capacidad de almacenamiento media ya que los datos tienen un volumen un tanto medio/bajo.

* + - 1. **Recurso del software**

El equipo de cómputo tiene un sistema operativo Windows 10 de 64 bits, a su vez para el procesamiento de los datos y el análisis exploratorio se cuenta con software de open source como lo son RStudio, Jupyter Notebook.

* + - 1. **Recurso humano**

Para este proyecto, solo se tiene a disposición una persona con conocimientos afines a las tecnologías de información y un rendimiento del estudio de la minería de datos.

* + 1. **Requerimientos**
* Se requiere realizar un análisis exploratorio sobre los biomarcadores suministrados en los datos de los pacientes.
* Visualizar la información de manera óptima para que sea entendible a la vista de personas que no tienen conocimiento de minería de datos.
* Implementar la conexión con un framework para que sea posible predecir el comportamiento predictivo.
  + 1. **Supuestos**
* Las herramientas para el análisis y la aplicación de los métodos deben funcionar en su totalidad.
* Los datos son brindados por una fuente confiable, por lo tanto se supone que la información allí suministrada es veraz.
* Los datos posean una variable categórica que permitirá la predicción.
* Todos los atributos responden a las preguntas que se han planteado.
  + 1. **Restricciones**
* Las etapas de la metodología deben cumplirse a cabalidad.
* Los datos faltantes deben llenarse con la media o moda.
* La evaluación se realizará una vez terminada todas las fases anteriores.
* Las herramientas de exploración de datos deben suministrar la información veraz.
* Los datos deben ser completamente procesados.
  1. **Orígenes de datos y almacenes de conocimientos**

Los datos a estudiar fueron encontrados en un trabajo de investigación complementario en donde realizaron un mismo enfoque, estos se encuentran disponible en el siguiente link <https://www.nature.com/articles/s42256-020-0180-7#MOESM3> específicamente en la información complementaria, cabe resaltar que son de libre acceso y que cuenta con 6121 registros.

* 1. **Terminología**
* **Biomarcadores:** Molécula biológica que se encuentra en la sangre, otros líquidos o tejidos del cuerpo y es un signo de un proceso normal o anormal, o de una afección o enfermedad. Un marcador biológico se utiliza a veces para determinar la respuesta del cuerpo a un tratamiento para una enfermedad o afección. También se llama biomarcador, marcador molecular y molécula distintiva.
* **Dataset:** Es una representación de datos residente en memoria que proporciona un modelo de programación relacional coherente independientemente del origen de datos que contiene. El data set contiene en sí, un conjunto de datos que han sido volcados desde el proveedor de datos.
* **COVID – 19:** es la enfermedad infecciosa causada por el coronavirus que se ha descubierto más recientemente.
  1. **Determinación de los objetivos de minería de datos**

Este proyecto tiene como finalidad aplicar metodologías y técnicas de minería de datos como lo son modelos predictivos y descriptivos, para lo cual se desarrollaron los siguientes objetivos.

* Construir un conjunto de datos que sea viable para la aplicación de técnicas de minería de datos
* Realizar un análisis exploratorio de los datos mediante técnicas de visualización de datos.
* Desarrollar un modelo predictivo para la gravedad de los pacientes infectados con COVID – 19 en Wuhan China.
* Desarrollar la conexión entre el modelo y un framework para facilitar el acceso al conocimiento para todas las personas.

1. **FASE 2: COMPRENSIÓN DE LOS DATOS**

El desarrollo de la segunda fase de la metodología CRISP – DM va enfocado a la búsqueda de información a través de los datos, para introducirnos un poco más al problema real en sí, desde los datos presentados y los informes estadísticos que podamos obtener de este, a su vez, nos va a permitir conocer la veracidad de los datos y la calidad que estos posean.

* 1. **Recopilación de datos iniciales**
     1. **Datos existentes**

Los datos obtenidos tratan sobre los pacientes infectados con COVID – 19 en Wuhan, China desde el 10 de enero hasta el 18 de febrero del 2020, está conformado por 81 variables de la cual 74 son biomarcadores y 7 conforman la edad, el sexo, clase, fecha de entrada, etc.

Tabla . Datos existentes

|  |  |  |
| --- | --- | --- |
| **ATRIBUTOS** | **TIPO DE DATO** | **DESCRIPCIÓN** |
| PATIENT\_ID | Int64 | Id del paciente |
| RE\_DATE | Datetime64 | Fecha de registro de datos |
| Age | Int64 | Edad |
| Gender | Int64 | Genero |
| Admission time | Datetime64 | Tiempo de admisión |
| Discharge time | Dattime64 | Tiempo de salida |
| Outcome | Int64 | Variable clase o predictora |
| Hypersensitive cardiac troponinI | Float64 | Troponina cardiaca hipersensible |
| Hemoglobin | Float64 | Hemoglobina |
| Serum chloride | Float64 | Cloruro de sodio |
| Prothrombin time | Float64 | Tiempo de protrombina |
| Procalcitonin | Float64 | Procalcitonin |
| Eosinophils (%) | Float64 | Eosinófilos (%) |
| Interleukin 2 receptor | Float64 | Receptor de interleucina 2 |
| Alkaline phosphatase | Float64 | Fosfatasa alcalina |
| Albumin | Float64 | Albúmina |
| Basophil (%) | Float64 | Basófilo (%) |
| Interleukin 10 | Float64 | Interleucina 10 |
| Total bilirubin | Float64 | Bilirrubina total |
| Platelet count | Float64 | Recuento de plaquetas |
| monocytes(%) | Float64 | monocitos (%) |
| antithrombin | Float64 | antitrombina |
| Interleukin 8 | Float64 | Interleucina 8 |
| indirect bilirubin | Float64 | bilirrubina indirecta |
| Red blood cell distribution width | Float64 | Ancho de la distribución de los glóbulos rojos |
| neutrophils(%) | Float64 | neutrófilos (%) |
| total protein | Float64 | proteina total |
| Quantification of Treponema pallidum antibodies | Float64 | Cuantificación de anticuerpos contra Treponema pallidum |
| Prothrombin activity | Float64 | Actividad de protrombina |
| HBsAg | Float64 | virus de la hepatitis B |
| mean corpuscular volume | Float64 | volumen corpuscular medio |
| hematocrit | Float64 | hematocrito |
| White blood cell count | Float64 | recuento de glóbulos blancos |
| Tumor necrosis factorα | Float64 | Factor de necrosis tumoral α |
| mean corpuscular hemoglobin concentration | Float64 | Concentración de hemoglobina corpuscular media |
| fibrinogen | Float64 | fibrinógeno |
| Interleukin 1β | Float64 | Interleucina 1β |
| Urea | Float64 | Urea |
| lymphocyte count | Float64 | recuento de linfocitos |
| PH value | Float64 | Valor de PH |
| Red blood cell count | Float64 | Recuento de glóbulos rojos |
| Eosinophil count | Float64 | Recuento de eosinófilos |
| Corrected calcium | Float64 | Calcio corregido |
| Serum potassium | Float64 | Potasio sérico |
| glucose | Float64 | glucosa |
| neutrophils count | Float64 | recuento de neutrófilos |
| Direct bilirubin | Float64 | Bilirrubina directa |
| Mean platelet volume | Float64 | Volumen medio de plaquetas |
| ferritin | Float64 | ferritina |
| RBC distribution width SD | Float64 | Ancho de distribución de RBC SD |
| Thrombin time | Float64 | Tiempo de trombina |
| (%) lymphocyte | Float64 | (%) linfocito |
| HCV antibody quantification | Float64 | Cuantificación de anticuerpos del VHC |
| D-D dimer | Float64 | Dímero D-D |
| Total cholesterol | Float64 | Colesterol total |
| aspartate aminotransferase | Float64 | aspartato aminotransferasa |
| Uric acid | Float64 | Acido úrico |
| HCO3- | Float64 | Bicarbonato |
| calcium | Float64 | Calcio |
| Amino-terminal brain natriuretic peptide precursor (NT- | Float64 | Precursor del péptido natriurético cerebral amino-terminal (NT- |
| Lactate dehydrogenase | Float64 | Lactato deshidrogenasa |
| platelet large cell ratio | Float64 | Proporción de plaquetas de células grandes |
| Interleukin 6 | Float64 | Interleucina 6 |
| Fibrin degradation products | Float64 | Productos de degradación de fibrina |
| monocytes count | Float64 | recuento de monocitos |
| PLT distribution width | Float64 | Ancho de distribución PLT |
| globulin | Float64 | globulina |
| γ-glutamyl transpeptidase | Float64 | γ-glutamil transpeptidasa |
| International standard ratio | Float64 | Razón estándar internacional |
| basophil count(#) | Float64 | recuento de basófilos (#) |
| 2019-nCoV nucleic acid detection | Float64 | Detección de ácido nucleico 2019-nCoV |
| mean corpuscular hemoglobin | Float64 | Hemoglobina corpuscular media |
| Activation of partial thromboplastin time | Float64 | Activación del tiempo de tromboplastina parcial |
| Hypersensitive c-reactive protein | Float64 | Proteína C reactiva hipersensible |
| HIV antibody quantification | Float64 | Cuantificación de anticuerpos del VIH |
| serum sodium | Float64 | sodio sérico |
| thrombocytocrit | Float64 | trombocitocrito |
| ESR | Float64 |  |
| glutamic-pyruvic transaminase | Float64 | transaminasa glutámico-pirúvica |
| eGFR | Float64 | Receptor de tirosina cinasa |
| creatinine | Float64 | creatinina |

Tabla . Datos de los 110 pacientes

|  |  |  |
| --- | --- | --- |
| **ATRIBUTO** | **TIPO DE DATO** | **DESCRIPCIÓN** |
| PATIENT\_ID | Int64 | Identificación de paciente |
| RE\_DATE | Datetime64 | Fecha de registro |
| Lactate dehydrogenase | Float64 | Lactato deshidrogenasa |
| Hypersensitive c-reactive protein | Float64 | Proteína C reactiva hipersensible |
| (%)lymphocyte | Float64 | (%) Linfocitos |
| Admission time | Datetime64 | Fecha de admisión |
| Discharge time | Datetime64 | Fecha de salida |
| outcome | Int64 | Variable clase |

* + 1. **¿Existen datos suficientes para obtener conclusiones generales o realizar predicciones precisas?**

El conjunto de datos con el que se va a trabajar cuenta con un total de 6121 registros con los datos de 375 pacientes los suficientes para realizar los análisis necesarios para la realización del proyecto.

* + 1. **¿Dispone de atributos suficientes para su método de modelado?**

Se disponen de 74 biomarcadores y 7 atributos entre ellos incluidos la variable predictora lo que nos deja entrevisto que la información es más que suficiente.

* + 1. **¿Ha considerado cómo se gestionan los valores perdidos en cada origen de datos?**

Para este proyecto es necesario entender que existe campos faltantes en todos los datos de los pacientes porque son biomarcadores respectivos y no todos cumplen con esos biomarcadores o simplemente no tomaron muestra de ellos.

* 1. **Resúmenes estadísticos de los datos numéricos**

En la tabla 2, se presentan un reporte estadístico por todas las variables numéricas que se encuentran en el Dataset que son aproximadamente 77 variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ATRIBUTOS** | **MEDIA** | **MEDIANA** | **VARIANZA** | **DESVIACIÓN** |
| Age | 59 | 62 | 268 | 16.373 |
| gender | 1.39 | 1 | 0.238 | 0.487 |
| outcome | 0.47 | 0 | 0.249 | 0.499 |
| Hypersensitive cardiac troponinI | 1223.22 | 20.6 | 29012149.39 | 5386.29 |
| Hemoglobin | 123.12 | 125 | 562.91 | 23.725 |
| Serum chloride | 103.13 | 102.1 | 59.95 | 7.74 |
| Prothrombin time | 16.674 | 14.8 | 90.13 | 9.49 |
| Procalcitonin | 1.106 | 0.1 | 21.70 | 4.65 |
| Eosinophils (%) | 0.62 | 0.1 | 1.136 | 1.066 |
| Interleukin 2 receptor | 907.16 | 676.5 | 618641.70 | 786.53 |
| Alkaline phosphatase | 82.46 | 69.5 | 2166.02 | 46.540 |
| Albumin | 32.00 | 32.2 | 39.01 | 6.246 |
| Basophil (%) | 0.209 | 0.2 | 0.047 | 0.217 |
| Interleukin 10 | 16.06 | 5.9 | 4666.64 | 68.312 |
| Total bilirubin | 16.69 | 10.7 | 704.990 | 26.55 |
| Platelet count | 184.310 | 178 | 10905.012 | 104.42 |
| monocytes(%) | 6.155 | 5.7 | 16.59 | 4.073 |
| antithrombin | 85.31 | 86 | 333.44 | 18.260 |
| Interleukin 8 | 83.08 | 16 | 304835.98 | 552.11 |
| indirect bilirubin | 6.88 | 5.4 | 49.27 | 7.019 |
| Red blood cell distribution width | 13.07 | 12.6 | 2.979 | 1.726 |
| neutrophils(%) | 77.601 | 82.4 | 276.72 | 16.634 |
| total protein | 65.303 | 65.9 | 58.085 | 7.6213 |
| Quantification of Treponema pallidum antibodies | 0.132 | 0.05 | 0.590 | 0.768 |
| Prothrombin activity | 78.55 | 81 | 489.585 | 22.12 |
| HBsAg | 8.30 | 0.01 | 1836.36 | 42.852 |
| mean corpuscular volume | 90.38 | 90.1 | 42.25 | 6.500 |
| hematocrit | 36.544 | 36.6 | 28.216 | 5.311 |
| White blood cell count | 15.59 | 7.72 | 4585.57 | 67.716 |
| Tumor necrosis factorα | 11.577 | 8.66 | 169.96 | 13.03 |
| mean corpuscular hemoglobin concentration | 342.803 | 343 | 316.49 | 17.79 |
| fibrinogen | 4.29 | 4.12 | 3.561 | 1.887 |
| Interleukin 1β | 6.509 | 5 | 48.98 | 6.999 |
| Urea | 9.58 | 5.984 | 89.026 | 9.43 |
| lymphocyte count | 1.0167 | 0.8 | 4.23 | 2.05 |
| PH value | 6.48 | 6.5 | 0.526 | 0.725 |
| Red blood cell count | 9.28 | 4.14 | 1316.913 | 36.28 |
| Eosinophil count | 0.038 | 0.01 | 0 | 0.06 |
| Corrected calcium | 2.354 | 2.36 | 0.016 | 0.128 |
| Serum potassium | 4.508 | 4.41 | 0.72 | 0.851 |
| glucose | 8.889 | 6.99 | 27.22 | 5.21 |
| neutrophils count | 7.809 | 5.85 | 36.810 | 6.067 |
| Direct bilirubin | 9.886 | 4.8 | 457.813 | 21.39 |
| Mean platelet volume | 10.909 | 10.8 | 1.18 | 1.087 |
| ferritin | 1379.14 | 711.0 | 11890710.5 | 3448.29 |
| RBC distribution width SD | 42.44 | 40.9 | 39.65 | 6.29 |
| Thrombin time | 18.17 | 16.8 | 78.966 | 8.88 |
| (%) lymphocyte | 15.39 | 11.45 | 167.107 | 12.926 |
| HCV antibody quantification | 0.117 | 0.06 | 0.05 | 0.234 |
| D-D dimer | 7.943 | 2.15 | 84.70 | 9.20 |
| Total cholesterol | 3.68 | 3.63 | 0.96 | 0.98 |
| aspartate aminotransferase | 46.528 | 27 | 10655.5 | 103.22 |
| Uric acid | 276.12 | 243.7 | 22939.40 | 151.45 |
| HCO3- | 23.14 | 23.5 | 19.42 | 4.40 |
| calcium | 2.07 | 2.08 | 0.026 | 0.162 |
| Amino-terminal brain natriuretic peptide precursor(NT-proBNP) | 3669.36 | 585 | 93795150.2 | 9684.78 |
| Lactate dehydrogenase | 474.22 | 340 | 136116.21 | 368.93 |
| platelet large cell ratio | 31.765 | 30.9 | 73.277 | 8.56 |
| Interleukin 6 | 112.30 | 19.265 | 287529.12 | 536.21 |
| Fibrin degradation products | 61.354 | 17.9 | 4296.24 | 65.54 |
| monocytes count | 0.525 | 0.41 | 2.867 | 1.693 |
| PLT distribution width | 13 | 12.4 | 7.78 | 2.78 |
| globulin | 33.24 | 32.7 | 30.31 | 5.50 |
| γ-glutamyl transpeptidase | 55.34 | 34 | 4818.33 | 69.414 |
| International standard ratio | 1.31 | 1.14 | 0.65 | 0.80 |
| basophil count(#) | 0.0170 | 0.0170 | 0.0002 | 0.017 |
| 2019-nCoV nucleic acid detection | -1.0 | -1.0 | 0.0 | 0.0 |
| mean corpuscular hemoglobin | 30.997 | 30.9 | 8.45 | 2.907 |
| Activation of partial thromboplastin time | 41.520 | 39.2 | 138.600 | 11.77 |
| Hypersensitive c-reactive protein | 76.24 | 51.5 | 6565.11 | 81.025 |
| HIV antibody quantification | 0.100 | 0.09 | 0.001 | 0.04 |
| serum sodium | 141.5 | 140.4 | 53.72 | 7.32 |
| thrombocytocrit | 0.212 | 0.21 | 0.0086 | 0.09 |
| ESR | 33.689 | 28 | 603.74 | 24.571 |
| glutamic-pyruvic transaminase | 38.86 | 24 | 7009.31 | 83.721 |
| eGFR | 81.56 | 87.9 | 1037.19 | 32.20 |
| creatinine | 109.93 | 76 | 17988.03 | 134.11 |

En la tabla 3, se encuentra un resumen estadístico para las variables numéricas de los otros 110 pacientes que se encuentran en el otro Dataset

Tabla . Resumen estadistico datos 110

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ATRIBUTO** | **MEDIA** | **MEDIANA** | **VARIANZA** | **DESVIACIÓN** |
| Lactate dehydrogenase | 376.65 | 258 | 86475.49 | 294.0671 |
| Hypersensitive c-reactive protein | 46.084 | 9.1 | 5208.059 | 72.166 |
| (%)lymphocyte | 18.147 | 17.8 | 167.874 | 12.956 |
| outcome | 0.359 | 0 | 0.230 | 0.479 |

* 1. **Exploración de los datos**

El análisis exploratorio es una tarea fundamental porque nos permite observar el comportamiento de nuestros datos, identificar características; es la forma de analizar a detalle cómo están distribuidos nuestros datos, para este proyecto utilizaremos herramientas de visualización como gráficos, estableciéndonos algunas preguntas que son realmente importantes.

**¿Cómo es la distribución de la variable clase?**

En nuestro Dataset la variable “outcome” es considerada esa variable predictora que nos permite saber cómo están distribuidos nuestros datos.

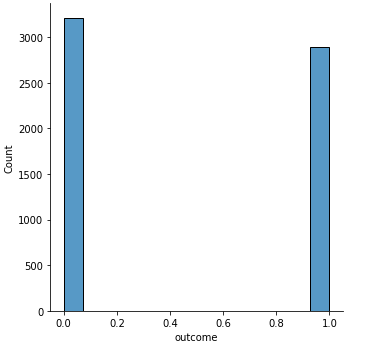


Figura . Distribución de la variable clase

La figura 2, es la representación de los datos dentro del Dataset, cuantos se encuentran identificados como pacientes graves y cuantos no.

**¿Qué tanto influye la variable “Lactate dehydrogenase” en los pacientes según la variable “outcome” y la edad?**

Realizando una investigación previa sobre las variables, se encontró que la “Lactate dehydrogenase” es una prueba que mide el nivel de lactato deshidrogenasa (LDH), que se encuentra presente en la sangre o a veces en otros líquidos del cuerpo, la LDH cumple una función importante para la producción de energía en el cuerpo. Se encuentra en casi todos los tejidos del cuerpo entre ellos, los de la sangre, el corazón, los riñones, el cerebro y los pulmones. Cuando se producen daños en algunos de estos tejidos, liberan LDH en la sangre, lo que significa que si los niveles de LDH están altos en la sangre significa que ciertos tejidos del cuerpo han sufrido daño por una enfermedad, lo cual puede ser sumamente peligroso. [10]

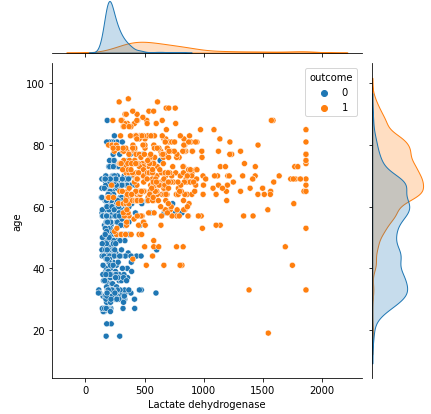


Figura . Relación entre LDH, edad y variable clase

La figura 3, es la mejor muestra para observar que los pacientes que se encuentran más propensos a tener daños en sus tejidos por el nivel de LDH en su sangre, están en un promedio de 60 a 90 años sin omitir también que se observa que son pacientes con más índice de gravedad por estar contagiados.

**¿Cómo se encuentra distribuido la variable “D-D Dimer” en los pacientes según su edad y variable clase “outcome”?**

En las series principales de pacientes con COVID – 19 publicadas hasta la fecha, los pacientes con peor pronóstico han mostrado niveles de D-dímero mucho más altos que los pacientes con enfermedad menos grave. En una serie de pacientes un valor de D-dímero en el momento del ingreso mayor de 1,0 mg/L fue uno de los principales factores de mal pronóstico [11]. Por esta razón el valor D-dímero está asociada a la gravedad de pacientes con COVID – 19.

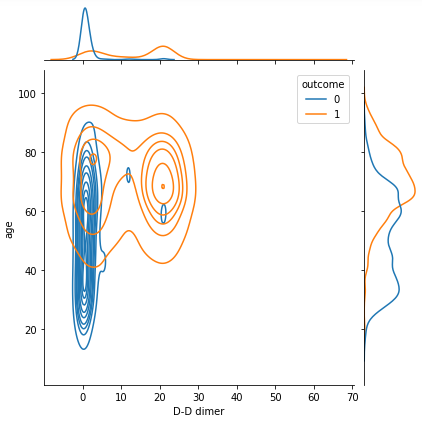


Figura . Relación entre la edad, D-D dimer y la variable clase

La figura 4, nos permite observar con claridad que los pacientes que se encuentran infectados con COVID – 19 entre la edad de 40 y 70 años presentan un nivel de D-dímero mayor de 10 mg/L en la sangre lo que da una mayor probabilidad de su gravedad por tener la enfermedad.

**¿Cómo se encuentra el “Prothrombin time” (tiempo de protombina) en los pacientes según su género y la edad?**

El tiempo de protombina es una de las varias pruebas que evalúan si su sangre se está coagulando correctamente. La coagulación de la sangre es necesaria para ayudar a detener un sangrado. La prueba de tiempo de protrombina puede determinar la falta de determinados coagulantes si la sangre no se coagula normalmente. La prueba de tiempo de protrombina se puede usar para ayudar a diagnosticar trastornos hereditarios y otras afecciones que pueden afectar la coagulación de la sangre [12]. La mayoría de los pacientes graves con coronavirus presentan síndrome de dificultad respiratoria, disfunción de la coagulación, insuficiencia multiorgánica, etc. El tiempo de protombina permite detectar que tan grave es la disfunción de la coagulación en la sangre.

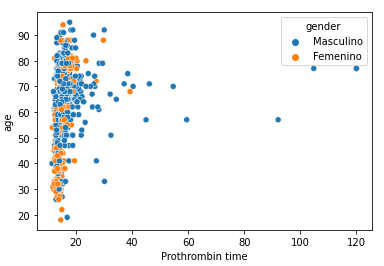


Figura . Relación entre tiempo de protombina, edad y género

La figura 5, nos representa la relación entre el tiempo de protombina, edad y género, en donde es muy fácil observar que los géneros se encuentran bastante distribuidos con la enfermedad pero los pacientes masculinos se presentan más entre 50 y 90 años, así mismo son los que tienden a tener un tiempo de protombina más alto por lo que su difusión de coagulación está más presente y los hace más propenso a aumentar la gravedad de la enfermedad.

* 1. **Verificación de la calidad de los datos**

Realizar la verificación de la calidad de datos trata de realizar una revisión de estos para observar valores perdidos o los que tiene valores faltantes ya sea por error de almacenamiento o codificación. En este proyecto hay que tener en cuenta que los valores faltantes no significan que el Dataset este mal construido puesto que como las variables son biomarcadores de exámenes de sangre, estos varían entre los pacientes y a algunos puede que se le hayan aplicados dichos exámenes como no. Por lo anterior, la siguiente tabla presenta el total de valores faltantes en los atributos.

Tabla . Valores faltantes

|  |  |
| --- | --- |
| **ATRIBUTOS** | **VALORES FALTANTES** |
| Age | 0 |
| Gender | 0 |
| Admission time | 0 |
| Discharge time | 0 |
| Outcome | 0 |
| Hypersensitive cardiac troponinI | 5599 |
| Hemoglobin | 5131 |
| Serum chloride | 5131 |
| Prothrombin time | 5444 |
| Procalcitonin | 5647 |
| Eosinophils (%) | 5149 |
| Interleukin 2 receptor | 5838 |
| Alkaline phosphatase | 5176 |
| Albumin | 5172 |
| Basophil (%) | 5149 |
| Interleukin 10 | 5839 |
| Total bilirubin | 5176 |
| Platelet count | 5149 |
| monocytes(%) | 5148 |
| antithrombin | 5776 |
| Interleukin 8 | 5838 |
| indirect bilirubin | 5200 |
| Red blood cell distribution width | 5183 |
| neutrophils(%) | 5149 |
| total protein | 5200 |
| Quantification of Treponema pallidum antibodies | 5827 |
| Prothrombin activity | 5447 |
| HBsAg | 5827 |
| mean corpuscular volume | 5149 |
| hematocrit | 5149 |
| White blood cell count | 4979 |
| Tumor necrosis factorα | 5838 |
| mean corpuscular hemoglobin concentration | 5149 |
| fibrinogen | 5540 |
| Interleukin 1β | 5838 |
| Urea | 5170 |
| lymphocyte count | 5149 |
| PH value | 5722 |
| Red blood cell count | 5149 |
| Eosinophil count | 5192 |
| Corrected calcium | 5126 |
| Serum potassium | 5331 |
| glucose | 5176 |
| neutrophils count | 5244 |
| Direct bilirubin | 5176 |
| Mean platelet volume | 5244 |
| ferritin | 5823 |
| RBC distribution width SD | 5183 |
| Thrombin time | 5540 |
| (%) lymphocyte | 5148 |
| HCV antibody quantification | 5827 |
| D-D dimer | 5476 |
| Total cholesterol | 5175 |
| aspartate aminotransferase | 5171 |
| Uric acid | 5172 |
| HCO3- | 5172 |
| calcium | 5127 |
| Amino-terminal brain natriuretic peptide precursor (NT- | 5631 |
| Lactate dehydrogenase | 5172 |
| platelet large cell ratio | 5244 |
| Interleukin 6 | 5834 |
| Fibrin degradation products | 5776 |
| monocytes count | 5149 |
| PLT distribution width | 5244 |
| globulin | 5176 |
| γ-glutamyl transpeptidase | 5176 |
| International standard ratio | 5447 |
| basophil count(#) | 5149 |
| 2019-nCoV nucleic acid detection | 5605 |
| mean corpuscular hemoglobin | 5149 |
| Activation of partial thromboplastin time | 5538 |
| Hypersensitive c-reactive protein | 5369 |
| HIV antibody quantification | 5828 |
| serum sodium | 5131 |
| thrombocytocrit | 5244 |
| ESR | 5723 |
| glutamic-pyruvic transaminase | 5175 |
| eGFR | 5170 |
| creatinine | 5170 |

Si analizamos la Tabla 2, los valores faltantes de los atributos varían en un valor no máximo 5600 puesto que para los 375 pacientes se realizaron varios exámenes en un mismo día pero con diferentes horas lo que permite que para el registro de los datos, los biomarcadores que no fueron realizados en la misma hora se registran como valores nulos, por lo tanto no hay que realizar tratamiento para dichos datos.

|  |  |
| --- | --- |
| **ATRIBUTO** | **VALORES FALTANTES** |
| Lactate dehydrogenase | 397 |
| Hypersensitive c-reactive protein | 472 |
| (%)lymphocyte | 384 |
| outcome | 0 |

* 1. **MATRIZ DE COVARIANZA**

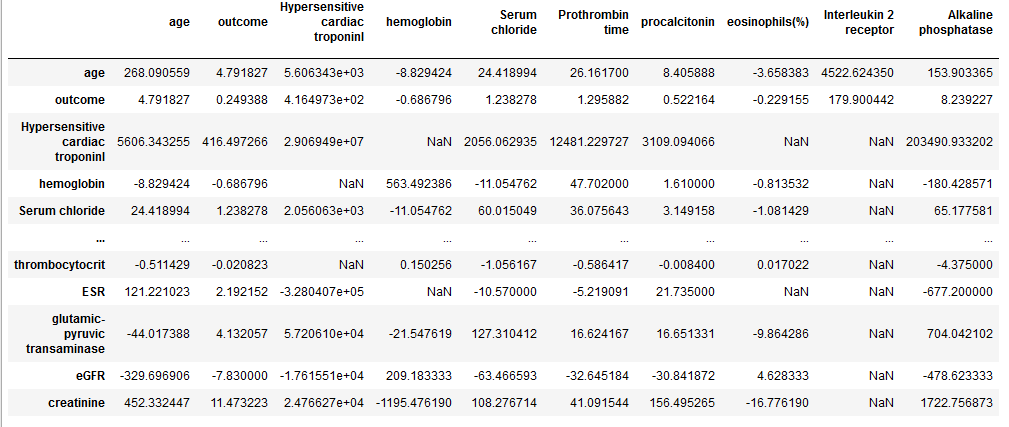


Figura . Matriz de covarianza

La figura 6, nos muestra la matriz de covarianza de los datos en donde es posible observar que entre mayor sea el número entre ellos más dependencia tienen, en este caso es posible observar que la Prothombin time y la Hypersensitive cardiac troponinI poseen mucha dependencia entre ellas.

1. **FASE 3: PREPARACIÓN DE LOS DATOS**

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

* 1. **Seleccionar los datos**

.

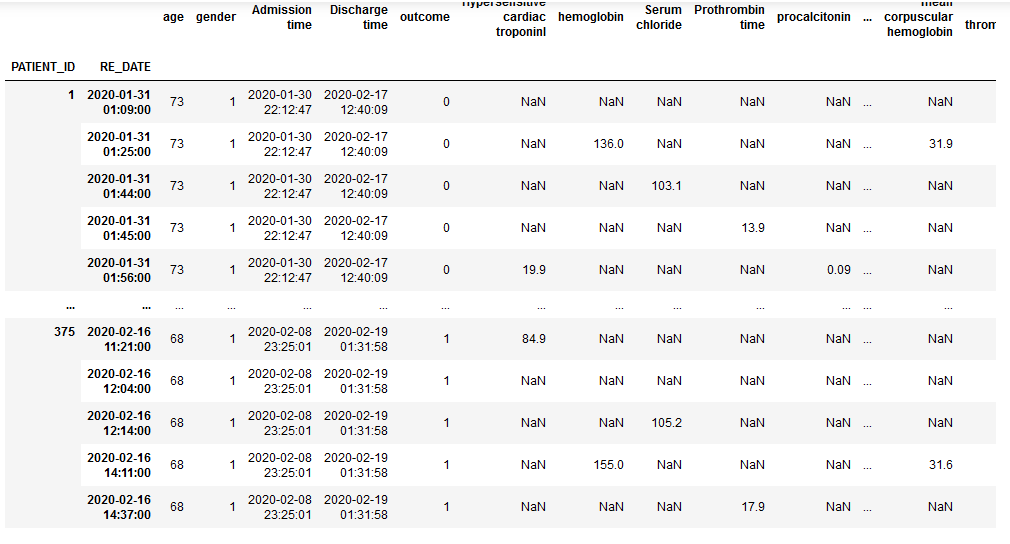


Figura . Datos de 375 pacientes con COVID - 19

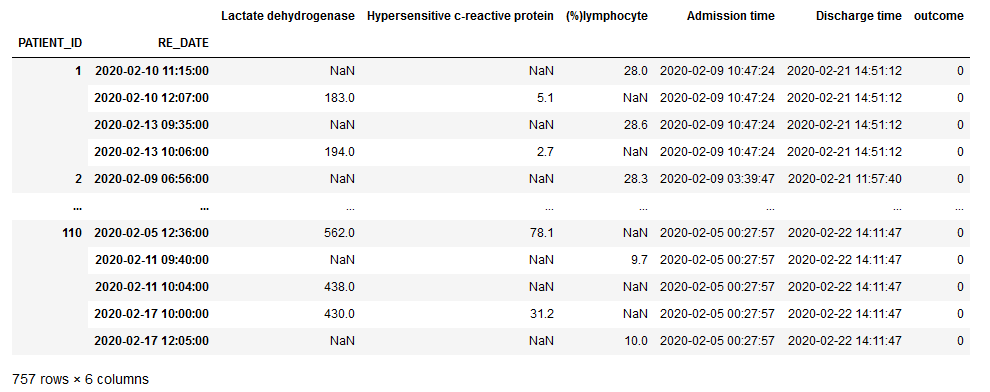


Figura . Datos de 110 pacientes con COVID - 19

Realizando un análisis a nuestro Dataset se llego a la conclusión de que todos los datos son importantes para la realización del modelo, por tal razón no se realizara la eliminación de ninguno de sus atributos ni de sus registros.

* 1. **Limpiar los datos**

Para limpiar los datos fue necesario realizar una función que me normalizara los datos por “Discharge time” y “RE\_DATE”, es decir que me uniera en un solo registro por fecha de admisión y fecha de registro para disminuir los datos

* 1. **Integrar los datos**

Para seleccionar los datos fue importante primero agrupar por identificación de pacientes (PATIENT\_ID) todos los datos que se encuentren tanto en los archivos de 370 pacientes y 110 pacientes, esto con el fin de que me quede un solo archivo y pueda realizar los siguientes procedimientos

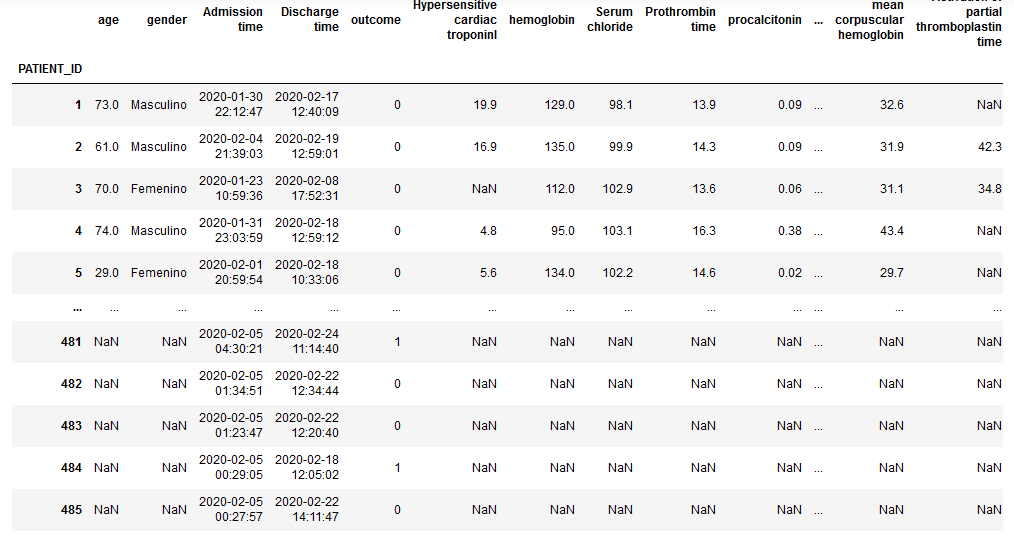


Figura . Datos limpios y procesados

La figura 8, me muestra el Dataset con los datos concatenados, procesados por registro, fecha de ingreso e identificación de paciente, ahora es posible realizar los modelos pertinentes, si nos damos cuenta el número de registro fue disminuido por lo que quiere decir que la integración fue realizada correctamente.

**Graficas generadas después de realizar el preprocesamiento de los datos**

* **Distribución de las clases**

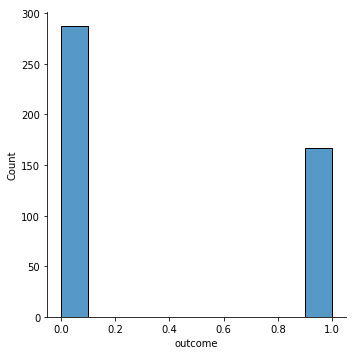


Figura . Distribución nueva de clases

En la figura 9, observamos nuevamente la distribución de los datos y nos damos cuenta como se realizó la variación pertinente una vez los datos fueron concatenados y pre procesados.

* **¿Qué tanto influye la variable “Lactate dehydrogenase” en los pacientes según la variable “outcome” y la edad?**

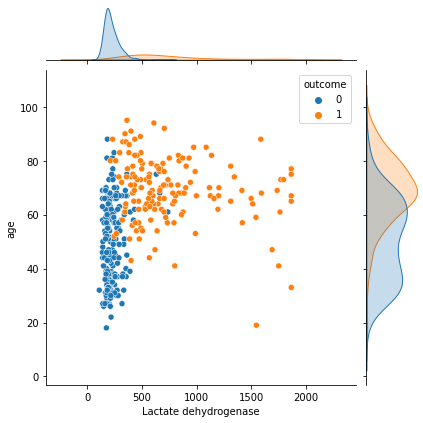


Figura . Relación entre LDH, edad y clase

Se puede observar notoriamente la diferencia existente entre la figura 10 y la figura 3, los datos al ser concatenados y limpiados su representación van a ser mucho mejor y sigue mostrando el análisis realizado respectivamente en la figura 3.

* **¿Cómo se encuentra distribuido la variable “D-D Dimer” en los pacientes según su edad y variable clase “outcome”?**

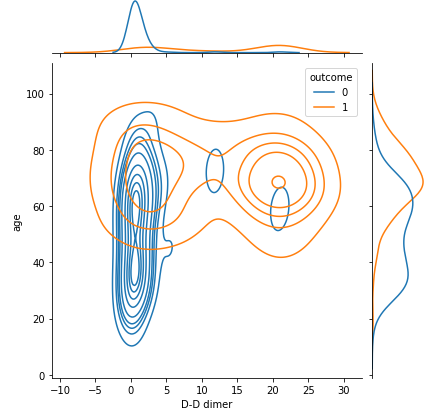


Figura . Relación entre la edad, D-dímero y outcome

La figura 11 muestra la representación de la figura 4 con los datos ya pre procesados, la diferencia es notoria, la reducción de los datos se puede observar y el análisis es mucho más fácil de entender.

* **¿Cómo se encuentra el “Prothrombin time” (tiempo de protombina) en los pacientes según su género y la variable clase?**

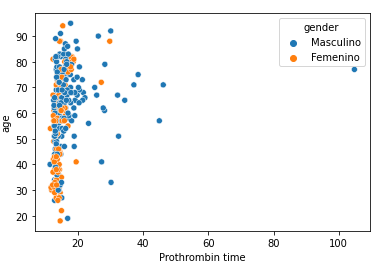


Figura . Relación entre gender, protombina y edad

La figura 12 se compara con la figura 5 y es posible ver diferencias mínimas entre la distribución de sus datos, por lo que es fácil entender que si se realizó un buen preprocesamiento.

# Referencias

|  |  |
| --- | --- |
| [1] | Connet Elsevier, «ELSEVIER,» 22 08 2018. [En línea]. Available: https://www.elsevier.com/es-es/connect/ehealth/que-es-el-machine-learning-salud-digital. |
| [2] | K. Darlington, «OpenMind,» 22 05 2020. [En línea]. Available: https://www.bbvaopenmind.com/tecnologia/inteligencia-artificial/esta-ayudando-la-inteligencia-artificial-contener-la-pandemia-covid-19/. |
| [3] | Centro Cochrane Iberoamericano, «Cochrane Iberoamérica,» 08 04 2020. [En línea]. Available: https://es.cochrane.org/es/una-revisi%C3%B3n-sistem%C3%A1tica-destaca-las-limitaciones-de-los-modelos-predictivos-para-el-diagn%C3%B3stico-y. |
| [4] | N. Z. R. Z. M. D. T. X. J. P. E. P. J. H. Y. Z. X. X. H. X. F. Z. G. W. Haochen Yao, «Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests,» 2020. |
| [5] | . Z. G. X.-L. W. G. S. T. J. Z. H. X. C. C. R. W. X. H. T.-n. H. Li Yan, «A machine learning-based model for survival prediction in patients with severe COVID-19 infection,» *Semanthic Scholar,* 2020. |
| [6] | . R. K. A.-M. Z. D. H. T. I. M. E. H. Chowdhury, «An early warning tool for predicting mortality risk of COVID-19 patients using,» *Semanthic Scholar,* p. 1, 2020. |
| [7] | . G. K. V. D. P. A. Karthikeyan, «Machine learning based clinical decision supportsystem for early COVID-19 mortality prediction,» *medrxiv,* p. 21, 2020. |
| [8] | L. Z. H. G. J. Yan, «An interpretable mortality prediction model for COVID-19 patients,» *Nat Mach Intell,* pp. 283 - 288, 2020. |
| [9] | Ministerio de Salud y Protección Social, «El reto del COVID - 19 para los sistemas de salud en el mundo,» 08 06 2020. [En línea]. Available: https://www.minsalud.gov.co/Paginas/El-reto-del-covid-19-para-los-sistemas-de-salud-en-el-mundo.aspx. |
| [10] | MedlinePlus, «MedlinePlus,» 26 02 2020. [En línea]. Available: https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-lactato-deshidrogenasa-ldh/. |
| [11] | Centro Cochrane Iberoamericano , «Cocharne Iberoamérica,» 06 04 2020. [En línea]. Available: https://es.cochrane.org/es/recursos/evidencias-covid-19/el-valor-d-d%C3%ADmero-est%C3%A1-asociado-la-gravedad-de-los-pacientes-con-la. |
| [12] | The StayWell Company, [En línea]. Available: https://mychart.geisinger.org/Staywel/html/Lab%20Tests/167,pt-prothrombin-time-ES.html. |
| [13] | Organización Mundial de la Salud, «Brote de enfermedad por coronavirus (COVID -19),» 2020. [En línea]. Available: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019. [Último acceso: 01 10 2020]. |