ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.874

Prediction System to Control the Spread of COVID-19 in People Who Work in Person in the Los Olivos District

Iván Wilber, Paredes Reyes[1]
1 Universidad Tecnológica del Perú, *Perú, E22100602@postgradoutp.edu.pe*

Abstract– In Peru, the COVID-19 pandemic affects various private and state sectors, due to the spread of the virus. The current government decreed quarantines to reduce infections, which implied that part of the population would have to stay at home , but a group of people carry out face-to-face activities and also cause contagion to younger people or older adults.The objective of this research work is to create a prediction mobile application that allows to identify if a person who carries out economic activities in person is exposed to contagion, depending on the transport chosen and the time to reach their destination.In order to develop the solution, data was collected from pages of the Peruvian government in charge of controlling the spread of the epidemic and when analyzing the data, the variables for prediction were identified through decision trees. Azure machine learning, SQL language, was used for the development of the project. The Kanban methodology was used to develop the project and the CRIPS-DM methodology for data analysis.The result was a mobile application to carry out a test on possible COVID-19 contagion.
Keywords: Propagation, COVID-19, Machine learning, Decision tree, azure machine learning.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Prediction System to Control the Spread of COVID-19 in People Who Work in Person in the Los Olivos District

Iván Wilber, Paredes Reyes[1]
1 Universidad Tecnológica del Perú, *Perú, E22100602@postgradoutp.edu.pe*

Abstract– In Peru, the COVID-19 pandemic affects various private and state sectors, due to the spread of the virus. The current government decreed quarantines to reduce infections, which implied that part of the population would have to stay at home , but a group of people carry out face-to-face activities and also cause contagion to younger people or older adults.The objective of this research work is to create a prediction mobile application that allows to identify if a person who carries out economic activities in person is exposed to contagion, depending on the transport chosen and the time to reach their destination.In order to develop the solution, data was collected from pages of the Peruvian government in charge of controlling the spread of the epidemic and when analyzing the data, the variables for prediction were identified through decision trees. Azure machine learning, SQL language, was used for the development of the project. The Kanban methodology was used to develop the project and the CRIPS-DM methodology for data analysis.The result was a mobile application to carry out a test on possible COVID-19 contagion.
Keywords: Propagation, COVID-19, Machine learning, Decision tree, azure machine learning.

Resumen. - En el Perú la pandemia del COVID-19 afecta a diversos sectores privados y estatales, a causa de la propagación del virus. El gobierno de turno decretó cuarentenas para disminuir los contagios, lo cual implicaba que parte de la población tendría que permanecer en su domicilio, pero un grupo de personas realizan actividades presenciales y ocasionan el contagio a otro grupo de menor edad o adultos mayores. El objetivo del presente trabajo de investigación es crear una aplicación móvil de predicción que permita identificar si una persona que realiza actividades económicas presenciales está expuesta al contagio, dependiendo el tipo de transporte que elija y el tiempo en llegar a su destino. Para poder desarrollar la solución se recopiló data de páginas del gobierno peruano encargados de controlar la propagación de la epidemia y al analizarla se identificó las variables para la predicción mediante árboles de decisiones.Para el desarrollo del proyecto se utilizó azure machine learning ,lenguaje SQL y la metodología CRIPS-DM para el análisis de datos.
Palabras clave: Propagación, COVID-19, Machine learning, Árbol de decisión, azure machine learning.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

I. INTRODUCCION

El machine learning (ML) es una técnica de inteligencia artificial (IA) que pretende simular las características humanas del pensamiento a través de experiencias anteriores, para esto se le proporciona a la computadora gran cantidad de datos con el objetivo de entrenar un conjunto de hipótesis. La computadora obtiene datos y después los divide en datos de entrenamiento y de prueba. Después que el algoritmo es entrenado se puede evaluar la precisión del modelo para realizar predicciones [1]. También, ML se clasifica de acuerdo con la forma que entrenan los modelos y la manera que interpretan los datos. Entonces se organizan en tres categorías (aprendizaje supervisado, aprendizaje por refuerzo, aprendizaje no supervisado) y se subclasifican de acuerdo con la función que cumplen al ser aplicadas [2].

Un nuevo coronavirus en China se reportó en diciembre del 2019 con varios casos de neumonía de diagnóstico desconocido que causan síndromes respiratorios agudos que tenían como foco de contagio un mercado de animales vivos. Las autoridades sanitarias chinas decidieron el 7 de enero del 2020 que el nuevo virus es de la familia coronaviridae y lo denominaron SARS-COV-2. La Organización Mundial de la Salud (OMS) declaró este nuevo virus como una pandemia el 11 de marzo del 2020 [3].

En el caso del Perú la emergencia sanitaria inició el 11 de marzo del 2020 por D.S N° 008-2020-SA (2020). Esta cuarentena a pesar de tener larga duración no sirvió para rastrear focos de contagio y permitir mejorar el sector salud. Lo que ocasionó que se incremente la propagación del virus en el país [4] .En el Perú, en el Distrito limeño de Los Olivos, la propagación del COVID-19 ocasionó la disminución de actividades presenciales ocasionando la pérdida de inversiones, disminución de actividades económicas, puestos de trabajo y actividades educativas [5].

II. MATERIALES Y MÉTODOS

A. Dataset

El conjunto de datos fue recopilado de la Plataforma Nacional de Datos Abiertos del gobierno peruano que contiene datos de personas como la edad, sexo, zonas, estado de vacunación, fecha de fallecimiento, estado de contagio. Los datos que se dispone permitieron analizar características de las personas que están ocasionando el aumento de contagios por COVID-19 en el distrito de Los Olivos. La tabla 1, muestra los principales dataset y recursos del modelo E-R [6].

TABLA 1
PRINCIPALES TABLAS DE LA BASE DE DATOS DEL MINSA

TABLA MODELO E-R	DATASET
TB_ETNIA	Tabla General Etnias
TB_ANEMIA	Morbilidad: Anemia
TB_SEG_PAC_COVID19	Seguimiento del paciente COVID19
TB_ATEPAC_COVID19	Atención del paciente COVID19
TB_MORBI_IRAS	Morbilidad: Infecciones respiratorias agudas altas
TB_PERSONA	Tabla General de Personas
TB_SOSPECHOSOS_F00_COVID	Sospechoso de Covid-19
TB_MORBI_EMERGENCIAS	Morbilidad en Emergencia Hospitalaria
TB_MORBI_EDAS	Morbilidad: Enfermedades infecciosas intestinales
TB_SEGUIMIENTO_CLINICO_F300	SISCOVID F300 Seguimiento Clinico
TB_TIPO_ESTADP_SEG_COVID	
TB_PRUEBAS_NOMOLECULARES	SISCOVID F100 Pruebas
TB_TIPO_PRUEBA_COVID	
TB_RESULTADO_PRUEBA_COVID	
TB_EESS	Establecimientos de Salud
TB_UBIGEO	Código equivalentes de UBIGEO del Peru
TB_EESS_COVID19	Disponibilidad de camas para atenciones Covid19
TB_CENTRO_VACUNACION	Centros de Vacunación
TB_PROGRAMACION	Programación de vacunas
TB_VACUNA	Catálogo de Fabricante de Vacunas
TB_VACUNADOS_COVID19	Vacunacion
TB_GRUPO_RIESGO	Tabla de grupo de riesgo
TB_POBLACION	Población Peru

B. Revisión de la Literatura

Árboles de decisión

Se usa al existir diversas variables calculadas para determinar el resultado más acertado. Con este tipo de árbol se puede aplicar en un proceso binario de categorías y subcategorías para representar diversas variables que rodean a un resultado [7].
Tipos de árboles de decisión:

- Árbol de clasificación

Con este tipo de árbol se usa en procesos binarios de categorías y subcategorías para representar diversas variables que rodean a un resultado[7] .

- Árbol de regresión

Se usa cuando se tiene diferentes partes de información para conseguir un resultado exclusivo. El árbol de regresión se fracciona en partes de información y luego se subdividen en otros subgrupos [7].

- Bosque de árboles de decisión

Se crean cuando ya existen árboles de decisión, luego se asocian entre sí para obtener una predicción más exacta. Se usa para evaluar el resultado con base al rumbo que estén siguiendo los árboles de decisión[7].

III. METODOLOGIA

La metodología Kanban se usó para gestionar el proyecto mediante la visualización de flujos. A continuación mencionamos las fases del Kanban que se emplearón:

A. Fase I - Análisis

Los datos disponibles en la plataforma de datos abiertos permite analizar características de las personas que están ocasionando el aumento de contagios por COVID-19 en el Distrito de Los Olivos que tiene por código ubigeo el número 150117 tal y como se muestra en la figura 1.

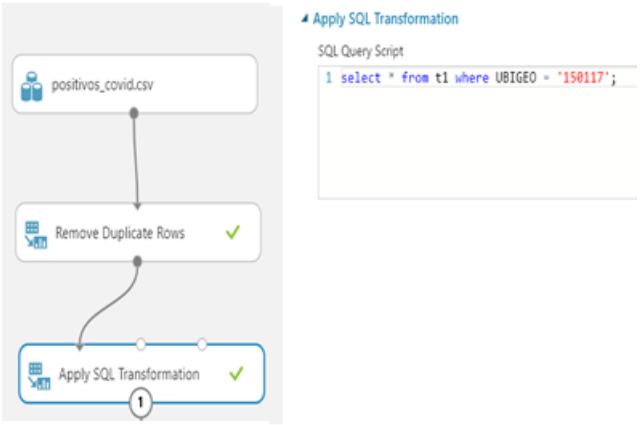

Fig.1 Selección mediante SQL de ubigeo del Distrito de Los Olivos

B. Fase II - Diseño

En está fase se diseñan los prototipos para cada iteración. Se observa en la figura 2 el test sobre el COVID-19 donde el usuario ingresa sus datos personales, medio de transporte y tiempo que le toma llegar a su destino.

Fig.2 Diseño de prototipo de test del COVID-19

C. Fase III - Desarrollo

En esta fase se programa cada iteración. Se uso como base de datos hojas de cálculo googlesheet alojadas en google drive y desarrollado en Appsheet. En la figura 3 se visualiza las tablas data_covid y menú en appsheet que se elaboraron a partir del modelo predictivo.

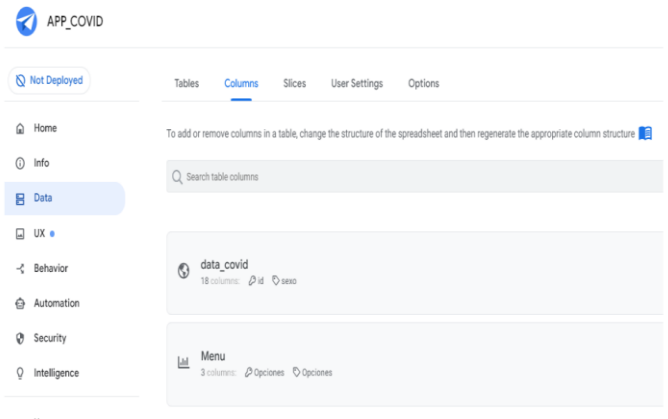

Fig. 3 Descripción tabla data_covid y Menu

D. Fase IV - Pruebas

En está fase se realizan las pruebas de software de los distintos components (base de datos , aplicación móvil , modelo predictivo)

La metodología para lograr el procesamiento de datos es la metodología CRISP-DM y describiremos a continuación sus diferentes fases [8]:

1) Comprensión del negocio
2) Comprensión de los datos
3) Preparación de los datos
4) Modelado
5) Evaluación

Fase 1

El objetivo que se plantea en el trabajo de investigación es desarrollar un modelo predictivo con los datos publicados en la Plataforma Nacional de Datos Abiertos del gobierno peruano, en la cual se encuentra publicada información de los organismos del estado encargados del control de la pandemia del COVID-19.

Fase 2

Luego de comprender los requerimientos de la investigación, se procede con el levantamiento de información. Los datos de casos de COVID-19 son del 08 de marzo del 2020 hasta la actualidad que se encuentran en formato de archivo csv tal y como se visualiza en la figura 4 que contiene información personas: contagiadas, fallecidas, edad, Distritos, sexo, vacunas, etc.

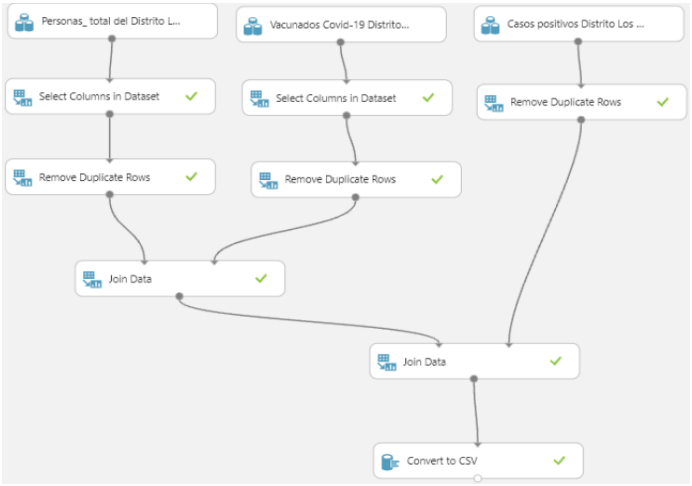

Fig. 4 Dataset obtenido de datos del Minsa

Las variables independientes que se utilizarón para predecir el modelo predictivo son: Tasa de mortalidad, Tasa de ataque, Probabilidad de contagio, medio de transporte, tiempo de desplazamiento, edad, vacunados, sexo.
La variable dependiente que vamos a predecir es contagio que depende de las variables independientes.

Fase 3

Se usó Jupyter Notebook y la plataforma azure machine learning, se importaron los datos para explorarlos y mediante pruebas estadísticas y gráficos de distribución para encontrar relación entre variables independientes y dependientes. En la figura 5 observamos que existe gran cantidad de personas con COVID 19 se encuentra en el rango edad de 30 a 50 años que son las personas que mayormente realizán actividades presencialmente.

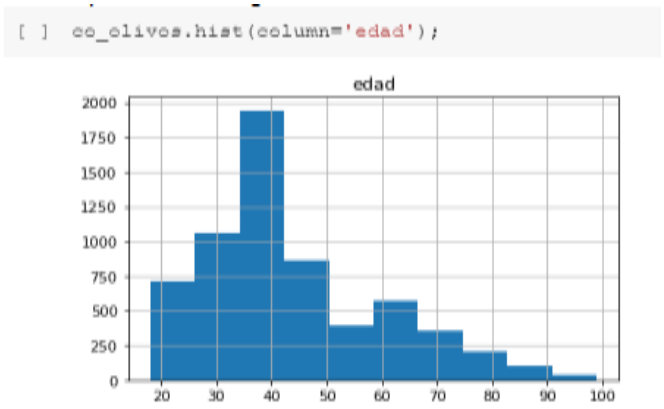

Fig. 5 Rango de edad de personas contagiadas con COVID-19 del Distrito de Los Olivos

En la figura 6 visualizamos que en su mayoría las personas utilizan como medio de transporte la combi y el bus.

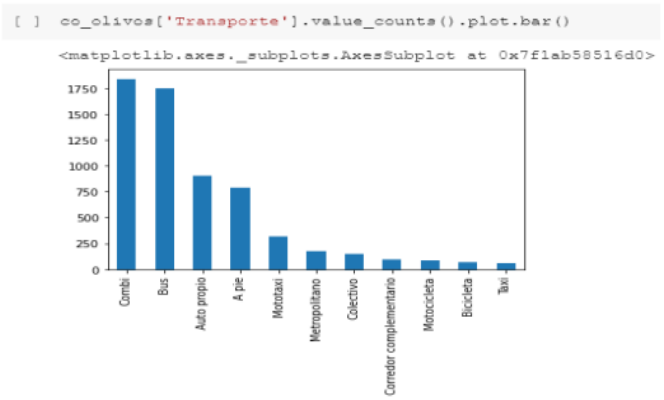

Fig. 6 Medio de transporte empleado en el Distrito de Los Olivos

En la figura 7 se observa que las personas mayormente demoran 30 minutos para arribar a su centro laboral.

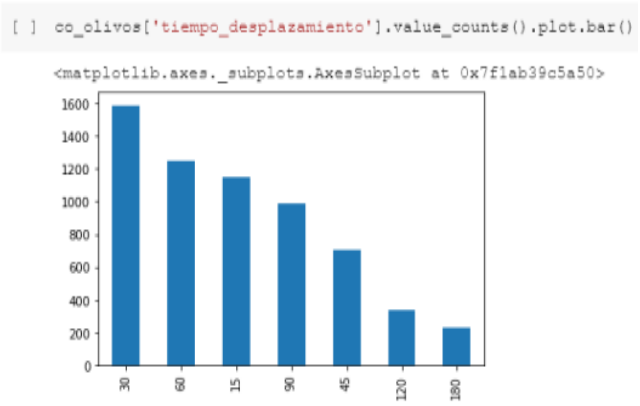

Fig. 7 Minutos empleados para llegar al centro laboral en el Distrito de Los Olivos

Fase 4

Al tratarse de un problema de clasificación binaria vamos a validar con la técnica de modelado árbol de decisiones (decisions Tree). Al contar con los algoritmos dividiremos el dataset (conjunto de datos) en forma aleatoria en dos partes: conjunto de datos de entrenamiento y conjunto de datos de prueba. Se ejecuta el experimento en el módulo Split que se observa en la figura 8 donde se utilizó el 70% de datos para entrenamiento y el 30% de datos para comprobar la eficacia del modelo.

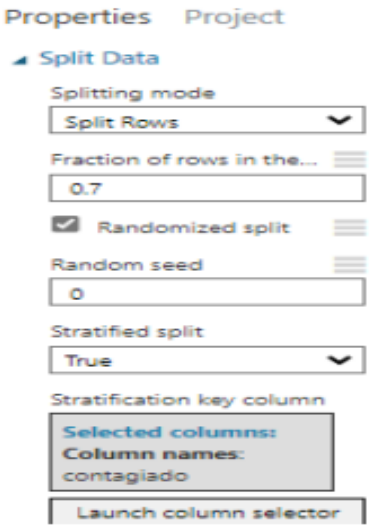

Fig. 8 Modo Split Data

Se procede a entrenar los modelos escogidos con la herramienta train model como se visualiza en la figura 9, se utiliza para entrenar de manera supervisada con los datos de entrenamiento

ingresados y se agrega score model para realizar la respectiva puntuación del modelo entrenado.

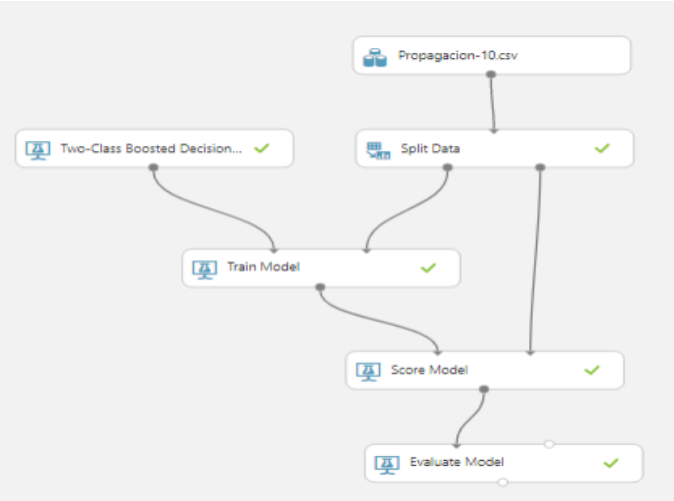

Fig. 9 Modelo predictivo de árbol de decisiones

Después de desarrollar el modelo entrenado en machine learning nos genera un archivo .csv el cual procedemos adjuntarlo como base de datos a la plataforma No-code de AppSheet , tal como se observa en la figura 10.

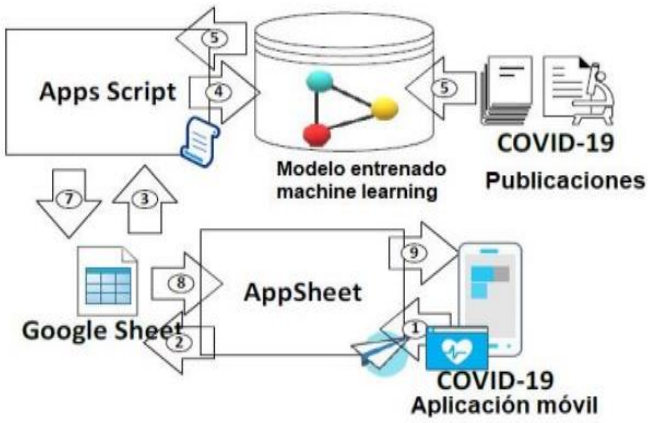

Fig. 10 Arquitectura lógica del Sistema

Tambien estudios relacionados con el trabajo de investigación:

El sistema de monitorización que ha erradicado el COVID-19 en China: ¿Se podría implantar en la UE? : En el trabajo se estudia una aplicación QR la cual utiliza técnicas de big data e información de organismos estatales y particulares para controlar y prevenir la propagación del COVID-19 en China. Al utilizar la aplicación móvil por primera vez se registra la temperatura corporal de la persona (las personas voluntariamente registran sus datos en la aplicación móvil). Hay colores que ha fijado el gobierno para este app: verde (la persona puede trasladarse libremente por el espacio público), amarillo (la persona tiene que hacer confinamiento por siete días o rojo (confinamiento automático de 14 días) [10].

Desarrollo y aplicación de técnicas de Machine Learning para la predicción de contagios por COVID-19: La tesis estudia sobre la predicción de contagios por COVID-19 en España. Como variables de entrada se ha empleado temperaturas máxima y mínima, extensión de una provincia, densidad poblacional, personas que han llegado de viaje por vía aérea. Con las variables obtenidas se determina cuales han influido en la cantidad de contagios [11].

IV. RESULTADO, DISCUSION Y CONCLUSION

Como resultado de desarrollar el modelo predictivo en árboles de decisión, se realizó el análisis de datos de personas expuestas al COVID-19 entre los meses de marzo del 2020 a noviembre del 2021 del Distrito de Los Olivos. El dataset de personas que realizan actividad presencial en el Distrito de los Olivos consta de 6238 filas y 18 columnas. En la figura 11 se visualiza tanto la curva AUC y la matriz ROC. Se obtuvo el siguiente resultado:
True positive: 1488 personas se contagiaron.
False positive: 136 personas se clasificaron erróneamente como contagiados
True negative: 3243 personas no se contagiaron
False negative: 123 personas se clasificaron erróneamente como no
contagiados.
Lo que género:
Exactitud: 94.8%
Precisión: 91.6%

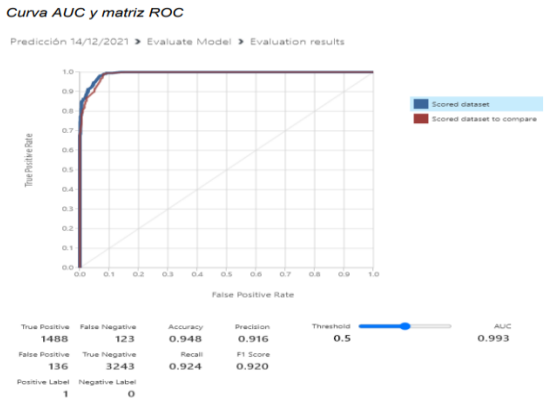

Fig.11 Matriz ROC

En la figura 12 se visualiza el dashboard del total de vacunados, al ingresar podemos validar el número de personas que están o no vacunadas. Al seleccionar sí o no en el dashboard nos muestra un reporte con el resultado, la hora, la fecha y el lugar donde se realizó el test.

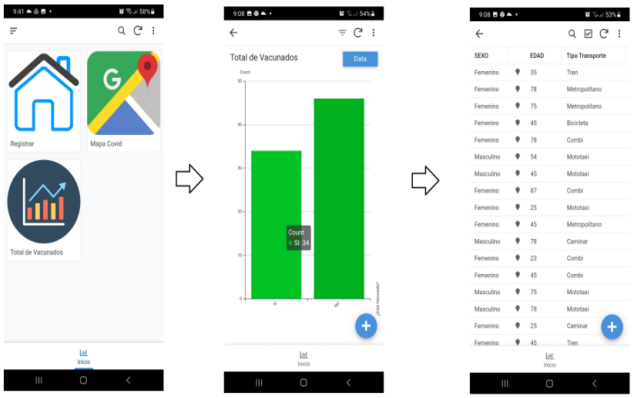

Fig.12 Dashboard visualizar resultado

REFERENCIAS

[1] V. Arias, . J. Salazar, . C. Garicano, . J. Contreras, . G. Chacón, M. Chacín-González, . R. Añez, . J. Rojas y V. Bermúdez-Pirela, «Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos,» *Revista Latinoamericana de Hipertensión,* vol. 14, nº 5, 2019.

[2] G. Rodríguez, «INDUSTRIAL IoT. MACHINE LEARNING EN LA INDUSTRIA 4.0,» *UPCommons,* 2020.

[3] J. Díaz-Pinzón, «Proyección de la propagación del COVID-19 en Colombia,» *Revista Med,* vol. 28, nº 1, 2020.

[4] J. De la Puente, «La gran depresión y el fracaso peruano. Balance de la primera ola del coronavirus,» *Revistas - Universidad de San Martín de Porres,* 2021.

[5] Plataforma Nacional de Datos Abiertos, 2023. [En línea]. Available: https://www.datosabiertos.gob.pe/search/field_topic/covid-19-917?sort_by=changed. [Último acceso: Noviembre 2022].

[6] S. Liu, J. McGree, Z. Ge y Y. Xie, Computational and Statistical Methods for Analysing Big Data with Applications, Academic Press, 2016.

[7] INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA, «ESTADÍSTICAS de la Criminalidad, Seguridad Ciudadana y Violencia,» 2022.

[8] C. Aranguren y . A. Flores, Artists, *Optimización del sistema web de la Veterinaria Dueñas para identificar casos de insuficiencia renal mediante árbol de decisiones.* [Art]. USMP, 2021.

[9] Autoridad Nacional para la Innovacion Gubernamental de Panama, «Marco de implementacion de proyectos de analisis de datos,» 2021.