Universidad del Rosario

MACC Matemáticas Aplicadas y Ciencias de la Computación

Tercera entrega proyecto probabilidad 2021-I

Germán David Plazas Cayachoa, David Alfonso Oviedo Salamanca, Santiago Rodríguez Morales.

Escuela de Ingeniería, Ciencia y Tecnología, Universidad del Rosario

1. Resumen ejecutivo.

El virus del COVID-19 ha generado un gran impacto en el país y especialmente en la ciudad de Bogotá. Sin embargo, los ciudadanos aún no conocen ciertos datos de suma importancia. Debido a esto, se decidió estudiar más a fondo diversos factores sociales y específicos de la población de la ciudad capital. Por lo anterior, este proyecto busca concluir que existen relaciones importantes de cómo afecta el virus a los habitantes de la ciudad según su sexo, edad y localidad de residencia. Como objetivo extra, se desea buscar un modelo para poder predecir el número de casos acumulados que dejará esta enfermedad en alguna fecha futura.

Para concluir que existen relaciones entre nuestras variables, se realizó un análisis estadístico descriptivo comparando algunas de los factores de nuestro problema considerado. Este proceso se dividió en 7 análisis y en cada uno se concluyó.

En los análisis considerados se concluyó que: Primero, el virus ha afectado más en cuanto a muertes y casos a las localidades con mayor población sin tener en cuenta factores económicos. Segundo, los hombres han fallecido en mayor cantidad a causa de este virus que las mujeres y más aún en edades entre 60 y 80 años, a pesar de que el número de contagiados sea similar entre ambos sexos. Tercero, no se encontró una diferencia notoria entre el promedio de edad de fallecimiento entre localidades. Cuarto, se encontró que en las localidades de menos recursos fue más demorado el tiempo de diagnóstico del virus en comparación a otras localidades. Finalmente, se encontró una relación entre la edad y el número de fallecidos. Todos estos análisis nos llevaron a la conclusión general.

Conclusión general: Gracias a las conclusiones de los análisis realizados se evidenció que existen relaciones importantes entre factores sociales como la edad, sexo y localidad, así como características económicas y poblacionales asociadas a cada zona de la ciudad. Desafortunadamente, el virus ha reflejado problemáticas de la ciudad que deberían de ser atendidas. En adición a esto, los habitantes de la ciudad deberían de ser conscientes de ciertos comunes con el virus con el objetivo de que el autocuidado sea bien practicado. Finalmente, en cuanto a nuestro objetivo adicional, el modelo para predicción de casos en una fecha debe de ser revisado teniendo en cuenta que los últimos registros son afectados por la situación actual del país.

2. Descripción final del problema considerado.

Desde que se inició la pandemia por el COVID-19 se ha visto la necesidad de conocer el virus en todas sus facetas. Sabemos por hecho que Colombia es un país muy desigual en el ámbito socioeconómico y sanitario, por lo cual es importante investigar que tanto ha afectado el virus a la salud de los bogotanos teniendo en cuenta diversos factores sociales e individuales de cada persona.

Para el proyecto se utiliza la base de datos extraída de SaludData Bogotá, actualizada al 12 de abril de 2021, siendo una fuente confiable la cual contiene los casos confirmados de COVID-19 en la ciudad [1]. Esta base de datos contiene distintas variables que servirán a la hora de construir un análisis objetivo.

Nuestro problema principal para este proyecto trata de evidenciar una relación entre factores sociales e intrínsecos y el número de casos/fallecidos por COVID-19 en la ciudad de Bogotá. Como un factor influyente social se busca estudiar la relación entre la localidad de residencia y los casos de COVID. Como factores intrínsecos buscamos evidenciar como afecta el COVID a las personas por edad y sexo. Adicionalmente vamos a utilizar un modelo de regresión lineal con el que predeciremos el número de casos acumulados a futuro.

3. Conjunto definitivo de datos seleccionado

Los datos que deseamos analizar se encuentran recopilados en el repositorio GitHub `https://github.com/DavPlazas/ProyectoProbabilidad1` Estos datos serán utilizados para realizar el análisis descriptivo el cual permite obtener representaciones comprensibles que actúan de manera conclusiva a nuestro objetivo principal. En la carpeta Archivos.xlsx se encuentran 7 archivos de Excel en los cuales están distribuidos distintos datos a conveniencia de los análisis descriptivos que serán interpretados.

El archivo Casos_sexo.xlsx contiene el sexo masculino o femenino de todos los pacientes contagiados por COVID.

El archivo Casos_acumulados_por_fecha.xlsx contiene las columnas: fecha de diagnóstico, casos por día, casos acumulados y los días transcurridos desde el primer caso detectado. Estos datos sirven para obtener el modelo de regresión lineal de los casos acumulados.

El archivo ProyectoEdadesMuertos.xlsx contiene dos columnas con las edades de todos lo hombres y mujeres fallecidos respectivamente.

El archivo TiempoDiagnostico.xlsx contiene las columnas de nombre de localidad, su respectivo id y el tiempo promedio de diagnóstico en días.

El archivo promedio_edad_muertos_localidades.xlsx contiene los datos de la localidad, su id, y el promedio de edad de los muertos por localidad.

El archivo Numero_casos_x_localidad.csv contiene los datos de localidad, id, y el número de casos y muertes en cada una de ellas. Estos sirven para obtener el mapa de calor de

Bogotá con el número de casos y muertes por localidad.

El archivo Edades_y_fallecidos.xlsx contiene la edad y los números de fallecidos por cada edad.

El archivo Muertes_por_Localidad.xlsx contiene los datos de las localidades de las muertes.

Muertes_sexo.xlsx contiene el sexo masculino o femenino de todos los pacientes fallecidos por COVID.

En la carpeta archivos varios se encuentran las imágenes de las gráficas a utilizar y la imagen (polígono) para el mapa de el distrito de Bogotá y sus localidades[2], y en la carpeta Code se incluye todo el código en R para el análisis descriptivo.

4. Análisis estadístico descriptivo.

En primer lugar, queremos visualizar de una manera amigable el estado de el número total de casos y defunciones por localidad. Para esto, utilizamos el polígono que nos representa el mapa del distrito de Bogotá y sus localidades. Finalmente, generamos un mapa de calor el cual sirve de ayuda para entender mejor el contexto general en el que se encuentra la ciudad dividida por localidades.

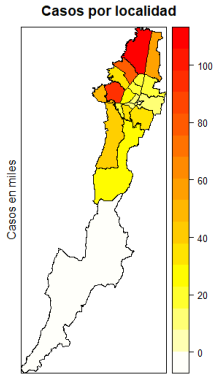

Figura 1: Mapa de calor de los casos postivos por localidad.

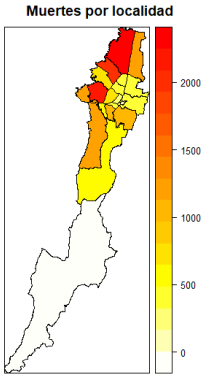

Figura 2: Mapa de calor de los fallecimientos por localidad.

Como segundo punto, queremos analizar que tanto afecta el virus en base al sexo de la persona y ver que tanta diferencia hay entre ambos. Para esto, utilizamos los archivos Casos_sexo.xlsx y Muertos_sexo.xlsx que nos brindan los datos para realizar los diagramas de Torta.

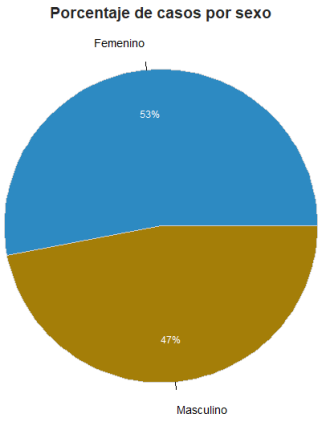

Figura 3: Proporción de casos en hombres y mujeres.

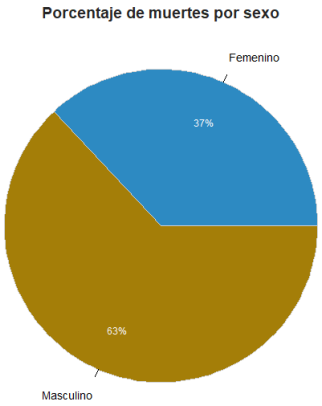

Figura 4: Proporción de falleciminetos en hombres y mujeres.

De esta manera, también queremos ver como se comporta el virus en base a la edad de ambos sexos; visualizando las muertes por grupos de edad y discriminando por hombres y mujeres. Esto lo logramos generando un histograma de dos variables cuyos datos obtuvimos del archivo ProyectoEdadesMuertos.xlsx

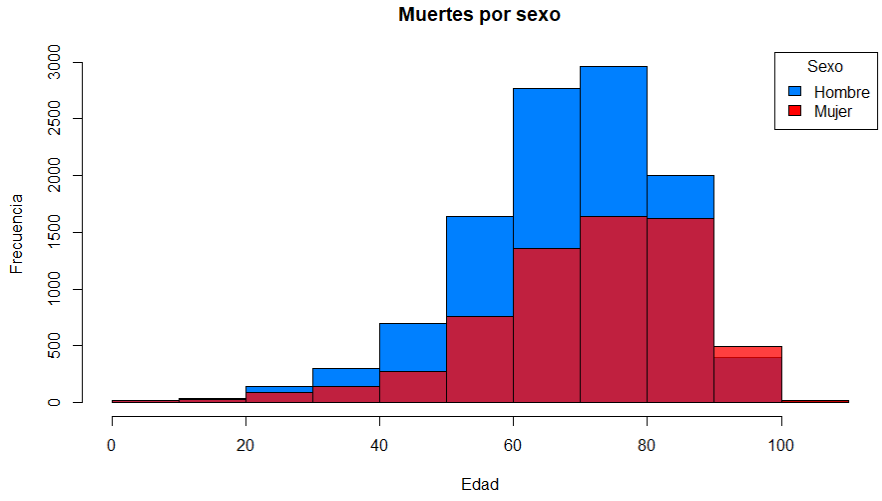

Figura 5: Histograma de muertes por edades y sexos

Para tener una mejor idea sobre los muertos por localidad realizamos dos diagramas de barras que nos brindan información sobre el número total de defunciones por localidad y también el promedio de edad de los muertos por cada una de las localidades. Con el promedio de edad de cada uno de los fallecidos se busca obtener con mayor precisión información sobre la realidad de diferentes sectores de la ciudad. Para los siguientes dia-

gramas de barras utilizamos los datos de los archivos Numero_casos_x_localidad.csv y promedio_edad_muertos_localidades.xlsx

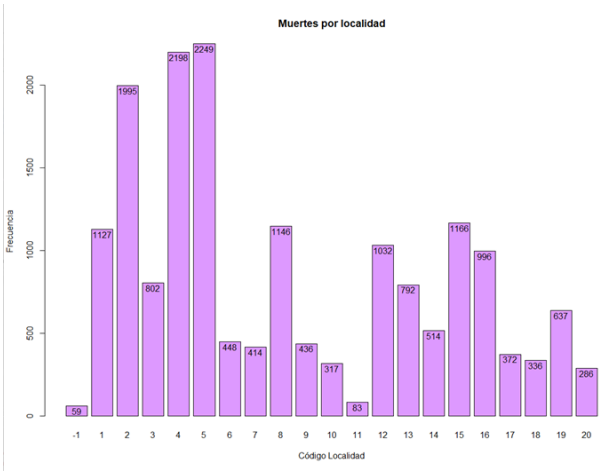

Figura 6: Diagramas de barras número de fallecidos por localidad

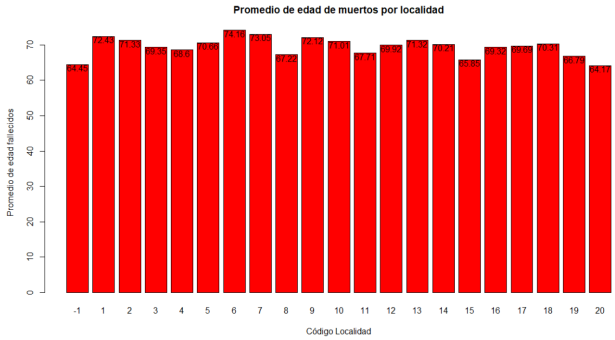

Figura 7: Diagrama de barras promedio de edad de los fallecidos por localidad

1: Usaquén.	8: Cuidad Bolívar.	15: Bosa.
2: Engativá.	9: Barrios Unidos.	16: San Cristóbal.
3: Fontibón.	10: Los Mártires.	17: Santa Fe.
4: Kennedy.	11: La candelaria.	18: Antonio Nariño
5: Suba.	12: Rafael Uribe Uribe.	19: Usme.
6: Teusaquillo.	13: Puente Arándano.	20: Fuera de Bogotá.
7: Chapinero.	14: Tunjuelito.	21: Sumapaz.

-1: Sin dato

Adicionalmente a los análisis descriptivos anteriormente descritos, se decide obtener información que nos brinde una idea más amplia en cuanto a como se ha manejado la detección del virus por cada una de las diferentes localidades de la ciudad, ya que esta puede también variar. Mencionado esto, se decide hacer un diagrama de barras que visibilice el tiempo de diagnóstico promedio en cada una de las localidades ya que esta es una manera sencilla de entender la información y comparar los resultados de cada localidad. Este diagrama lo realizamos utilizando los datos del archivo TiempoDiagnostico.xlsx.

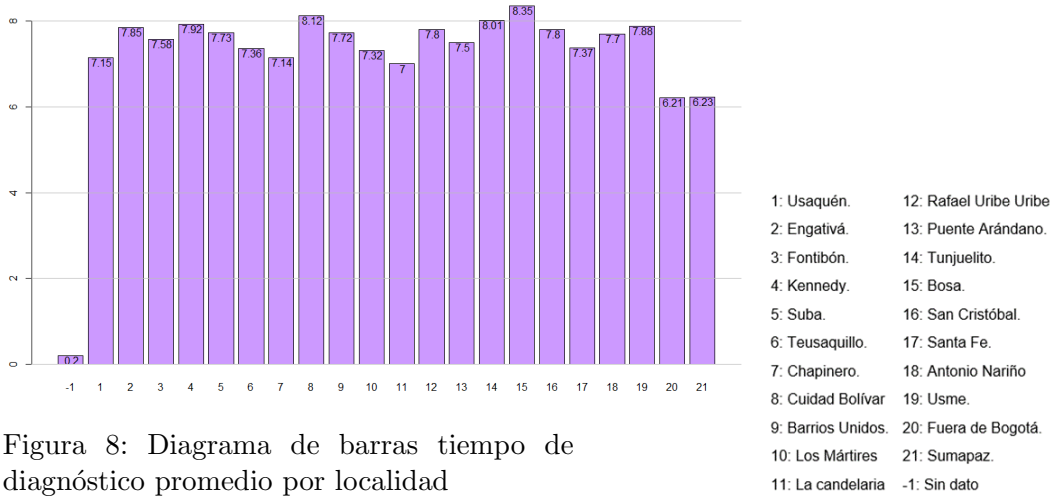

Figura 8: Diagrama de barras tiempo de diagnóstico promedio por localidad

Además, y de manera práctica, resulta conveniente aplicar lo aprendido durante el curso en cuanto al tema de regresión lineal para obtener una predicción de la cantidad de casos acumulados a futuro. Para esto es necesario trabajar con varios datos del archivo Casos-acumulados_por_fecha.xlsx. Esta información que queremos obtener la vamos a visualizar generando un diagrama de dispersión a lo largo del tiempo (días acumulados) y el total de casos que se van acumulando. Dicho diagrama resulta muy útil para poder entender el comportamiento del virus en la ciudad a lo largo del tiempo.

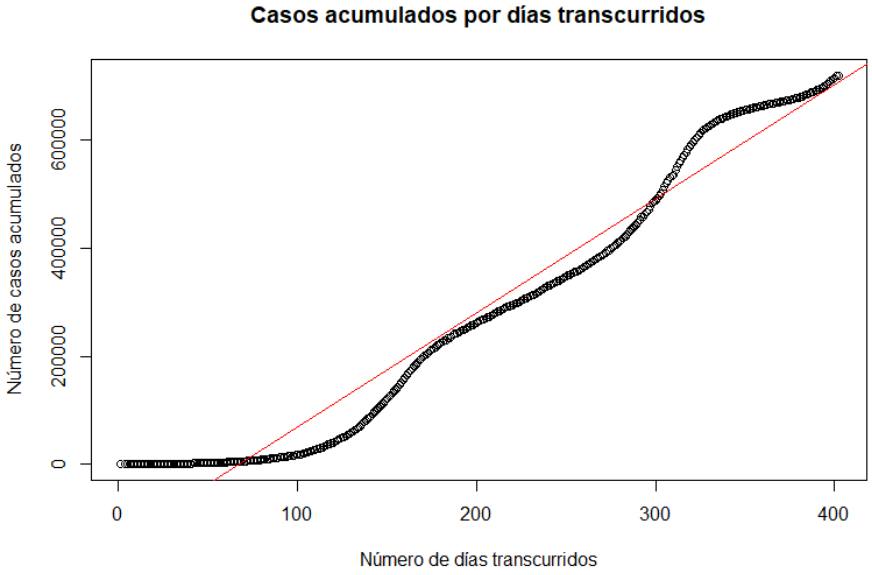

Figura 9: Diagrama de dispersión. Casos acumulados por días transcurridos

Como estudio concluyente se decidió realizar una prueba de hipótesis para determinar si

el número de casos acumulados depende del número de días transcurridos, como parece ser evidente en la Figura 9. Para ello, tomamos como hipótesis nula que el valor de β_1 es cero, y como hipótesis alternativa, tomamos que β_1 es diferente de cero. Anterior a utilizar un estadístico de prueba, se calcularon las variables necesarias para realizar una prueba de hipótesis ($\bar{X}$, $\bar{Y}$, S_{XX}, S_{XY}, S_{YY}, SSE, S). Se calculó el estadístico de prueba T y el valor de $T_{\alpha/2}$ arronjándonos como resultados 96.97 y 1.96 respectivamente, considerando α como 0.05. Dado que T está en nuestra región de rechazo, rechazamos la hipótesis nula, por lo que concluimos que el número de casos acumulados dependen del número de días transcurridos en la pandemia.

Finalmente, se decidió realizar un nuevo diagrama de dispersión del cuál haremos una regresión y prueba de hipótesis para poder obtener la relación entre la edad de los fallecidos y el número de muertos acumulados por edad. Para esto, es necesario trabajar con los datos del archivo Edad_y_fallecidos. Dicho diagrama resulta muy útil para poder entender y visualizar mejor cómo se comporta la cantidad de fallecidos según la edad.

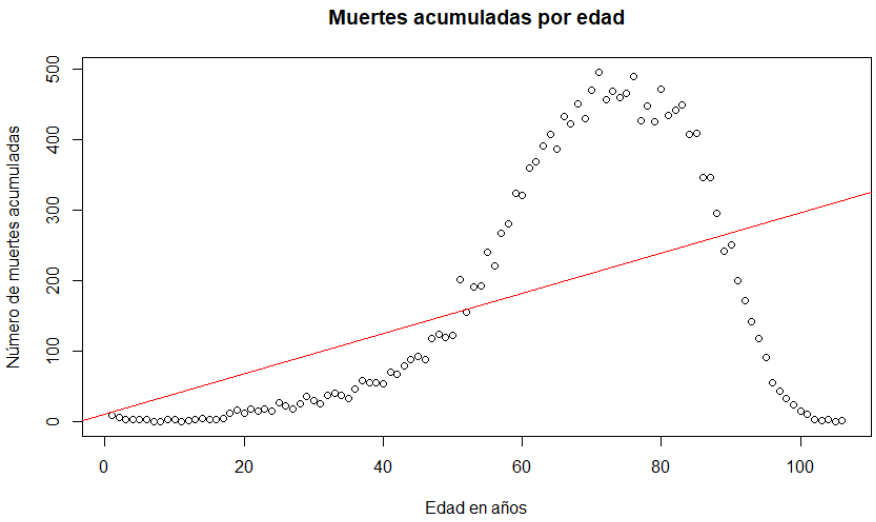

Figura 10: Diagrama de dispersión. Muertes acumuladas por edad

En base a la Figura 10, se decidió realizar una prueba de hipótesis para determinar si el número de muertes acumuladas depende de la edad de la población. Para lo anterior, consideramos como hipótesis nula que el valor de β_1 es cero, y como hipótesis alternativa, tomamos que β_1 es diferente de cero. Para el estadístico de prueba, tal cual como se hizo para nuestra anterior prueba de hipótesis, se calcularon las variables necesarias para realizar una nueva prueba ($\bar{X}$, $\bar{Y}$, S_{XX}, S_{XY}, S_{YY}, SSE, S). Se calculó el estadístico de prueba T y el valor de $T_{\alpha/2}$ arronjándonos como resultados 5.93 y 1.98 respectivamente, considerando α como 0.05. Dado que T está en nuestra región de rechazo, rechazamos la hipótesis nula, por lo que concluimos que el número de muertes acumuladas sí dependen de la edad de la población.

5. Conclusiones.

- **Figuras 1 y 2:** En base al mapa de calor de la Figura 1, se puede observar que Usaquén, Kennedy, Suba y Engativá son las localidades que más presentan casos de COVID-19 en Bogotá. Esto podría ser ocasionado por factores poblacionales, pues según el diario El Espectador y la Veeduría distrital [3], estas localidades son las que poseen un mayor número de habitantes, sin embargo, Usaquén y Suba no poseen una alta densidad poblacional a comparación de las otras localidades más afectadas. Ahora bien, en base a la Figura 2, no se reflejan mayores cambios con la Figura 1, pues los fallecidos han sido más abundantes en las localidades de mayor población (Usaquén, Kennedy, Suba y Engativá). Inicialmente, el grupo esperaba encontrar una gran diferencia entre localidades del sur de la ciudad con respecto al norte de ella, sin embargo, esto no fue así. En base a los anteriores datos, concluimos que el virus ha impactado a las localidades de mayor población en cuanto a casos y fallecimientos.

- **Figuras 3 y 4:** Concluimos según la visualización de la Figura 3 que hay una ligera mayoría de casos positivos de COVID-19 en mujeres que en hombres. Para ahondar más en el análisis que nos revela la proporción de casos, el diagrama de torta de la Figura 4 nos muestra la cantidad de fallecidos por sexos, que en contraste con la Figura 3 la cantidad de hombres fallecidos es superior a la de mujeres. Lo que nos revelan estos resultados dice mucho acerca de la salud o decisiones de vida que llevan en general ambos sexos en Colombia, o de que tan expuestos se encuentran al virus.

- **Figura 5:** En el histograma de la Figura 5 se puede observar la frecuencia de muertes por sexo y edad. Se evidencia en general que los hombres tienen una tasa de mortandad mayor a la de las mujeres por COVID-19. Estos resultados se podrían ver ampliamente explicados por distintos factores. En el año 2020, la pobreza monetaria alcanzó un porcentaje superior al 40 % en la ciudad de Bogotá [4], razón que explica en gran medida la necesidad de las personas para buscar el sustento del día a día. Siendo así, gran parte de los habitantes de la cuidad de Bogotá, se ven obligados a salir de sus casas generando una mayor probabilidad de contraer el virus ya sea por el contacto con el transporte público, o simplemente por el contacto fisico con las demás personas.

  Según Semana, el desempleo femenino en el 2020 fue del 20,4 % mientras que el desempleo para los hombres redondea un porcentaje del 13 % [5]. Esto demuestra el porqué la exposición de los hombres hacia el virus es mucho mayor que el de las mujeres. Sumado a esto, debemos tener en cuenta el factor genético, pues se ha demostrado que las mujeres son más resistentes que los hombres en circunstancias en las que la mortandad es realmente alta [6].

  En conclusión, en edades tempranas no se presenta una diferencia significativa entre los muertos por sexo, pero a medida que se aumenta la edad, se puede observar una diferencia significativa en las defunciones. Dado lo anterior, se puede evidenciar que entre las edades de 60 a 80 años se alcanza la mayor diferencia entre hombres y mujeres fallecidos en Bogotá.

- **Figuras 6 y 7:** Respecto a los diagramas de barras de las Figuras 6 y 7 logramos apreciar en primera instancia el número de fallecidos por localidad, lo que nos da una visión general cuantitativa sobre los fallecidos (esto teniendo en cuenta que hay localidades mucho más pobladas que otras). Luego, en la Figura 7, el diagrama nos revela de manera específica el promedio de edad de los fallecidos por cada una de las localidades, esto para poder concluir si entre localidades hay una diferencia significativa de edad, sin embargo como se puede apreciar no hay una diferencia notoria en el promedio de edad de los fallecidos por localidad.

- **Figura 8:** En este diagrama de barras, se analiza el tiempo promedio que se tarda en diagnosticar a un paciente COVID positivo desde que presenta síntomas. Las localidades de Bosa y Ciudad Bolívar presentan el mayor tiempo promedio en días para diagnosticar a un persona con COVID-19. Según Semana, Ciudad Bolívar, Usme, San Cristóbal y Bosa son las localidades mayor índice de pobreza en Bogotá [7]. Esto demuestra la gran desigualdad social y económica presente en el país, pues según este análisis, en los sectores más pobres de Bogotá, se presenta un tiempo mayor para diagnosticar el COVID-19 mientras que en otras zonas más favorecidas de la ciudad el tiempo de diagnóstico promedio es menor.

- **Figura 9:** En esta figura se puede observar un modelo de regresión lineal entre el número de días transcurridos desde el primer caso confirmado en la ciudad (Marzo 6 de 2020) y el número de casos confirmados acumulados. Este modelo lo utilizamos principalmente para poder predecir el número de casos acumulados en el futuro. Es de importancia aclarar que nuestro modelo no tiene en cuenta factores y/o eventos como vacunación y algomeraciones masivas. En nuestro experimento, en base a nuestro modelo, se predijo el número de casos acumulados al día Mayo 31 de 2021, arrojándonos como resultado 809215 casos. Sin embargo, a este día, ya se presentan 958281 casos acumulados en la ciudad [8]. Esta diferencia se puede explicar por la llegada de un nuevo pico de la pandemia en los meses de Abril y Mayo y la manifestaciones ciudadanas del último mes.

- **Figura 10:** En esta figura se puede observar un modelo de regresión lienal entre la edad y el número de muertes acumuladas producto del COVID-19 desde el inicio de la pandemia. Este modelo será utilizado principalmente para realizar una prueba de hipótesis que indique la relación entre estas variables, el cual fue desarrollado en la sección de estadístico de pruebas arronjándonos como resultado que sí existe una relación entre las variables.

6. URL GitHub.

URL del repositorio: `https://github.com/DavPlazas/ProyectoProbabilidad1`

Referencias

[1] Secretaría de Salud de Bogotá, Çasos confirmados de COVID-19 en Bogotá D.C.", Bogotá, 12 de Abril 2021. Base de Datos. Recuperado de: https://saludata.saludcapital.gov.co /osb/index.php/datos-de-salud/enfermedades-trasmisibles/covid19/

[2] M. Suárez, "Localidades de Bogotá D.C. - Archivo Carpeta Zip", Bogotá D.C., 2009. Recuperado de: https://sites.google.com/site/seriescol/shapes

[3] J. Torres en El Espectador, "Las 20 localidades de Bogotá en datos", ELESPECTADOR.COM, 2021. Disponible en: https://www.elespectador.com/bogota/las-20- localidades-de-bogota-en-datos-article-804728/.

[4] Ramos, F., 2021. Pobreza monetaria. Dane.gov.co. Recuperado de: https://www.dane.gov.co/ index.php/ estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria.

[5] Revista Semana. 2021. Pobreza en Colombia en 2020 fue de 42,5 % , hay 21 millones de pobres. Recuperado de: https://www.semana.com/economia/macroeconomia/articulo/pobreza-en-colombia-en-2020-fue-de-425/202129/.

[6] BBC Mundo. 2018. Por qué las mujeres son más resistentes que los hombres - BBC News Mundo. Recuperado de: https://www.bbc.com/mundo/noticias-42648551.

[7] Revista Semana. 2018. Así se construyó el mapa de los 403 barrios más pobres de Bogotá. Recuperado de: https://www.semana.com/nacion/articulo/el-mapa-de-los-403-barrios-mas-pobres-de-bogota/669419/.

[8] Ministerio de Salud. Coronavirus COVID-19. Consultado el 30-05-2021: https://covid19.minsalud.gov.co/