Análisis de Datos de Covid-19 en México

Luis Fernando Cisneros Chavez [1]

[1]*Universidad Nacional Autónoma de México*
Posgrado en Ciencia e Ingeniería de la Computación
Merida, México

Introducción

El análisis de datos del COVID-19 en México se ha convertido en una herramienta fundamental para comprender la propagación y el impacto de la pandemia en el país. México ha experimentado importantes desafíos en el manejo de la enfermedad, lo que ha llevado a la necesidad de recopilar, analizar y utilizar datos para tomar decisiones informadas y diseñar estrategias efectivas de respuesta.

El análisis de datos del COVID-19 en México abarca una variedad de aspectos, incluyendo el número de casos confirmados, las tasas de mortalidad, la distribución geográfica de los casos y la capacidad de pruebas realizadas. Estos datos se recopilan a nivel nacional, estatal y municipal, y se actualizan regularmente para proporcionar una imagen actualizada de la situación epidemiológica. El análisis de datos del COVID-19 en México es un proceso en constante evolución, ya que se integran nuevos datos, se mejoran las técnicas de modelado y se desarrollan herramientas de visualización más sofisticadas. La información derivada de estos análisis juega un papel crucial en la toma de decisiones estratégicas, la asignación de recursos y la implementación de políticas públicas para combatir la propagación del virus y proteger la salud de la población mexicana.

Datos a emplear

La Secretaría de Salud comparte con licencia de uso libre Datos preliminares a través de la Dirección General de Epidemiología. La información contenida corresponde únicamente a los datos que se obtienen del estudio epidemiológico de caso sospechoso de enfermedad respiratoria viral al momento que se identifica en las unidades médicas del Sector Salud en el País.

Del total de los datos se emplearon las siguientes columnas para su estudio: SEXO, FECHA_DEF, INTUBADO, NEUMONIA, EDAD, EMBARAZO, DIABETES,EPOC, ASMA, INMUSUPR, HIPERTENSION, OTRA_COM, CARDIOVASCULAR, OBESIDAD, RENAL_CRONICA, TABAQUISMO, OTRO_CASO, CLASIFICACION_FINAL y FECHA_DEF

Eliminando las de carácter étnico, nacionalidad, indígena y origen de los datos.

Análisis Exploratorio de los Datos

análisis exploratorio de datos es un proceso fundamental para comprender la estructura, las relaciones y las características de un conjunto de datos.

Los pasos que se siguieron para la exploracion fueron los siguientes:

- Obtener los datos.
- Familiarizarse con los datos: Se examinaron los datos para comprender su estructura y formato. Esto incluyó verificar el tipo de variables presentes (numéricas, categóricas, etc.), la cantidad de registros y las características específicas de cada columna.
- Limpieza de datos: Se limpiaron los datos eliminando valores atípicos y datos faltantes. Esto garantiza que los resultados del análisis sean más confiables.
- Resumen estadístico: Se obtuvo una visión general de la distribución de los datos.
- Visualización de datos: Utilizaron gráficos y visualizaciones para explorar los datos.
- Segmentación de datos: Se dividieron los datos en subgrupos basados en características específicas.

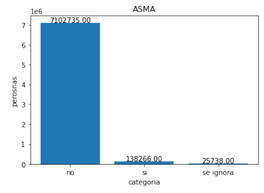
Referencias

[1] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. *Ensemble machine learning: Methods and applications*, pages 157–175, 2012.

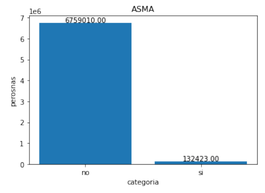
[2] Subdirector de Notificación y Registros Epidemiológicos. Bases de datos covid 19 en méxico, 2023.

Resultados del análisis exploratorio

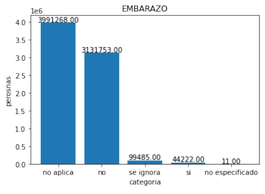
Análisis de Comorbilidades



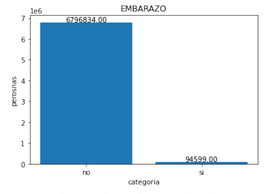
(a) Datos de asma



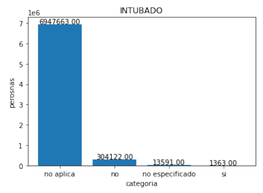
(b) Datos de asma limpios



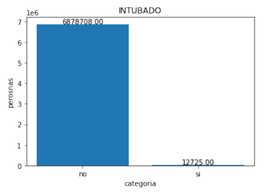
(c) Datos de embarazo



(d) Datos de embarazo limpios

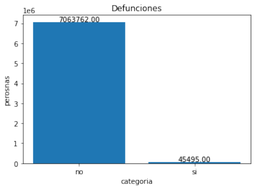


(e) Datos de intubación

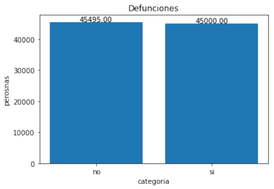


(f) Datos de intubación limpios

Análisis de Mortalidad



(g) Datos de defunciones



(h) Datos de defunciones equilibrado

Metodología

Para el análisis se empleó Random Forest, este es un algoritmo de aprendizaje automático supervisado que se utiliza para tareas de clasificación. Se basa en el concepto de "ensemble learning" (aprendizaje por conjunto), que combina múltiples modelos de aprendizaje para obtener resultados más precisos y robustos.

Un Random Forest está compuesto por un conjunto de árboles de decisión individuales, donde cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento y produce una predicción. Luego, la predicción final del Random Forest se determina a través de un proceso de votación o promediado de las predicciones individuales de los árboles.

Algunas de las ventajas del algoritmo Random Forest incluyen:

- Capacidad para manejar grandes conjuntos de datos con muchas características
- Buena capacidad de generalización.
- Tolerancia a valores atípicos y datos faltantes.
- Capacidad para evaluar la importancia relativa de las características en el modelo.
- Rendimiento eficiente en términos de tiempo de entrenamiento y predicción.

Resultados análisis de comorbilidad

Se realizo un análisis empleando las columnas SEXO, INTUBADO, NEUMONIA, EDAD, EMBARAZO, DIABETES, EPOC, ASMA, INMUSUPR, HIPERTENSION, OTRA_COM, CARDIOVASCULAR, OBESIDAD, RENAL_CRONICA, TABAQUISMO, OTRO_CASO; como características y CLASIFICACION_FINAL como predictor. El objetivo es medir que tanto contribuyen las comorbilidades y características como la edad para enfermar de Covid-19.

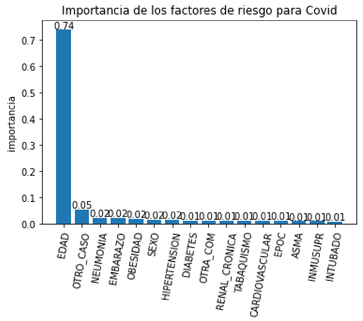
Se empleo un Random Forest entrenado con 7109257 de casos con parámetros 'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': 0.666, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 500, 'oob_score': True

El resultado de la evaluación con datos para verificar fue de: 0.5562

Los principales factores de riesgo para padecer covid-19 son:

predictor	importancia
EDAD	0.739672
OTRO_CASO	0.054760
NEUMONIA	0.024534
EMBARAZO	0.022625
OBESIDAD	0.020619
SEXO	0.015441
HIPERTENSION	0.015111
DIABETES	0.014147
OTRA_COM	0.014052
NAL_CRONICA	0.013153
TABAQUISMO	0.012573
DIOVASCULAR	0.011580
EPOC	0.011491
ASMA	0.011114
INMUSUPR	0.011049
INTUBADO	0.008079

(i) Tabla factores de riesgo



(j) Gráfica factores de riesgo

Resultados análisis de mortalidad

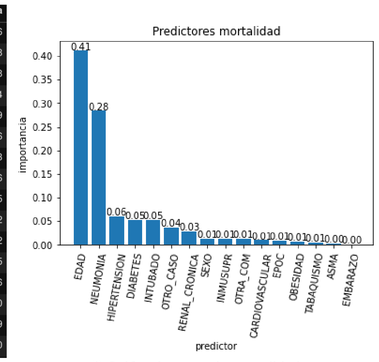
Se realizo un análisis empleando las columnas SEXO, INTUBADO, NEUMONIA, EDAD, EMBARAZO, DIABETES, EPOC, ASMA, INMUSUPR, HIPERTENSION, OTRA_COM, CARDIOVASCULAR, OBESIDAD, RENAL_CRONICA, TABAQUISMO, OTRO_CASO; como características y FECHA_DEF como predictor. El objetivo es medir que factores incrementan la mortalidad al padecer covid-19.

Se empleo un Random Forest entrenado con 90495 de casos con parámetros iguales al anterior. En este caso fueron menos ya que se busco equilibrar la característica de defunciones.

El resultado de la evaluación empleando los datos de entrenamiento: 0.924599
Evaluación con datos para verificar: 0.9082

predictor	importancia
EDAD	0.411276
NEUMONIA	0.283938
HIPERTENSION	0.061368
DIABETES	0.052354
INTUBADO	0.052199
OTRO_CASO	0.036566
RENAL_CRONICA	0.028693
SEXO	0.013546
INMUSUPR	0.013245
OTRA_COM	0.012572
CARDIOVASCULAR	0.009932
EPOC	0.008735
OBESIDAD	0.006816
TABAQUISMO	0.005070
ASMA	0.002999
EMBARAZO	0.000690

(k) Tabla factores de mortalidad



(l) Gráfica factores de mortalidad

paaleeleepipedo@gmail.com