Proyecto Final.

Algoritmo de Aprendizaje Automático para Clasificación de EPS y reasignación de usuarios de EPS SURA.

Asignatura: Programación.

Docente: Andrés Quintero Zea

Integrantes:

Sebastián Escobar Restrepo

Laura Gutiérrez Arias

Jerónimo Valencia Vallejo

Universidad EIA

Escuela de Ingenierías y ciencias básicas

Envigado, Antioquia

2024-1

ÍNDICE

1. **Título del Proyecto**

Algoritmo de Aprendizaje Automático para Clasificación de EPS y reasignación de usuarios de EPS SURA.

2. **Introducción**

Este proyecto tiene como objetivo aplicar los conocimientos adquiridos durante el semestre en la materia de Programación. Se realizará un algoritmo de aprendizaje automático para clasificación de la Población Activa del régimen contributivo según las EPS. Se entrena sin los pertenecientes a la EPS SURA y después, se le imputan algunos pertenecientes a esta para ver a cuál de las demás EPS sería más pertinente para una reacomodación en el contexto del retiro de la misma entidad del sistema de salud Colombiano.

3. **Marco Teórico**

Una EPS es una organización que gestiona los recursos del Sistema General de Seguridad Social en Salud (SGSSS). Su función principal es asegurar que los afiliados reciban los servicios médicos y asistenciales estipulados en el Plan Obligatorio de Salud.

En este contexto, hay diversas Entidades Promotoras de Salud en nuestro país, y cada una dispone de una red de proveedores de servicios médicos con los cuales tiene acuerdos. Estos proveedores incluyen hospitales, clínicas, consultorios, laboratorios, entre otros.

En este sistema existen dos tipos de afiliación: el régimen contributivo y el régimen subsidiado. En el régimen contributivo se incluyen las personas que tienen un contrato de trabajo, los servidores públicos, los pensionados y jubilados, así como los trabajadores independientes con capacidad de pago.

Por otro lado, el régimen subsidiado está destinado a las personas que no tienen la capacidad de pagar el monto total de la cotización por lo que se apoyan de subsidios del gobierno.

Existen diversos factores a tener en cuenta en el momento de escoger una EPS. Los principales son:

- Cobertura: se busca una EPS con mayor alcance en la zona de residencia o trabajo.
- Costos: las EPS ofrecen distintos planes para las diversas necesidades y presupuestos. Esto se relaciona con si es cotizante (paga por sus servicios), beneficiario (se beneficia por conducto del cotizante) o adicional (no cumple los requisitos de los anteriores, sin embargo se inscribe en el núcleo familiar del cotizante.
- Requisitos de afiliación: factores como la edad, estado de salud, situación laboral, entre otros, son factores que tienen en cuenta las entidades al momento de aceptar a un solicitante de los servicios ofrecidos.

Lo anterior justifica que, aunque las EPS no están directamente relacionadas con las características individuales de un beneficiario, sí pueden influir en la elección de la entidad a la que pertenece.

Ahora bien, el sistema de salud colombiano ha enfrentado en los últimos años una crisis sin precedentes. Se han liquidado más de 20 EPS, muchas otras están bajo intervención de la Superintendencia Nacional de Salud (Supersalud) y varias más están en riesgo de ser las próximas afectadas por esta situación crítica.

Las causas de estas crisis son multifactoriales e incluyen corrupción, ineficiencia, desfinanciamiento y desigualdad en el acceso a la atención médica (Chamorro et al., 2024).

Cuando el presidente Gustavo Petro seleccionó la reforma del sistema de salud como su principal objetivo político a principios de 2023, el sistema de salud ya había pasado por una de sus pruebas más difíciles: la pandemia de 2020. Durante ese periodo, las necesidades de recursos aumentaron significativamente mientras la economía se desplomaba. Por esta razón, el gobierno de Iván Duque permitió a las EPS utilizar parte de sus reservas en los meses más críticos de la crisis sanitaria. Después de que la reforma propuesta por Petro, que esencialmente buscaba eliminar a las EPS o reducirlas significativamente, fuera rechazada en abril pasado, el impacto de ese gasto excepcional se convirtió en un argumento utilizado por el gobierno para criticarlas. Esto llevó a que la Superintendencia de Salud interviniera en varias de ellas, marcando el inicio del fin del sistema de salud vigente en Colombia durante las últimas tres décadas.

Una de las tangentes de esta crisis que más impacto e incertidumbre ha causado, especialmente en el contexto antioqueño, es el retiro del sistema de salud de la EPS Suramericana anunciado el martes 28 de Mayo luego de intentar encontrar sin éxito soluciones al desfinanciamiento progresivo del sistema que afecta a toda la cadena de valor (Peña, 2024).

Ante esta situación, se ve necesaria la reasignación progresiva de los más de 5 millones de afiliados de la entidad.

4. Metodología

4.1. Descripción de la base de datos

De acuerdo con la resolución 4622 del 2016, la Base de Datos Única de Afiliados almacena y publica en la plataforma ‘Datos Abiertos’ información concerniente a los afiliados del Sistema de Salud Colombiano. Para nuestro caso, tomamos la población activa del régimen contributivo para relacionar ciertas características de los usuarios con su respectiva EPS.

El dataset está conformado por 305 mil registros de 14 columnas: género, grupo etario, código y nombre de la entidad, régimen, tipo y estado del afiliado, condición del afiliado, zona de afiliación, departamento, municipio, división del, división y nivel del sisbén, y cantidad de registros por entrada. Estos datos se han registrado desde el 2 de julio de 2020,

y la última actualización fue el 6 de mayo de 2024, por lo que se puede considerar que están actualizados los datos.

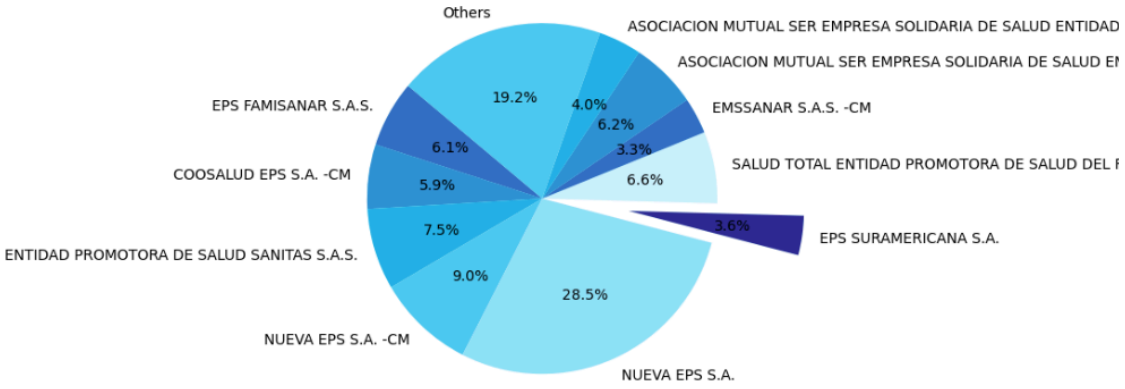
Ahora bien, es importante hacer hincapié en que estos datos son en su mayoría categóricos, sólo tienen una única columna numérica que es la del número de registros.

4.2. Exploración

En primera instancia se realizó un análisis de los valores únicos por columna para ver que valores posibles se podían encontrar en los datos.

Posteriormente se hicieron otros análisis como distribución de entradas nulas, cálculo de algunas medidas estadísticas (las aplicables para variables categóricas), y en última instancia se buscó que porción de la muestra eran los pertenecientes a la EPS SURA. Los resultados se pueden ver en la figura 1. Por motivos de extensión se muestra la figura cortada, la figura completa está en el archivo de exploración de datos.

Figura 1: Porcentaje de entradas pertenecientes a la EPS SURA



4.3. Limpieza y Preprocesado

Se eliminaron algunas columnas en concordancia con lo encontrado en la fase exploratoria: valores nulos y relación de las columnas con la clasificación por entidad. Siendo así, se eliminaron las columnas de Código de la entidad, Régimen, Municipio, Nivel del Sisbén, Cantidad de registros.

Como se había mencionado anteriormente, este dataset tenía en su mayoría variables tipo objeto por lo que se utilizaron 2 métodos de codificación:

- One-Hot: para variables más booleanas, crea columnas separadas para cada valor de la columna original y asigna valores de 1 o 0 según corresponda. Se le hizo esta categorización a: Tipo de afiliado, Estado del afiliado, Condición del beneficiario, Zona de Afiliación y Género ya que estas tenían como máximo 3 valores únicos.
- De Etiquetas: se le asigna un número a cada valor de texto correspondiente. Se le aplicó a: Grupo etario, Nombre de la entidad, Departamento, tps_nvl_ssb_id ya que estas tenían un número significativamente mayor de valores. Acto seguido se realizó un mapeo para ver qué valores se le asignaron a cada variable.

Es importante aclarar que con el paso anterior, se hizo una imputación y codificación a la vez ya que para los datos tipo NaN, se les asignó también un valor. Es decir, los NaNs, también entrarán dentro de la clasificación. Teniendo esto en cuenta se consideró borrar ese 86% de datos faltantes ya que aún quedaría un dataframe de más de 5000 instancias, sin embargo se hizo un análisis y todos los beneficiarios de EPS Sura tenían en esa columna un NaN por lo que afectaría el propósito del proyecto, así que se eliminó también esta columna.

4.4. Modelos de Aprendizaje

Como parámetro del proyecto, se deben escoger 2 modelos de aprendizaje para comparar su rendimiento y escoger el mejor en el contexto del data-frame empleado.

4.4.1. Random Forest

Se escogió de primera mano este modelo ya que dentro de los ofrecidos por scikit learn, era el más pertinente para este conjunto de datos, ya que es muy rápido.
En esencia, este algoritmo de clasificación genera diferentes árboles de decisión al azar, para tomar decisiones sobre las clasificaciones para que cada árbol sea completamente independiente y la decisión final sea lo más fuerte posible.

Paralelamente se hizo un algoritmo de validación cruzada, el cual funciona al aplicar un rango de hiper parámetros y de manera aleatoria busca la mejor combinación de los mismos según un número de iteraciones hasta encontrar la más pertinente, imprimiendo el accuracy y los hiper parámetros de la mejor iteración.

4.4.2. XG Boost

Este algoritmo toma una serie de árboles de decisión al igual que en el método anterior (aunque este toma menos) y se busca secuencialmente los árboles más débiles y se entrenan más para fortalecerlos y emparejar la capacidad de cada árbol de decisión.
Lo anterior hace que este algoritmo sea significativamente más lento que el anterior, sin embargo, aumenta la precisión al fortalecer cada uno de los árboles según el entrenamiento que necesiten para los datos que se le entregan.

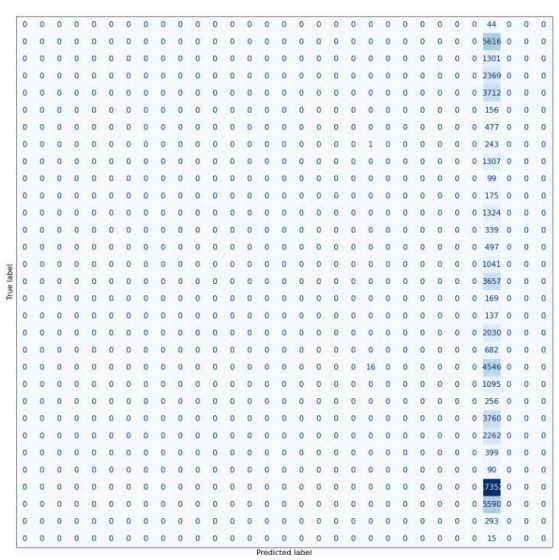
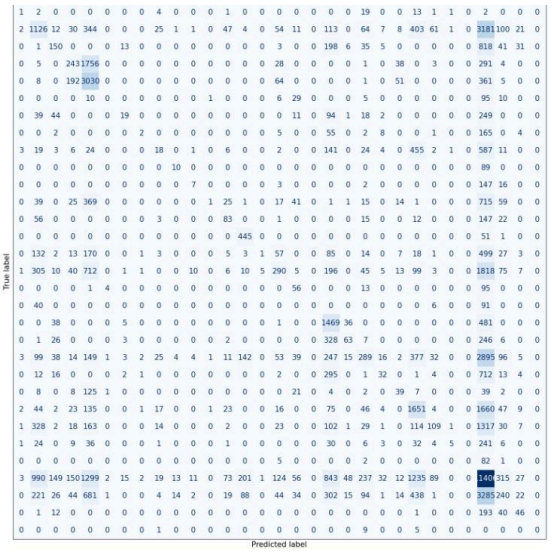
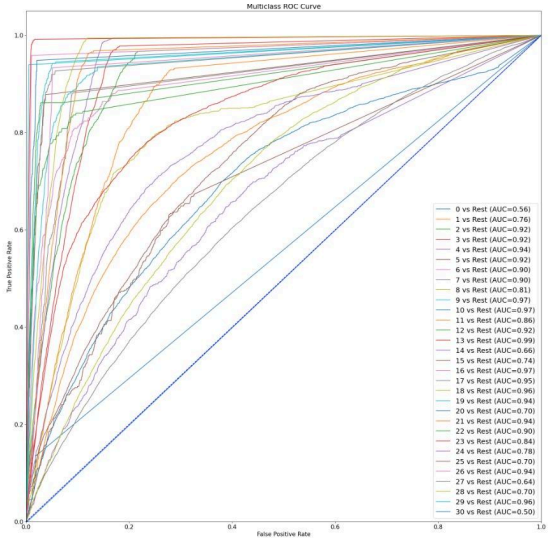
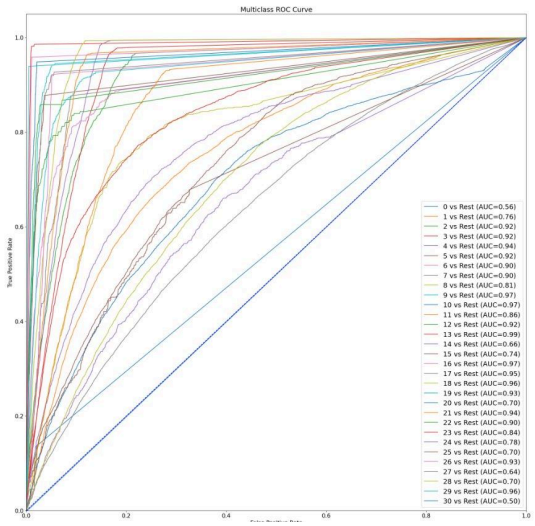
A este método también se le aplicó el algoritmo de validación cruzada para obtener los mejores hiper parámetros y su respectiva accuracy.

4.5. Evaluación

Una vez entrenados los modelos se evaluaron para ver cuál de los dos funcionaría mejor para el conjunto de datos.
Para esto se hizo un reporte de clasificación que evaluaba la precisión con la que se predecía cada una de las EPS, y una matriz de confusión para relacionar el las predicciones con los resultados reales.
Finalmente se hizo una curva ROC que relaciona la especificidad (False Positive Rate) con la sensibilidad (True Positive Rate) pero para cada una de las variables que se están prediciendo (EPS) y así ver cuál de ellas tiene una mejor predicción.

Las gráficas para cada método de aprendizaje se pueden observar en la siguiente tabla:

Tabla 1: *gráficas de evaluación para los modelos de aprendizaje automático*

Evaluación de Rendimiento de los modelos	
Random Forest	XG Boost
Matriz de confusión	
	
Curva ROC	
	

Las gráficas de tamaño completo se encuentran en el archivo 05 - comparación.ipynb.

4.6. Reasignación de EPS

Para utilizar el modelo de clasificación como aplicativo para reasignar usuarios de EPS Suramericana a otras entidades, se escogió uno de los usuarios de SURA y se le entregó al modelo para que predijera su EPS, es decir, la EPS que más le conviene según sus características. Esto se hizo creando un diccionario con los datos del usuario, se convirtió en un dataframe para entregarle al algoritmo y que realizara la predicción.

5. Conclusiones

Con base en la evaluación realizada, se puede concluir que el modelo XG Boost demostró un rendimiento superior en términos de precisión, convirtiéndose así en la mejor opción para el conjunto de datos analizado.

En lo que respecta a la reasignación de afiliados beneficiarios, se observa que para un perfil específico - afiliado activo, sin discapacidad, ubicado en zona urbana, género femenino, perteneciente al departamento del Valle - el modelo Random Forest asigna preferentemente a la EPS 27 (NUEVA EPS S.A.), mientras que el XG Boost predice la asignación hacia la EPS 18 (EMSSANAR S.A.S.). Dado que el modelo XG Boost exhibe un mejor desempeño, se recomienda que los usuarios de este perfil sean transferidos a EMSSANAR S.A.S.

Es importante destacar que el conjunto de datos presentaba un notable desbalance, lo que puede introducir sesgos en la reasignación, especialmente hacia NUEVA EPS, que tiene mayor peso dentro de los datos (ver figura 1).

En cuanto a los métodos utilizados, se observa que en una comparación directa 1 contra 1, la diferencia entre los modelos no es significativa. Sin embargo, a medida que se aumenta el número de iteraciones, se evidencia que el XG Boost requiere un mayor poder computacional.

6. Recomendaciones y Futuras Aplicaciones

El algoritmo de reasignación se hizo únicamente para un solo paciente por lo que cada usuario se debe reasignar manualmente. Como futuras mejoras para el proyecto se sugiere hacer un algoritmo que aplique lo mismo pero para todos los usuarios de EPS sura automáticamente.

Para mejorar el rendimiento de los modelos, también se recomienda usar unos datos mejor balanceados y con mayor número de variables numéricas para tener una relación más limpia de las variables.

7. Referencias

Análisis y diseño de estructuras con STAAD.Pro - Nivel 1. (s. f.). Udemy.

https://www.udemy.com/course/staadpro-nivel-1/?couponCode=OF52424

Castañeda, C. A. P. (2024, 28 mayo). EPS Sura se retira del sistema de salud en Colombia:

¿qué pasará con los más de cinco millones de afiliados? *El Tiempo*.

https://www.eltiempo.com/salud/eps-sura-se-retira-del-sistema-de-salud-en-colombia-que-pasara-con-los-mas-de-cinco-millones-de-afiliados-3347138

Crisis del sistema de salud colombiano: un análisis urgente. (s. f.). Universidad Central. https://www.ucentral.edu.co/noticentral/crisis-sistema-salud-colombiano-analisis-urgente

De la Cruz Sebastián Serrano Vega, J. (s. f.). *2.10.- Problemas estáticamente determinados e indeterminados | Estática*. https://ecosistema.buap.mx/forms/files/dspace-29/210_problemas_estticamente_determinados_e_indeterminados.html

ESTACO S.A, & Angel, C. (s. f.). *Escuela de Ingeniería de Antioquia: Escaleras Bloque A* [Memorias de cálculo].

Estévez, S. (2008). DISEÑO ESTRUCTURAL DE LA CUBIERTA METÁLICA PARA DOS CANCHAS DE ECUAVOLEY. . https://bibdigital.epn.edu.ec/bitstream/15000/651/1/CD-1590%282008-07-15-01-14-35%29.pdf

Ministerio de Ambiente, Vivienda y Desarrollo Territorial. (1997). NSR-10 reglamento colombiano de construcción sismo resistente. *OMISION ASESORA PERMANENTE PARA EL REGIMEN DE CONSTRUCCIONES SISMO RESISTENTES*. https://repositorio.ucp.edu.co/bitstream/10785/6670/1/CDPEARQ207.pdf

Pixelpro. (2024, 18 abril). ¿Cómo elegir una EPS en Colombia? *Colombiana de Transplantes*. https://colombianadetrasplantes.com/web/institucional/como-elegir-una-eps-en-colombia/

Población Base de Datos Única de Afiliados BDUA del régimen contributivo | Datos Abiertos Colombia. (2024, 6 mayo).

https://www.datos.gov.co/Salud-y-Protecci-n-Social/Poblaci-n-Base-de-Datos-nica-de-Afiliados-BDUA-del/tq4m-hmg2/about_data

RAE. (2a. C.). Kilopondio. En *Diccionario de la Lengua Española*. https://dle.rae.es/kilopondio