Programa de Ciencia de Datos
Curso: Big Data
Profesor: Juan Manuel Esquivel
Estudiante: María Yorleni Alfaro Alfaro

Proyecto de Big Data

Investigación Preliminar – Primer Problema

Fuentes de datos analizadas

Para esta investigación se procedió a buscar y a analizar información de diferentes fuentes tanto a nivel laboral como fuentes de datos abiertas.

A nivel laboral no fue posible obtener información dado que la mayor parte de la información es de carácter confidencial, y la información considerada como pública es información que ya se encuentra agregada por lo cual no es funcional por los requerimientos del proyecto.

También se analizaron diferentes fuentes de datos de instituciones como el INEC, Programa Estado de la Nación, CCSS, Mideplan, otros ministerios del gobierno, etc.

Producto de este análisis se decidió escoger las siguientes dos fuentes de datos, para el desarrollo del proyecto:

- Base de datos de educación primaria, tomada del Programa Estado de la Nación, que muestra una serie de características de las escuelas públicas y privadas en Costa Rica.
- Fuente de datos con el índice de desarrollo social distrital de Costa Rica, tomado del Mideplan, que muestra el índice de desarrollo social para cada distrito segmentado por las siguientes dimensiones: económica, participación, salud y educativa.

El motivo por el cual se eligió la fuentes de datos de educación primaria es porque considero que contiene información muy valiosa que puede servir para realizar análisis de información de diferentes ámbitos como excelencia en la educación primaria (ya que muestra información de aprobados, reprobados, repitentes, etc), problemas sociales (estudiantes que han abandonado la escuela, embarazos en estudiantes, casos atendidos por violencia, robos, destrucción de materiales, etc), infraestructura en las escuelas (cantidad de aulas, laboratorios, comedores, bibliotecas, sodas, computadoras, etc). Todo lo anterior nos permite ver de una manera más detallada la realidad de Costa Rica en estos ámbitos.

El motivo por el cual se eligió la fuente de datos que contiene el índice de desarrollo social distrital, es porque aporta información muy relevante al set de datos de escuelas primarias, dado que nos permite conocer la realidad de cada uno de los distritos donde se encuentran ubicados los centros educativos, a nivel económico, educativo, de salud, etc., y esto nos ayuda a comprender mejor algunos de los ámbitos que se ven reflejados en el set de datos de escuelas primarias, como la excelencia en la educación, problemas sociales o infraestructura, ya que estos se encuentran muy ligados al desarrollo social que tenga cada distrito.

Descripción detallada de los datos

Educación Primaria

El set de datos de educación primaria, del Programa Estado de la Nación, contiene información de diferentes características de las escuelas en Costa Rica, desde el año 2000 hasta el año 2016, por lo tanto, se decidió tomar la información únicamente del año 2015, eliminando todo el resto de las columnas que no corresponden a este año. Además muchas de las columnas estaban dadas por año y por grado (1º , 2º , 3º , 4º, 5º y 6º) , por lo que se decidió utilizar únicamente las columnas que contienen el total por centro educativo (eliminando las que contienen los total por cada grado).

En la siguiente tabla se muestra el detalle de cada uno de los campos del set de datos de educación primaria:

Columna	Tipo de dato	Descripción
llave	Numérico	Llave única del registro
nombre_ins	String	Nombre del centro educativo
creacion00	Numérico	Año de creación de la institución
direg15	Numérico	Dirección Regional Educativa
cdpr15	Numérico	Código de Provincia
cdcan15	Numérico	Código de Cantón
cddis15	Numérico	Código Distrital
regplan15	Numérico	Región de Planificación
zona15	Numérico	Código de Zona
mit_15	Numérico	Matrícula Inicial Total en el año 2015
mih_15	Numérico	Matrícula Inicial de Hombres en el año 2015
rt_15	Numérico	Repitentes Total 2015
rh_15	Numérico	Repitentes Hombres 2015
aprobt_15	Numérico	Aprobados Total 2015
aprobh_15	Numérico	Aprobados Hombres 2015
reprot_15	Numérico	Reprobados Total 2015
reproh_15	Numérico	Reprobados Hombres 2015
desa_15	Numérico	Abandono Total 2015
desah_15	Numérico	Abandono Hombres 2015
desert_15	Numérico	Exclusión intra-anual Total 2015
deserh_15	Numérico	Exclusión intra-anual Hombres 2015
act_15	Numérico	Estudiantes con Adecuación de Acceso Total 2015
ach_15	Numérico	Estudiantes con Adecuación de Acceso Hombre 2015
nst_15	Numérico	Estudiantes con Adecuación No Significativa 2015
nsh_15	Numérico	Estudiantes con Adecuación No Significativa Hombres 2015
sit_15	Numérico	Estudiantes con Adecuación Significativa Total 2015
sih_15	Numérico	Estudiantes con Adecuación Significativa Hombre 2015
embt_15	Numérico	Estudiantes Embarazadas total 2015

embmenor_15	Numérico	Estudiantes Embarazadas Menor de 18 años 2015
embmayor_15	Numérico	Estudiantes Embarazadas con 18 años o más 2015
aat15	String	Aulas para impartir lecciones I y II ciclos total 2015
aab15	String	Aulas para impartir lecciones I y II ciclos buenas 2015
apt15	Numérico	Aulas para impartir lecciones Educación Preescolar total 2015
apb15	Numérico	Aulas para impartir lecciones Educación Preescolar buenas 2015
aest15	Numérico	Aulas para impartir lecciones Aula Integrada total 2015
aesb15	Numérico	Aulas para impartir lecciones Aula Integrada 2015
aaet15	Numérico	Aulas para impartir lecciones Aula Edad total 2015
aaeb15	Numérico	Aulas para impartir lecciones Aula Edad buenas 2015
anat15	Numérico	Aulas que no se utilizan para impartir lecciones total 2015
anab15	Numérico	Aulas que no se utilizan para impartir lecciones buenas 2015
inft15	Numérico	Laboratorio de Informática Total 2015
infb15	Numérico	Laboratorio de Informática Bueno 2015
olat15	Numérico	Otro Laboratorio Total 2015
olab15	Numérico	Otro Laboratorio Total Bueno 2015
salt15	Numérico	Sala de Profesores Total 2015
salb15	Numérico	Sala de Profesores Bueno 2015
comt15	Numérico	Comedor Total en 2015
comb15	Numérico	Comedor Bueno en 2015
bibt15	Numérico	Biblioteca Total en 2015
bibb15	Numérico	Biblioteca Bueno en 2015
gimt15	Numérico	Gimnasio en 2015
gimb15	Numérico	Gimnasio Bueno en 2015
talt_ai15	Numérico	Taller de Artes Industriales Total 2015
talb_ai15	Numérico	Taller de Artes Industriales Bueno en 2015
otalt15	Numérico	Otros Talleres Total 2015
otalb15	Numérico	Otros Talleres Bueno en 2015
sodt15	Numérico	Soda Total 20015
sodb15	Numérico	Soda Bueno 20015
indt15	Numérico	Inodoros Total 2015
indb15	Numérico	Inodoros Bueno 2015
lavt15	Numérico	Lavatorios Total 2015
lavb15	Numérico	Lavatorios Bueno 2015
sant15	Numérico	Servicio Sanitario Accesible Total 2015
sanb15	Numérico	Servicio Sanitario Accesible Bueno 2015
tvt15	Numérico	Televisión Total 2015
tvb15	Numérico	Televisión Bueno 2015
vbt15	Numérico	Proyector de video (Video Beam) Total 2015
vbb15	Numérico	Proyector de video (Video Beam) Bueno 2015
dvdt15	Numérico	DVD Total 2015
dvdb15	Numérico	DVD Bueno 2015
cetoi15	Numérico	Computadoras escritorio con internet

cetos15	Numérico	Computadoras escritorio sin internet
cepei15	Numérico	Computadoras escritorio con internet Pedagógico
cepes15	Numérico	Computadoras escritorio sin internet Pedagógico
cepai15	Numérico	Computadoras escritorio con internet Pedagógico y Administrativo
cepas15	Numérico	Computadoras escritorio sin internet Pedagógico y Administrativo
ceadi15	Numérico	Computadoras escritorio con internet Administrativo
ceads15	Numérico	Computadoras escritorio sin internet Administrativo
cptoi15	Numérico	Computadoras portátil con internet
cptos15	Numérico	Computadoras Portátil sin internet
cppei15	Numérico	Computadoras portátil con internet Pedagógico
cppes15	Numérico	Computadoras portátil sin internet Pedagógico
cppai15	Numérico	Computadoras portátil con internet Pedagógico y Administrativo
cppas15	Numérico	Computadoras portátil sin internet Pedagógico y Administrativo
cpadi15	Numérico	Computadoras portátil con internet Administrativo
cpads15	Numérico	Computadoras portátil sin internet Administrativo
bib15	Numérico	Servicio de Biblioteca
sal15	Numérico	Servicio de Salud
pla15	Numérico	Servicio de Planes de Emergencia
aux15	Numérico	Servicio de Primeros Auxilios
serv_int15	Numérico	Servicio de Internet
expto_15	Numérico	Expulsiones por agresión total 2015
expdef_15	Numérico	Expulsiones por agresión definitivas 2015
exptem_15	Numérico	Expulsiones por agresión temporales 2015
agrve_15	Numérico	Casos atendidos entre estudiantes de violencia verbal 2015
agrvep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia verbal 2015
agrveo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia verbal 2015
agrfe_15	Numérico	Casos atendidos entre estudiantes de violencia física 2015
agrfep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia física 2015
agrfeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia física 2015
agree_15	Numérico	Casos atendidos entre estudiantes de violencia escrita 2015
agreep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia escrita 2015
agreeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia escrita 2015
agrre_15	Numérico	Casos atendidos entre estudiantes de robos 2015
agrrep_15	Numérico	Casos atendidos entre estudiantes y docentes de robos 2015
agrreo_15	Numérico	Casos atendidos entre estudiantes y otro personal de robos 2015
agrde_15	Numérico	Casos atendidos entre estudiantes de destrucción de materiales 2015
agrdep_15	Numérico	Casos atendidos entre estudiantes y docentes de destrucción de materiales 2015
agrdeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de destrucción de materiales 2015
agroe_15	Numérico	Casos atendidos entre estudiantes de otros tipos de violencia 2015
agroep_15	Numérico	Casos atendidos entre estudiantes y docentes de otros tipos de violencia 2015

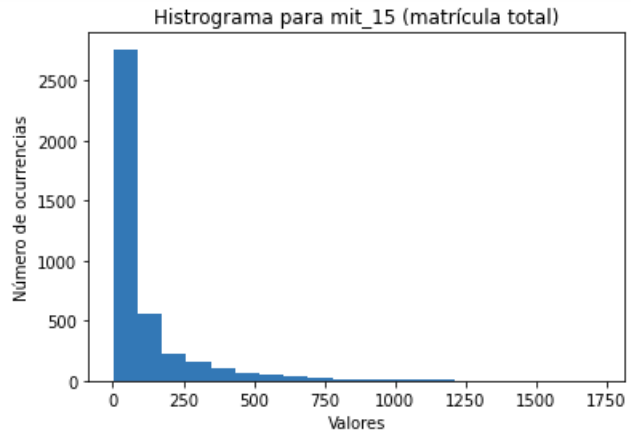
agroeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de otros tipos de violencia 2015
int_15	Numérico	Matrícula Inicial en Inglés Total 2015
inht_15	Numérico	Matrícula Inicial en Inglés Hombre 2015
rit_15	Numérico	Matrícula Inicial en Inglés Radio Interactiva Total 2015
rih_15	Numérico	Matrícula Inicial en Inglés Radio Interactiva Hombre 2015
frt_15	Numérico	Matrícula Inicial en Francés Total 2015
frh_15	Numérico	Matrícula Inicial en Francés Hombre 2015
itt_15	Numérico	Matrícula Inicial en Italiano Total 2015
ith_15	Numérico	Matrícula Inicial en Italiano Hombre 2015
extrant_15	Numérico	Matrícula Inicial Alumnos Extranjeros 2015
extranh_15	Numérico	Matrícula Inicial Alumnos Extranjeros Hombres 2015

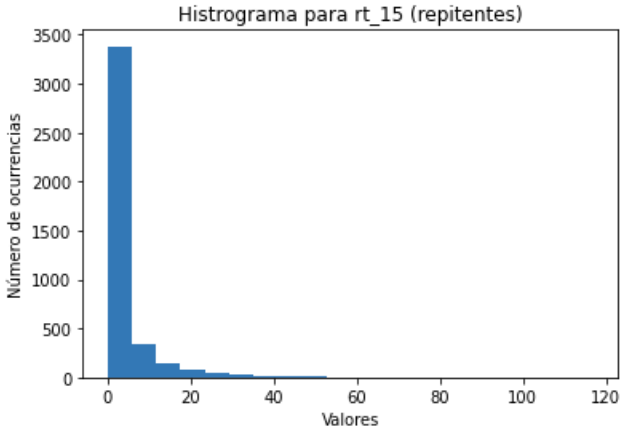
Cantidad de datos de la fuente: 4266 filas.

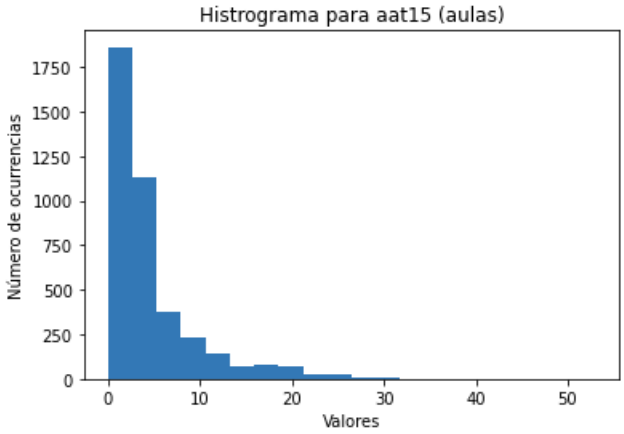
La fuente de datos fue tomada directamente de la página web del Programa Estado de la Nación (https://estadonacion.or.cr/base-datos/), específicamente la que tiene el nombre “Educación Primaria” (la cual corresponde a un archivo csv).

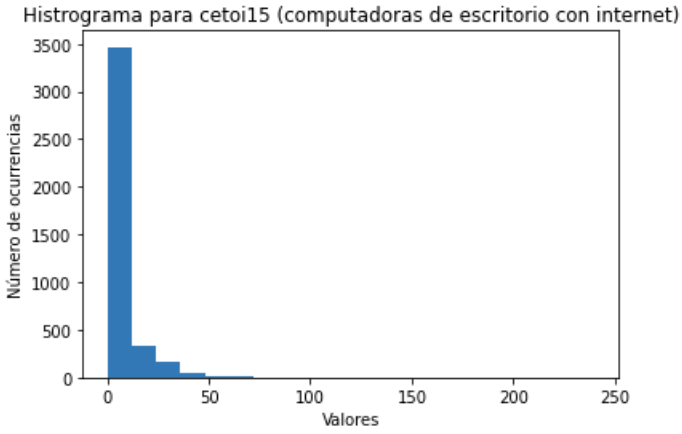
Histogramas

Los siguientes histogramas permiten entender mejor la distribución de algunas de las columnas:









Índice de Desarrollo Social Distrital

En la siguiente tabla se muestra el detalle de cada uno de los campos del conjunto de datos que contiene el índice de desarrollo social distrital:

Columna	Tipo de dato	Descripción
Codigo	Numérico	Código del distrito

Distrito	String	Nombre del distrito
Dimension_Economica	Float	Índice de desarrollo social a nivel económico
Dimension_Participacion	Float	Índice de desarrollo social a nivel de participación
Dimension_Salud	Float	Índice de desarrollo social a nivel de salud
Dimension_Educativa	Float	Índice de desarrollo social a nivel educativo
IDS	Float	Índice de desarrollo social a nivel general

Cantidad de datos de la fuente: 477 filas.

La fuente de datos fue tomada directamente de la página web del Mideplan (https://documentos.mideplan.go.cr/share/s/W-V1r-h_T-eeTH6FVpBu4Q). Este archivo se encuentra en formato xlsx pero se guardó dicha información como archivo csv.

Unión de ambos conjuntos de datos

Ambos conjuntos de datos (escuelas primarias e índice de desarrollo social distrital) se unirán por el código de distrito, específicamente a través de la columna "cddis15" (que contiene el código de distrito) de la base de datos de escuelas primarias, con la columna "Codigo" del conjunto de datos que contiene el índice de desarrollo social distrital.

Objetivo predictivo

Dado que la recomendación es que el objetivo sea de predicción binaria, se agregará una nueva columna al set de datos llama "Promoción_Alta" (Sí o No), la cual indica si la cantidad de estudiantes aprobados es mayor o igual al 95% es alta (valor de columna "Sí"), y si es menor al 95% no es alta (valor de columna "No"). Esta columna se calculará a partir de los datos de las columnas "mit_15" (que contiene la matrícula total para cada centro educativo) y "aprobt_15" (que contiene la cantidad total de aprobados), para obtener el porcentaje de aprobación y a partir de ahí indicar si es ese porcentaje corresponde a una promoción alta (>=95%) o no (< 95%). El umbral de 95% se elige de acuerdo con el set de datos, dado que se tiene aproximadamente un 46% de escuelas con una promoción menor al 95% y 54% con una promoción mayor o igual a 95%.

Por lo tanto, el atributo que se utilizará como variable objetivo del modelo de aprendizaje automático, será esta nueva columna llamada "Promoción_Alta".

La predicción de este atributo se realizará basada en la información que se tiene en el dataset, referente a cada centro educativo como por ejemplo estudiantes reprobados, repitentes, matrícula, etc., además de información del índice de desarrollo social del distrito donde se encuentra ubicado el centro educativo, así como la infraestructura de cada centro educativo (cantidad de aulas, laboratorios, comedor, biblioteca, gimnasio, sodas, computadoras, inodoros, lavatorios, etc.).