Centro de Investigación y de Estudios
Avanzados del Instituto Politécnico
Nacional

UNIDAD ZACATENCO
DEPARTAMENTO DE COMPUTACIÓN

Minería de datos para determinar la calidad educativa de las Escuelas de nivel básico en México

TESIS

Que presenta

Sergio Daniel Romero García

Para obtener el grado de

Maestro en Ciencias en Computación

Directores de la tesis:
Dr. Sergio Víctor Chapa Vergara
M.C Erika Hernández Rubio

Ciudad de México Octubre 2019

Agradecimientos

A mis padres, por todo el apoyo brindado.

A mi hermana, en tu memoria.

A mis hermanos, por su apoyo incondicional.

A Rebeca, por todo lo que me has brindado.

A mis directores, el Dr.Sergio y la M.C Erika, por su tiempo y apoyo durante todo este tiempo.

A mis sinodales, el Dr.Amilcar y el Dr. Juan, por sus grandes aportaciones.

Al personal administrativo del CINVESTAV, Sofía, Érika, Felipa.

Al CONACYT y al CINVESTAV.

Resumen

Se implementará una metodología basada en algoritmos de inteligencia artificial (IA) con información geoestadística y de las escuelas en México, para determinar su nivel de calidad educativa. Para lograr este objetivo se hará uso de la información pública que se tiene, entre las cuales son ENLACE y CEMABE. Con base en toda esta información , se creará una base de datos especializada, en el cual estará la información que va recibir el algoritmo de IA. Esta metodología ayudará a la toma de decisiones, por ejemplo para conocer qué escuela puede tener un mayor crecimiento, si debe recibir un mayor apoyo, entre otros. En el presente trabajo se clasifica el nivel de excelencia educativa usando algoritmos de *clustering*. Estos algoritmos son de aprendizaje no supervisado por lo cual las escuelas se van a clasificar en tres niveles.

Abstract

A methodology based on artificial intelligence (AI) algorithms will be implemented with geostatistical and school information in Mexico, to determine the level of educational quality of the schools. To achieve this goal, we will use public information that is available, among which are ENLANCE and CEMABE. Based on all this information, we will create a specialized database, in which it will be the information that the AI algorithm will receieve. This methodology will help for decision making, for example to know which school can have to increase growth, if it supported, among others. In the present work classifies the level of educational excellence using clustering algorithms. These algorithms are unsupervised learning so schools are going to sort into three levels.

Índice general

Capítulo 1

Introducción

Uno de los principales problemas del sistema educativo en México según datos de la Secretaria de Educación Publica (SEP) es que no garantiza la educación a la mayor parte de los ciudadanos. Ésta es la razón por la que existen comunidades que no reciben la educación básica. El principal objetivo de la educación en México es que debe ser pública, laica, gratuita, de calidad e incluyente. Esto significa no sólo que el Estado deba garantizar el acceso a la escuela a todos los niños, niñas y jóvenes independientemente de su entorno socioeconómico, origen étnico o género, sino que la educación que reciban les proporcione aprendizajes y conocimientos significativos, relevantes y útiles para la vida. En México según datos del Instituto Nacional de Estadística y Geografía (INEGI) del año 2013; 2 de cada 10 estudiantes de nivel básico no cuentan con servicios básicos como agua, drenaje, mientras que alrededor del 16 % no tiene mobiliario básico como pizarrones, bancos.

De acuerdo al Instituto Nacional para la Evaluación de la Educación (INEE) el sistema educativo en México presenta cuatro problemas, acceso, deserción, inequidad y calidad. La falta de acceso a la educación tiene que ver con la parte geográfica y económica que tienen muchas familias en nuestro país, así como de la falta de planteles educativos sobre todo en las zonas rurales; por otro parte la deserción compete otros aspectos que van desde económicos, familiares e incluso del mismo sistema educativo; la inequidad representa la desigualdad de oportunidades para acceder a bienes y servicios como vivienda, educación o salud; finalmente la calidad, según resultados de la prueba PISA (Programme for International Student Assessment) del año 2015 el desempeño de México se encuentra por debajo del promedio en ciencias, lectura y matemáticas. En estas cuatro áreas, menos del 1 % de los estudiantes en México lograron alcanzar niveles de competencia de excelencia.

Se requieren escuelas dignas y funcionales que aseguren el derecho a aprender [Martha E(2016)] hay evidencias claras de las condiciones en que se encuentran muchos planteles, los cuales no están en condiciones de cumplir su propósito.

Por otro lado, el INEE, demandó mayor financiamiento a la educación básica y media superior, para contar con recursos humanos y materiales que garanticen mejores resultados de aprendizaje. De acuerdo al INEGI la escolaridad media de la población de 15 años o más, a nivel nacional, es de 9.2 grados, lo que significa que en promedio la población logra concluir la secundaria.

Otros factores son las desigualdades que se presentan en todo el sistema educativo, por lo que el Estado debe realizar una mayor redistribución de recursos materiales y humanos para que se garanticen en la escuela las mismas oportunidades de acceso, permanencia y aprendizaje para todos los niños y jóvenes del país, en especial a la población y regiones con mayores carencias y rezago educativo [Martha E(2016)]. La infraestructura educativa requiere actualización y mantenimiento, así como la incorporación de elementos que se derivan de los avances tecnológicos que facilitan e impulsan la tarea pedagógica, por lo que no es suficiente contar con la infraestructura necesaria, sino que ésta se actualice y se adecúe a efecto de dignificar las tareas del docente y particularmente el desarrollo de los alumnos en espacios que cuenten con las mejores condiciones.

1.1. Nivel de una escuela

El nivel de una escuela será considerado de calidad si logra sus metas y objetivos previstos [Martha E(2016)] [Gamboa and Bonals(2016)]. Llevado esto al aula, podríamos decir que se alcanza la calidad si el alumno adquiere los conocimientos básicos por ejemplos realizar las operaciones básicas de matemáticas (suma,resta,multiplicación,división). Un segundo punto de vista se refiere a considerar la calidad en términos de relevancia. En este sentido los programas educativos de calidad serán aquellos que incluyan contenidos valiosos y útiles: que respondan a los requerimientos necesarios para formar integralmente al alumno, para preparar excelentes profesionales, acordes con las necesidades sociales, o bien que provean de herramientas valiosas para el trabajo o la integración del individuo a la sociedad. Una tercera perspectiva del concepto de calidad se refiere a los recursos y a los procesos. Un programa de calidad será aquel que cuente con los recursos necesarios y además que los emplee eficientemente. Así, una buena planta física, laboratorios, programas de capacitación docente, un buen sistema académico o administrativo, apropiadas técnicas de enseñanza y suficiente equipo, serán necesarios para el logro de la calidad.

La educación de calidad es uno de los factores más influyentes para el avance y progreso de las personas, sociedades y países, que ha adquirido mayor importancia debido a los acelerados cambios científicos y tecnológicos [Gamboa and Bonals(2016)]. La educación en tema de economía es considerada como uno de los factores más importantes de la producción, en temas sociales como la base para erradicar las desigualdades, la pobreza y el analfabetismo [Gamboa and Bonals(2016)]. La educación es necesaria en todos los sentidos; para mejorar nuestro bienestar social, nuestra calidad de vida, acceder a mejores oportunida-

des de empleo, fortalecer nuestros valores y relaciones sociales. La importancia de la educación radica en ser mejores cada día y aprovechar los recursos que tenemos, es por eso que se ha tratado de medir el nivel de excelencia de una escuela.

En [Frederick.(1987)] identificaron la efectividad y la no efectividad de una escuela. Para ello los autores usaron un enfoque más social, donde el promedio de los alumnos por grado se comparaba con el promedio del grado de todo el país y con base en esto se determinaba si una escuela era efectiva o no.

En [Masters.(2012)], se estudió las dependencias que afectaban el desempeño de los estudiantes, entre cuales están:

- Las prácticas que toma la escuela por ejemplo, si hay colaboración para mejorar la enseñanza, apoyar a los alumnos que lo necesitan, entre otros.

- La influencias sociales como, la localización de la escuela, el lenguaje hablado en casa, etc.

En [Dos.(2013)], se hace un estudio sobre algunos modelos que se propusieron para medir el nivel de una escuela, entre los cuales se encuentran aquellos que miden el nivel de la escuela por medio de la felicidad de sus alumnos dentro y fuera de la escuela, entre otros.

1.2. Aprendizaje máquina

El *aprendizaje automático*, también llamado aprendizaje automatizado o aprendizaje de máquinas (del inglés, "Machine Learning") es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos [Gareth James(2013)].

El proceso de aprendizaje automático es similar al de la minería de datos. Ambos sistemas buscan entre los datos para encontrar patrones. Sin embargo, en lugar de extraer los datos para la comprensión humana como es el caso de las aplicaciones de minería de datos, el aprendizaje automático utiliza esos datos para detectar patrones en los datos y ajustar las acciones del programa en consecuencia.

Los diferentes algoritmos de Aprendizaje Automático se agrupan en una taxonomía en función de la salida de los mismos. Algunos tipos de algoritmos son [Trevor Hastie(2001)].

Aprendizaje supervisado: El algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados anteriores [Shai Shalev Shwartz(2014)].

Aprendizaje no supervisado: Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas [Ghahramani.(2004)].

1.3. Planteamiento del problema

Para lograr el objetivo de este trabajo necesitamos información de las escuelas de nivel básico en México. Como esta información proviene de muchas fuentes (INEGI, SEP, CONAGUA, etc), la cual viene con diferentes formatos, con diferentes estructuras, juntar esta información presenta varios problemas:

1. Los datos provienen de múltiples fuentes, por lo que se necesita manejar grandes volúmenes de información, además se necesita reformatearlos, limpiarlos y cargarlos en otra base de datos, este proceso se conoce como ETL [José Hernandez Orallo(2004)] (por sus siglas en inglés), además los datos se tienen que homogeneizar y estructurar.

2. Análisis de la base de datos que se creó, qué información es mínima y suficiente para medir el nivel de excelencia educativa de una escuela. Debido a que la información proviene de muchas fuentes, se requiere hacer uso de algoritmos de minería de datos como lo es *apriori* [Hegland(2008)], para explotar la información y con ésta se creará un *datamart* [Abraham Sillberschatz(2002)].

4. ¿Cómo se mide el nivel de excelencia educativa de una escuela? Se necesita el indicador del nivel de una escuela como puede ser nivel bajo, medio y alto. Se creará un modelo con base en los indicadores más importantes. Para ello se hará un estudio de los indicadores de las escuelas y se propondrán nuevos indicadores, como por ejemplo, el número de alumnos en promedio por salón de clases, el número de alumnos indígenas que tiene la escuela, si en la escuela se enseña un idioma extra.

5. Debido a que no se cuenta con toda la información de las escuelas de nivel de básico en México, se hará inferencias con los datos que se cuentan. Para ello se hará un estudio de los algoritmos de IA y se propondrá el algoritmo que se va usar; *Linear Discrimant Analysys (LDA), K-Nearest Neighbors (KNN), redes neuronales*[Andreas C.Muller(2017)], entre otros.

1.4. Objetivos

General

Se aplicaran algoritmos de inteligencia artificial, utilizando información geoestadística de las bases de datos de ENLANCE (Evaluación Nacional del Logro Académico en Centros Escolares), CEMABE (Censo de Escuelas, Maestros y Alumnos de Edu- cación Básica y Especial), para determinar el nivel de calidad educativa de las escuelas de nivel básico en México.

Particulares

- Hacer el ETL para el *datamart*.
- Plantear el modelo para la clasificación del nivel de una escuela.
- Aplicar los algoritmos de IA, usando el *datamart*.
- Interpretación de los resultados obtenidos.

1.5. Descripción del documento

Las secciones restantes del documento se han estructurado de la siguiente manera:

Capitulo 2: Se muestra el análisis de datos, los tipos de datos que existen, se explica el proceso ETL y los datos atípicos.

Capitulo 3: Se hace un estudio de los algoritmos que se usaron en el presente trabajo.

Capitulo 4: Se describe las bases de datos que se usaron en el presente trabajo, los lineamentos, que campos tienen, etc.

Capitulo 5: Se describe la metodología usada para clasificar a las escuelas y se muestran los resultados que usaron para ello.

Capitulo 6: Se discuten las conclusiones obtenidas.

Capítulo 2

Análisis de datos

El análisis de datos es la ciencia que se encarga de examinar un conjunto de datos con el propósito de obtener conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas [José Hernandez Orallo(2004)]. En la actualidad, la cantidad de datos que ha sido almacenada en las bases excede nuestra habilidad para reducir y analizarlos sin el uso de técnicas de análisis automatizadas. Muchas bases de datos comerciales, transaccionales y científicas crecen a una proporción exponencial [Migrant and Start(2006)]

El análisis de datos consiste en someter los datos a la realización de operaciones (pruebas de hipótesis, intervalos de confianza, entre otras), esto se hace con la finalidad de obtener conclusiones precisas que nos ayudarán a alcanzar nuestros objetivos, dichas operaciones no pueden definirse previamente ya que la recolección de éstos puede presentar dificultades. Actualmente, muchas industrias usan el análisis para sacar conclusiones y decidir acciones a implementar [Migrant and Start(2006)]. Cabe mencionar que la ciencia también usa el análisis de datos para comprobar o descartar teorías o modelos existentes.

2.1. Extracción, transformación y carga

Un problema habitual al que se enfrentan las organizaciones es cómo recopilar datos de varios orígenes, en varios formatos y moverlos a uno o a varios almacenes de datos. El destino puede no ser el mismo tipo de almacén de datos que el origen y, a menudo, el formato es diferente, o es necesario dar forma o limpiar los datos antes de cargarlos en su destino final. Con los años se han desarrollado varias herramientas, servicios y procesos para ayudarle a afrontar estos desafíos. Independientemente del proceso que se utilice, hay una necesidad común de coordinar el trabajo y aplicar cierto nivel de transformación de los datos.

Extracción, transformación y carga (ETL, por sus siglas en inglés) es una técnica que se utiliza para recopilar datos de varios orígenes, transformarlos según las reglas establecidas y cargarlos en un almacén destino. El trabajo de transformación en ETL tiene lugar en un motor especializado y, a menudo, implica el uso de tablas de almacenamiento temporal para conservar los datos temporalmente a medida que estos se transforman y, finalmente, se cargan en su destino [José Hernandez Orallo(2004)]. La transformación que tiene lugar a menudo conlleva varias operaciones como la calidad, filtrado, ordenación, agregación, combinación, limpieza, quitar duplicados y validación de datos.

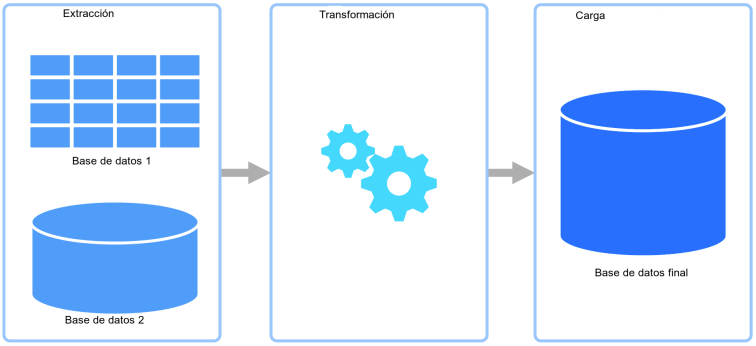


Figura 2.1: Ejemplo del proceso de ETL.

2.1.1. Calidad de los datos

No existen estandarizaciones para determinar la calidad de los datos que se tienen en una base, ni una medida única que define la calidad de los mismo [José Hernandez Orallo(2004)]. Por lo que mantener la exactitud y la calidad de todos los tipos en toda la base es complicado. La limpieza (que se refiere a quitar datos atípicos, repetidos, incorrectos) se requiere para asegurar la calidad de éstos. El primer paso es la identificación de los datos. Para ello hay que imaginar cuáles se necesitan, dónde se pueden encontrar y cómo conseguirlos. Una vez que se disponen de éstos, se deben seleccionar aquellos que sean útiles para los objetivos propuestos, éstos se preparan, poniéndolos en un formato adecuado.

La calidad de datos es un término usado para referirse a la exactitud y fiabilidad de éstos. Los datos deben estar completos, sin variaciones. En las organizaciones donde pueden ser manipulados, identificados y abordados, las posibles fuentes de daño a los datos son un aspecto importante en la calidad de los mismos. Los problemas con la calidad pueden comenzar con una fuente humana. Las personas que los registran pueden cometer errores, lo que lleva a variaciones entre los originales y los almacenados en un sistema. Del mismo modo, las personas pueden cometer errores durante la transferencia o la copia electrónica de datos, haciendo disparidad entre las diferentes versiones o referencias a un archivo.

Figura 2.2: La calidad de los datos evita datos duplicados, faltantes, alterados e incorrectos

Limpieza de datos es el proceso de alterarlos en un almacenamiento para asegurarse de que son exactos y correctos. La limpieza de ellos se centran en la revisión cuidadosa de conjuntos de datos y los protocolos asociados con cualquier tecnología de almacenamiento. La limpieza se compara con la purga de datos, en la que los datos viejos o inútiles se eliminan de un conjunto. Aunque la limpieza puede implicar la eliminación de los datos antiguos, incompletos o duplicados, la limpieza es diferente de la purga, ya que la purga generalmente se centra en limpiar el espacio para nuevos datos, mientras que la limpieza se centra en maximizar la precisión en un sistema. Un método de limpieza puede utilizar el análisis sintáctico u otros métodos para deshacerse de errores de sintaxis, errores tipográficos o fragmentos de registros. Un análisis cuidadoso de un conjunto de datos puede mostrar cómo la fusión de múltiples conjuntos lleva a la duplicación, en cuyo caso la limpieza se puede utilizar para solucionar el problema.

Muchos problemas relacionados con la limpieza de datos son similares a los problemas que tienen los archivistas, el personal administrativo de la base de datos y otros en torno a procesos como el mantenimiento de los mismos, minería de datos orientada y la metodología de extracción, transformación y carga (ETL), donde los datos viejos se vuelven a cargar en un nuevo conjunto. Estos problemas suelen considerar la sintaxis y el uso específico de comandos para realizar tareas relacionadas en la base de datos. La administración de las bases es un papel muy importante en muchas empresas y organizaciones que dependen de grandes conjuntos de datos y registros precisos para el comercio o cualquier otra iniciativa.

Al crear bases de datos, se debe prestar atención a la calidad de los mismos y a cómo mantenerlos. Una buena base de datos hará cumplir la integridad siempre que sea posible [José Hernandez Orallo(2004)]. Por ejemplo, un usuario podría accidentalmente intentar ingresar un número de teléfono en un campo de fecha. Si el sistema aplica calidad, detectará que el usuario cometió estos errores. Mantener la calidad significa asegurarse de que los datos permanezcan intactos y sin cambios a lo largo de todo su ciclo de vida. Esto incluye la captura de los

mismos, el almacenamiento, las actualizaciones, las transferencias, las copias de seguridad, etc. Cada vez que se procesan existe el riesgo de que se corrompan (accidental o maliciosamente).

Para que la integridad de los datos se mantenga, es necesario que no haya habido cambios o alteraciones en éstos. A medida que los datos son introducidos, almacenados, accedidos, movidos y actualizados, los puntos débiles en un sistema pueden comprometerlos. Fallas en una computadora pueden llevar a sobreescribirlos parcialmente. Las interrupciones en las diversas operaciones pueden dar lugar a problemas, como daños mecánicos, como la exposición a imanes o daños físicos causados por cortes de energía u otros eventos.

2.2. Tipos de datos

Cada vez que trabajemos con una base de datos, debemos definir los tipos de datos que con mayor eficiencia puedan almacenarlos. En los campos de las tablas que tiene nuestra base en general hay tres grandes tipos de contenidos [José Hernandez Orallo(2004)]:

- Datos cuantitativos
- Datos cualitativos
- Datos para almacenar fechas y horas.

Un dato nos permite describir un objeto [Migrant and Start(2006)]. Dicho objeto podemos llamarlo entidad, por ejemplo una casa en la que viven personas. La casa es la entidad y la cantidad de personas que viven en la casa son un dato, que en este caso es cuantitativo. Un dato por sí sólo no puede demostrar demasiado, siempre se evalúa el conjunto para poder examinar los resultados. Para examinarlos, primero hay que organizarlos o tabularlos [Migrant and Start(2006)].

2.2.1. Datos cuantitativos

Todo lo que se puede medir y contar, decimos que se puede cuantificar [Migrant and Start(2006)]. El concepto datos cuantitativos hace referencia precisamente a eso, a la información tangible. Éstos se refiere al flujo constante de valores posibles de la variable, estos datos no se restringen a valores enteros, ejemplo:

- Medir la altura de una persona. (Puedes mediar la altura en metros, centímetros y hasta dar una medida en milímetros, es decir, los datos son continuos).
- Edad (Puedes definir una edad en años, meses y hasta días)
- Medir el número de alumnos que tiene una escuela (LA variable sólo puede tomar valores enteros)

Hay dos tipos de datos cuantitativos, que también se conocen como datos numéricos: continuo y discreto [Migrant and Start(2006)]. Como regla general, los recuentos son discretos y las mediciones son continuas. Los datos discretos son un conteo que no se puede hacer más preciso. Por lo general, implica números enteros. Por ejemplo, el número de niños (o adultos, o mascotas) en su familia es información discreta, porque está contando entidades enteras e indivisibles: no puede tener 2.5 hijos o 1.3 mascotas. Los datos continuos, por otro lado, podrían dividirse y reducirse a niveles cada vez más finos. Por ejemplo, puede medir la altura de sus hijos en escalas progresivamente más precisas (metros, centímetros, milímetros y más), por lo que la altura es un dato continuo.

La mayoría de las veces se necesita analizar estos valores en un grupo de individuos, para ello, estas variables también se expresan en el valor promedio de los datos que conforman el grupo [Migrant and Start(2006)]. El análisis de datos cuantitativos es fundamental en la toma de decisiones basadas en la investigación, para ello se recurre a un análisis de comparaciones numéricas y estadísticas. Los datos cuantitativos nos dan las bases, el soporte y sobre todo confiabilidad a nuestra investigación. Los podemos ubicar por categorías, darle un orden de acuerdo a su importancia o unidad de medida. Si vamos a medir, entonces vamos a cuantificar, pero existen también métodos de investigación que nos dan la posibilidad de obtener datos cuantitativos como cualitativos, por ejemplo observaciones de campo, sondeos y las encuestas o preguntas donde los encuestados tengan como opción para contestar “SI” o “NO” [José Hernandez Orallo(2004)].

2.2.2. Datos cualitativos

Los datos cualitativos se expresan en forma de palabras o textos que ayudan a comprender ciertas acciones. Los enfoques cualitativos proporcionan información contextual, en profundidad sobre los “por qué” y “cómo” [José Hernandez Orallo(2004)]. La información cualitativa complementa y proporciona una mayor percepción de los datos cuantitativos. Por lo general, el análisis de datos cualitativos requieren más tiempo para procesar y ordenar los datos que para su recolección.

Encontramos distintas técnicas para analizar palabras o frases, una labor que comienza en realidad con la recolección, estos datos recolectados se analizan a partir de diferentes técnicas de análisis. Se trata de una actividad compleja [José Hernandez Orallo(2004)], cuyo fin último es dotar a los datos de sentido. Para ello, se utilizan procedimientos variados muy diversos, que raramente son estadísticos, por ejemplo podríamos pasar las variables cualitativas a cuantitativas. No en vano, la analítica de datos cualitativos se considera más un arte que una técnica [de Smith(2018)]. Su elección dependerá de nuestro objetivo, así como de las tareas y operaciones más adecuados. El proceso general de este tipo de análisis comienza por una recopilación selectiva de los datos, seguida de una reducción para su identificación, clasificación, síntesis y agrupamiento. Una vez que la información haya sido recolectada y ordenada, la codificaremos

para poder empezar a llegar a conclusiones una vez integremos la información. El proceso de codificación agrupa en categorías, temas o conceptos, con el objetivo de relacionarlos con el fin de la consulta o investigación, la codificación las dotará de sentido. Finalmente, se analizan los datos para alcanzar conclusiones que también deben verificarse.

2.3. Transformación de los datos

La transformación de datos es el proceso de convertir datos de un formato o estructura a otro formato o estructura [José Hernandez Orallo(2004)]. La transformación de datos es crítica para actividades como la integración y la gestión de los mismos. Es decir; se transforman los datos, como agregar datos de ventas o convertir formatos de fecha, editar cadenas de texto, unir filas y columnas, pasar datos cualitativos a cuantitativos y viceversa.

2.3.1. Transformación de datos cualitativos

Los datos cuantitativos para un mejor estudio de ellos pueden ser clasificados de manera cualitativa. La razón principal suele ser el intento de simplificar la interpretación de la variable en cuestión, de tal manera que la clasificación en categorías facilite la toma de decisiones, por ejemplo a la hora de solicitar pruebas complementarias. La conversión de una variable cuantitativa en cualitativa se denomina categorización.

Un procedimiento para la categorización se basa en escoger los valores de los cuartiles o de percentiles específicos de la distribución de los datos en nuestro estudio. Este método se suele utilizar para fijar intervalos de referencia de pruebas analíticas a partir de una muestra representativa de la población, eligiéndose dos percentiles centrados en torno a la mediana de la distribución, concretamente los valores 2.5 y 97.5, que definen por tanto un intervalo de referencia del 95 %.

Otro procedimiento es el calcular la media μ y la desviación estándar σ de los datos, posteriormente poner los datos en intervalos de tipo $I_0 = [\mu - \sigma, \mu]$, $I_1 = [\mu, \mu + \sigma]$, definiendo el número de intervalos que se quiera, ejemplo:

Nombre	Peso
Persona 1	75 KG
Persona 2	65 KG
Persona 3	80 KG
Persona 4	82 KG
Persona 5	102 KG
Persona 6	76 KG
Persona 7	55 KG
Persona 8	70 KG
Persona 9	73 KG
Persona 10	78 KG

Tabla 2.1: Tabla que muestra el peso en kilogramos de diez personas

De la tabla 2.1 tenemos que $\mu = 75{,}6$ y $\sigma = 11{,}56$, por lo que definimos 6 intervalos de la siguiente manera:

I_i	Intervalo
$I_0 = [\mu - 3 * \sigma, \mu - 2 * \sigma)$	I_0 =[40.92,52.48)
$I_1 = [\mu - 2 * \sigma, \mu - \sigma)$	I_1 =[52.48,64.04)
$I_2 = [\mu - \sigma, \mu)$	I_2 =[64.04,75.6)
$I_3 = [\mu, \mu + \sigma)$	I_3 =[75.6,87.16)
$I_4 = [\mu + \sigma, \mu + 2 * \sigma)$	I_4 =[87.16,98.72)
$I_5 = [\mu + 2 * \sigma, \mu + 3 * \sigma)$	I_5 =[98.72,110.28)

Tabla 2.2: Tabla que muestra los intervalos I_i de la tabla 2.1

Así transformando los datos de la tabla 2.1 con los intervalos de la tabla 2.2, obtenemos:

Nombre	Intervalo
Persona 1	I_2
Persona 2	I_2
Persona 3	I_3
Persona 4	I_3
Persona 5	I_5
Persona 6	I_3
Persona 7	I_1
Persona 8	I_2
Persona 9	I_2
Persona 10	I_3

Tabla 2.3: Tabla que muestra la transformación de la tabla 2.1

Así categorizamos los datos; por otro lado los intervalos de los extremos para nuestro ejemplo podrían ser de la forma, $I_0 = (-\infty, \mu - 2*\sigma]$ y $I_5 = [\mu + 2*\sigma, \infty)$, debido a que puede haber datos que estén más alejados de la media que tres desviaciones estándar, además podemos refinar más los intervalos por ejemplo teniendo intervalos de la forma $I_i = \pm i(0{,}5)(\sigma)\ i = 1, 2, 3, 4, \cdots, n$.

2.3.2. Transformación de datos cuantitativos

Los datos cuantitativos, son usados para nombrar o categorizar información. Este tipo de dato se caracteriza por no ser ordenado, incluso si se usan números para representarlos. Por ejemplo el nombre de las diferentes tipos de cancer que sufren los pacientes en un hospital es un dato cuantitativo. Aunque puedes ordenar todos los nombres alfabéticamente, carece de sentido conceptual que cáncer de "Colon" se encuentre antes que "Estomago" o que "Estomago" se encuentre después de "Hueso". Esto no nos ayuda a entender sobre las características de los pacientes que tienen cáncer.

Múltiples métodos de análisis de datos requieren valores numéricos como entrada, por lo que es necesario transformar variables de tipo categórico, la manera más sencilla de transformar estos datos es crear variables *dummy* (falsas, en español). Crear variables *dummy* implica transformar datos de un formato "alto", en el que cada columna contiene la información de una variable, a datos con un formato "ancho", en los que múltiples columnas contienen la información de las dos variables, codificada de manera binaria, esto es, con 0 y 1, por ejemplo:

Paciente	Tipo de cáncer
Paciente 1	Colon
Paciente 2	Colon
Paciente 3	Estomago
Paciente 4	Hueso
Paciente 5	Colon
Paciente 6	Hueso
Paciente 7	Estomago
Paciente 8	Estomago
Paciente 9	Hueso
Paciente 10	Colon

Tabla 2.4: Tabla que muestra el tipo de cáncer para una muestra de 10 pacientes

Paciente	cáncer.Colon	cáncer.Estomago	cáncer.Hueso
Paciente 1	1	0	0
Paciente 2	1	0	0
Paciente 3	0	1	0
Paciente 4	0	0	1
Paciente 5	1	0	0
Paciente 6	0	0	1
Paciente 7	0	1	0
Paciente 8	0	1	0
Paciente 9	0	0	1
Paciente 10	1	0	0

Tabla 2.5: Tabla que muestra la obtención de las variables dummies de la tabla 2.4

Como podemos observar se elimina la columna de "Tipo de cáncer"de nuestros datos y se crean columnas adicionales, en las cuales en cada una de ellas se coloca el tipo de cáncer que tiene el paciente.

2.4. Datos atípicos

Un valor atípico es un dato que es considerablemente diferente a los otros de la muestra [Vargas(2015)]. Con frecuencia, los valores atípicos en un conjunto de datos pueden producir anormalidades experimentales o errores en las mediciones tomadas, y debido a esto puede que los descarten del conjunto de datos. Si los valores atípicos del conjunto se ignoran, puede haber cambios importantes en las conclusiones obtenidas del estudio. Por eso, saber cómo calcular y evaluar los valores atípicos es importante para asegurar la comprensión apropiada de los datos.

Los valores atípicos son datos que son muy diferentes a la tendencia expresada por los otros valores del conjunto de datos [Vargas(2015)]. Es decir, se ubican distantes a los otros valores. Generalmente es muy difícil detectar esto en las

tablas de datos si es conjunto muy grande. Si este conjunto se expresa gráfico, los valores atípicos se ubican "distante" a los otros valores. Si, por ejemplo, la mayoría de los datos en un conjunto de datos formaran una línea recta, no se podría interpretar razonablemente que los valores atípicos fueran parte de esa línea, por ejemplo:

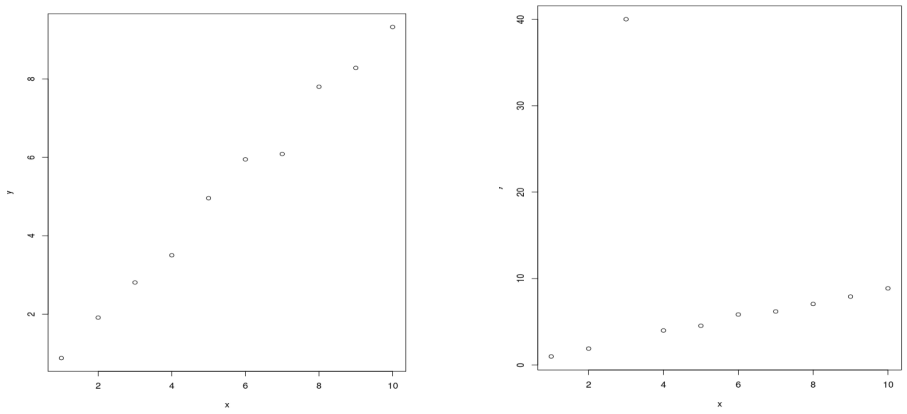


Figura 2.3: La parte izquierda muestra una simulación de 10 datos, la parte derecha muestra la misma simulación ahora introduciendo una dato atípico

Dato	Valor
1	0.8777
2	1.9107
4	3.5008
5	4.9578
6	5.9491
7	6.0856
8	7.8001
9	8.2818
10	9.3249
3	40

Tabla 2.6: Tabla que muestra los datos de menor a mayor de la parte derecha de la imagen 2.3

De la figura 2.3 tenemos un conjunto de 10 datos simulados. De la parte derecha vemos que todos los datos menos uno tienen un comportamiento de recta, al ver esta gráfica uno puede indicar que el valor en el eje y de 40 probablemente sea un valor atípico.

El método para calcular los datos atípicos en este trabajo es el algoritmo de Tukey [Vargas(2015)] que incorpora el uso de la media truncada y recorta la base inicial de trabajo en un 10 %, predefiniendo como atípicos todos los registros allí contenidos. El método de cálculo de los límites superior e inferior implica:

a) Ordenar los datos de mayor a menor, señalizando aquellos registros localizados en el 5 % superior y 5 % inferior.

b) La base de trabajo para la determinación del resto de atípicos se constituye por el restante 90 % de registros.

c) Determinación de dos grupos de trabajo para la misma base: primera y segunda parte, definidas por medio de la media aritmética.

d) Cálculo de la media aritmética para cada uno de los grupos, medias de las mitades, media superior y media inferior o medias truncadas.

e) Cálculo de los limites superior e inferior a partir de:
Límite superior: Media + 2,5 (media superior -media)
Límite inferior: Media -2,15 (media -media inferior)

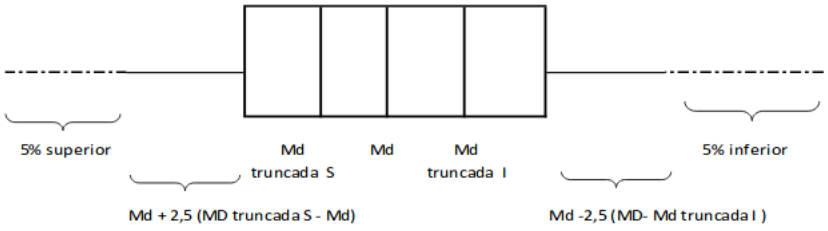


Figura 2.4: Representación gráfica del análisis de Tukey

La figura 2.4 representa: Los datos atípicos a revisar corresponden tanto al 5 % superior e inferior de la base, como aquellos que superan los límites superiores e inferiores. La media truncada de la parte superior de los datos se reconoce como la media truncada superior (S), en tanto que su contraparte responde como media truncada inferior (I).

Capítulo 3

Aprendizaje no supervisado

El aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones, a diferencia del aprendizaje supervisado en cual se tiene observaciones y soluciones a dichas observaciones, un ejemplo de aprendizaje supervisado puede ser: determinar las ventas esperadas en un año de cierta empresa con base a la información de lo que se ha consumido a lo largo de los años anteriores [Trevor Hastie(2001)] [Gareth James(2013)].

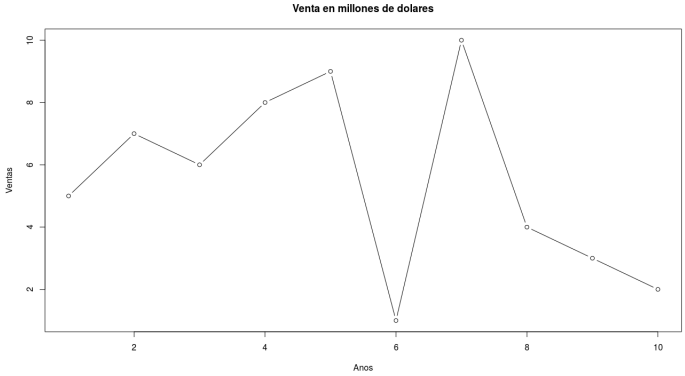


Figura 3.1: Ejemplo de aprendizaje supervisado: se quiere predecir las ventas del año once con base en los años anteriores.

Por otro lado el aprendizaje no supervisado se distingue por el hecho de que no hay un conocimiento a priori, por ejemplo, queremos saber la clase social de las personas (baja, media, alta), para ello sabemos sus ingresos percibidos anualmente y también se sabe sus egresos percibidos anualmente, pero no sabemos a que clase social pertenecen.

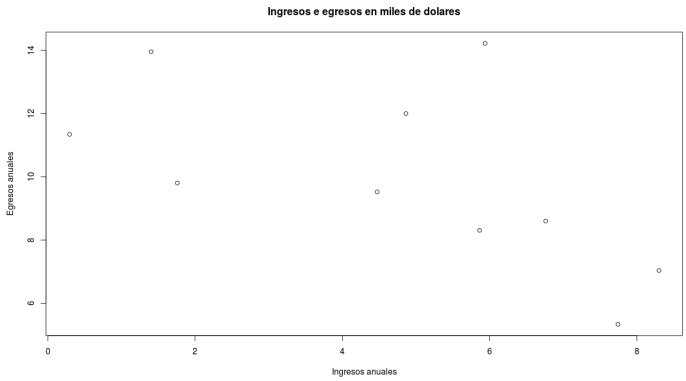


Figura 3.2: El eje x es el ingreso anual en miles de dolares, el eje y es el egreso anual en miles de dolares, por lo que cada punto representa una persona, note que no hay conocimiento de la clase social de la persona

Así entonces el aprendizaje supervisado permite: buscar patrones en datos históricos. Por otro lado, el aprendizaje no supervisado usa datos históricos que no están etiquetados. El fin es explorarlos para encontrar alguna estructura o forma de organizarlos.

Los tipos de algoritmo más habituales en aprendizaje no supervisado son:

- Algoritmos de clasificación
- Análisis de componentes principales
- Descomposición en valores singulares
- Análisis de componentes independientes

3.1. Clasificación no supervisada

Un algoritmo de clasificación no supervisada (mejor conocido como *clustering*) es un procedimiento de agrupación acuerdo con un criterio, por ejemplo queremos agrupar personas con respecto a su peso, entonces el criterio es el peso de la persona. Estos criterios son por lo general distancia o similitud o alguna otra característica que comparta la población a clasificar. La cercanía se define en términos de una determinada función de distancia, como la euclidiana.

El *clustering* se refiere a técnicas para encontrar subgrupos, o *clusters*, en un conjunto de datos. Cuando clasificamos usando una técnica de *clustering*, buscamos las particiones del conjunto de datos, tal que las observaciones en cada subgrupo sean bastante similares una de la otra. Donde se define que una observación es similar de la otra con base en la distancia Euclidiana.

Por ejemplo, supongamos que tenemos n observaciones, con p características cada una, donde los datos corresponde a muestras de pacientes con cáncer de pulmón, entonces lo que queremos saber es si ¿Hay tipos de cáncer de pulmón que no sean conocidos?, el *clustering* nos puede ayudar a responder esa pregunta.

Otra aplicación es la publicidad. Supongamos que tenemos acceso a un gran numero de datos de los hogares en México, entre los cuales están el salario promedio del hogar, la distancia más cercana a una área urbana, el número de niños que hay en el hogar, etc. Nuestra meta es hacer una segmentación de publicidad identificando los subgrupos de hogares que sean más receptivos al comprar un producto en particular.

Como el *clustering* es muy popular en muchos campos, existen un gran número de metodos de *clustering*, los más conocidos de estos metodos son $K-means$ *clustering* y *hierarchical clustering*. La descripción de los algoritmos se tomaron de [Trevor Hastie(2001)] y [Gareth James(2013)].

3.1.1. K-medias *clustering*

Es un algoritmo de particionamiento de un conjunto de datos en K subconjuntos (llamados *clusters*), donde en estos *clusters* no hay intersección. Para empezar este algoritmo, primero tenemos que especificar el número de K *clusters*, después el algoritmo asignará aleatoriamente a cada una de las observaciones en uno de los K *clusters*.

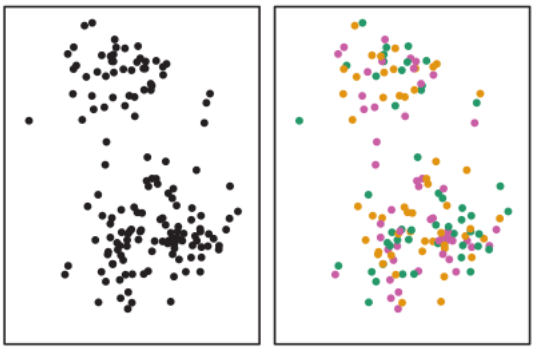
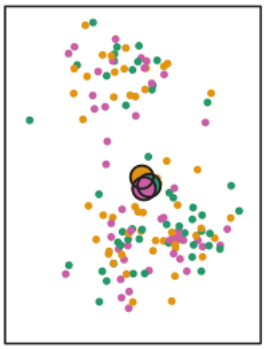
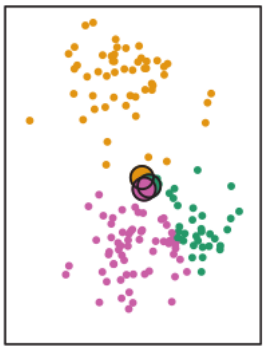


Figura 3.3: Un conjunto de datos simulados con 150 observaciones en dos dimensiones. Se muestran los primeros pasos del algoritmo con $K = 3$. De [Hastie and Tibshirani(2016)]

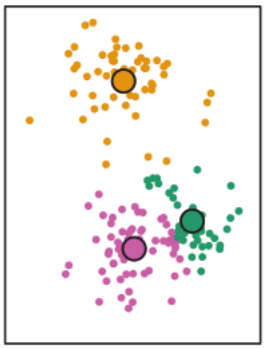
Después para cada *cluster* K, se va a calcular el centroide.



Posteriormente se asigna cada observación al *cluster* con el centroide más cercano.



Se repite este proceso hasta que ya no hay cambios en los *clusters*.



El algoritmo debe resolver un problema de optimización.

$$min \left(\sum_{i=1}^{k} \sum_{x_j \in C_i} || \, x_j - u_i \, || \right)$$

Algoritmo 1: $K - Medias\ clustering$

1. Aleatoriamente asignar un número de 1 a K a cada una de las observaciones.
2. Iterar hasta que no haya cambios en los *clusters*:
   (a) Para cada K *cluster*, calcular el centroide.
   (b) Asignar a cada observación al *cluster* con el centroide más cercano.

Donde x_j representa el dato j de nuestro conjunto de datos y u_i la media del *cluster* i y C_i representa el *cluster* i, así $K - medias$ encuentra una solución optima local en vez de una solución global, por lo que los resultados dependen de la selección aleatoria del paso 1 en el Algoritmo 1, por eso es importante ejecutar el algoritmo muchas veces para cambiar la aleatoriedad. Entonces la mejor solución es la que da la cantidad más pequeña. Esta cantidad representa la suma de la sumas de las distancias de cada observación a su *cluster* asignado. La figura 3.4 muestra la solución local de seis diferentes ejecuciones del algoritmo para la misma simulación, en este caso la mejor solución es el *clustering* con valor de 235.8

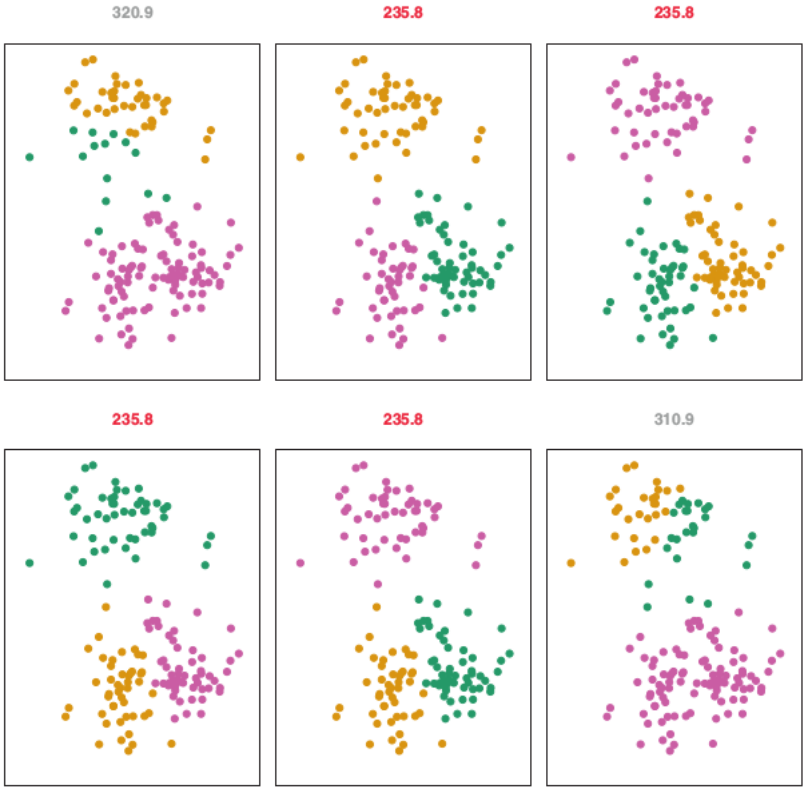


Figura 3.4: $K-Medias$ *clustering* se ejecutó seis veces en los datos de la figura 3.3 con $K = 3$, para cada diferente ejecución obtenemos un número diferente, entonces la solución optima es la que tiene la menor cantidad. En este caso hay tres *clustering* que dieron la solucion más pequeña que es de 325.8. De [Hastie and Tibshirani(2016)]

Como podemos ver, para ejecutar el algoritmo necesitamos saber cuantos *clusters* esperamos que haya en los datos. El problema de seleccionar K no es trivial.

3.1.2. *Clustering* jerárquico

Los resultados de aplicar el algoritmo $K-medias$ depende de la elección del número K. Por otro lado, el *clustering* jerárquico no requiere ninguna especificación del número de clases, este algoritmo requiere una medida de similitud entre las observaciones, basada en las características que comparten las observaciones a pares como lo es la distancia entre puntos. Como el nombre sugiere la representación del algoritmo produce *clusters* a cada nivel creado por la similitud de las observaciones, en el nivel más bajo cada observación es un *cluster* y en el nivel más alto hay un solo *cluster* en donde todas la observaciones fueron asignadas a

este *cluster*. Otra ventaja que tiene el algoritmo jerárquico sobre $K-medias$ es la representación basada en un árbol de las observaciones llamada dendrograma.

Tenemos el dendrograma de la figura 3.5, en la cual tenemos 45 observaciones, en donde cada hoja del árbol representa una observación, algunas ramas del árbol se fusionan, ésto representa que las observaciones de cada rama son similares (con base en la medida de similitud definida). La fusión de las ramas de la parte superior del árbol no son muy similares entre ellas. La altura de la fusión es medida en el eje vertical e indica la diferencia entre las observaciones. Así las observaciones que se fusionan en los niveles más bajos del árbol, son bastantes similares y las observaciones que se fusionan en la parte superior del árbol tiende a ser bastantes diferentes.

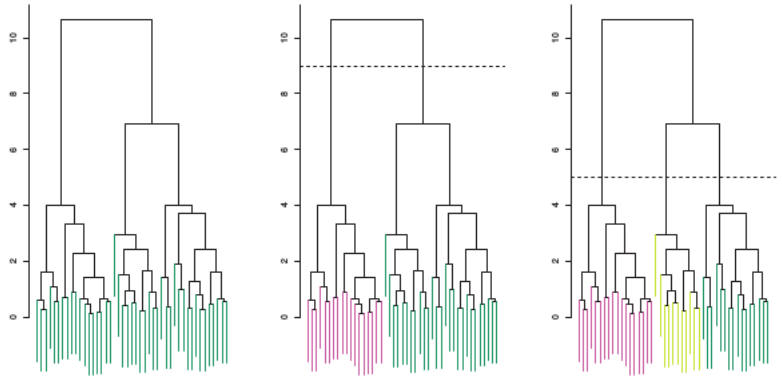


Figura 3.5: El dendrograma de la izquierda representa un solo *cluster*, el dendrograma del centro se corta a la altura de nueve (indicada por la linea), este corte da como resultado dos *clusters* mostrados en diferente color, el dendrograma derecho se corta a la altura de cinco lo que produce tres *clusters* diferentes. De [Hastie and Tibshirani(2016)]

Estas fusiones son muy importantes en un dendrograma, pero pueden ser fácilmente mal interpretadas, considere la parte izquierda de la figura 3.6 en la cual se muestra un dendrograma de nueve observaciones. Podemos observar que la observación 5 y 7 son bastante similares, esto se debe a que la fusión de éstas se produce en el nivel más bajo del árbol, por ejemplo la observación 1 y 6 también son bastante similares. Sin embargo se puede cometer el error de argumentar que la observación 9 y 2 son bastante similares lo cual es un error. Al ver la parte derecha de la figura 3.6 observamos que la observación 9 no es más similar a la observación 2, de lo que las observaciones 8, 5 y 7 son similares entre ellas.

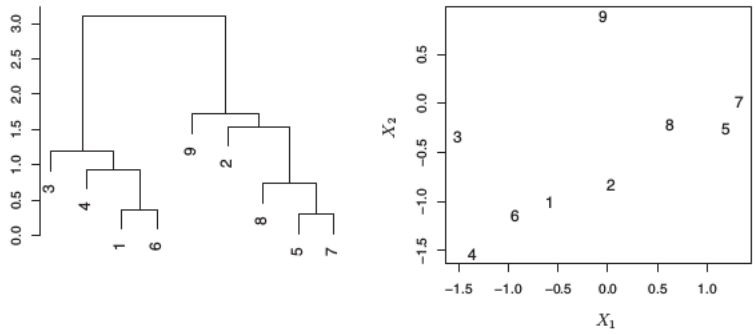


Figura 3.6: La parte izquierda representa un dendrograma usando la distancia euclidiana como medida de similitud. Las observaciones 5 y 7 son bastante similares, sin embargo la observación 9 no es más similar a la observación 2, de lo que las observaciones 8, 5 y 7 son similares entre ellas, esto se debe a que las observaciones 2, 8, 5 y 7 se fusionan con la observación 9 en la misma altura (aproximadamente 1.8). En la parte derecha se grafican los datos que usaron para crear el dendrograma, de aquí se puede observar que la observación 9 no es muy similar a la observación 2. De [Hastie and Tibshirani(2016)]

Para identificar el número de *clusters* que hay en un dendrograma, tenemos hacer un corte horizontal a través del dendrograma, por ejemplo en la figura 3.5 hay tres cortes, la parte izquierda: el corte se hizo a una altura mayor de la última fusión (aproximadamente en 11) dando como resultado un solo *cluster*, la parte de en medio: el corte se hizo a una altura aproximadamente de 9 dando como resultado dos *cluster* y la parte derecha haciendo un corte a la altura de aproximadamente 5 dando como resultado tres *clusters*. En otras palabras, la altura a la que se realice el corte en el dendrograma, tiene la misma función que el número K en el algoritmo de $K - medias$.

El dendrograma jerárquico se obtiene por un algoritmo no muy complicado. Empezamos por definir la medida de similitud entre las observaciones, para alcances de este trabajo solo se considera la distancia euclidiana. El algoritmo empieza en la parte más baja del dendrograma, en donde cada observación es su propio *cluster*. Las dos observaciones más similares se fusionan en una sola, el resultado de esta fusión es la observación que más lejos está del origen, así al final del primer paso tenemos $n-1$ *clusters*. Posteriormente se repite el proceso, se fusionan las observaciones que son más similares entre ellas, así que al final tenemos $n-2$ *clusters*, el algoritmo continua de esta manera hasta que solo haya un *cluster*. En la figura 3.7 se muestra los primero pasos del algoritmo.

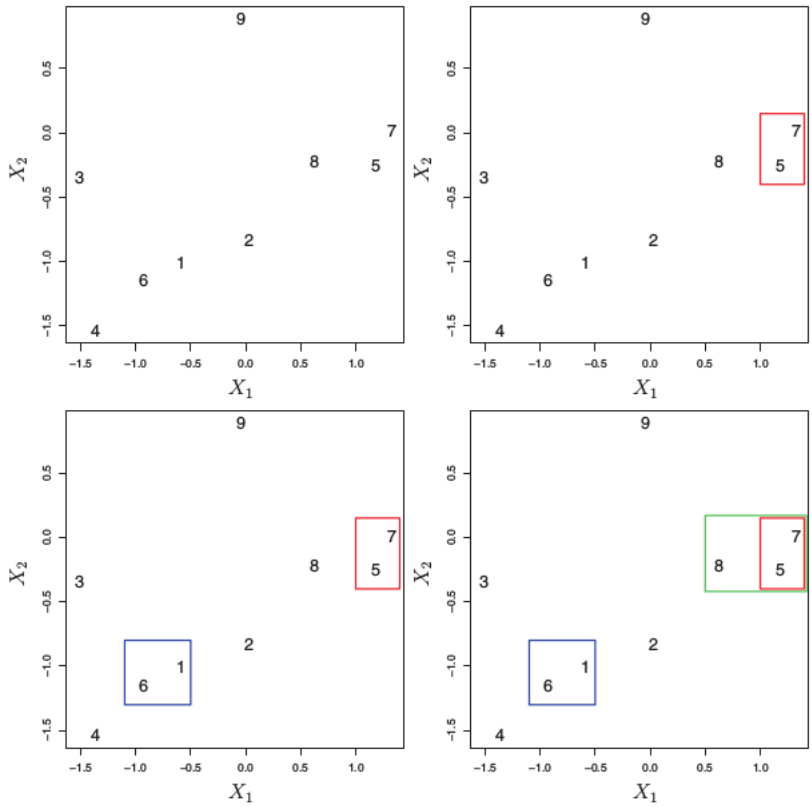


Figura 3.7: Una ilustración de los primeros pasos del algoritmo para crear el dendrograma jerárquico. La parte superior izquierda: hay nueve observaciones distintas. La parte superior derecha: los *clusters* que son más similares se fusionan en un solo *cluster*. La parte inferior izquierda: los dos siguientes *clusters* más similares se fusionan en un solo. La parte inferior derecha: los dos siguientes *clusters* que son más similares se fusionan en un solo *cluster*. De [Hastie and Tibshirani(2016)]

3.2. Análisis de componentes principales

Cuando tenemos una gran conjunto de variables, el análisis de componentes principales (PCA por su siglas en inglés) nos puede ayudar para reducir el tamaño de dichas variables, dicha reducción se conoce como reducción de la dimensionalidad. Esta reducción se calcula como las direcciones de las componentes principales. Estas direcciones se definen como líneas en un subespacio en donde estas lineas son lo más cercanas como es posible a los datos.

Algoritmo 2: *clustering* Jerárquico

1. Empezamos con n observaciones y como medida de similitud la distancia euclidiana. Calculamos las $\binom{n}{2}$ distancias de las observaciones para medir la similitud, tratamos cada observación como un *cluster*.
2. Para $i = n, n-1, \ldots, 2$:
   (a) Identificamos las dos *clusters* que son más similares (con la distancia euclidiana más pequeña). Fusionamos esos dos *clusters* (nos quedamos con el *cluster* que este más lejos del origen).
   (b) Calculamos la distancia euclidiana entre los $i-1$ *clusters* restantes.

Supongamos que queremos visualizar n observaciones con p variables cada observación, al ser un espacio de p variables solo se podría visualizar para $p < 3$, por lo que podríamos examinar los datos usando una gráfica de dos variables, sin embargo hay $\binom{p}{2}$ posibles gráficas por lo que si $p = 15$ hay, 105 posibles gráficas que se pueden formar. Así si p es grande, entonces no es posible visualizar los datos, por otro lado cada una de estas gráficas tiene solo una pequeña parte del total de la información. Por lo que se necesita un mejor método para tratar los datos cuando p es grande, en general nos gustaría una reducción de la dimensionalidad en la que esta reducción represente de manera correcta a los datos, por ejemplo queremos obtener una representación de dos dimensiones, en donde esta representación capture la mayoría de la información, entonces podríamos graficar las observaciones de p variables en un espacio de dos dimensiones.

PCA nos proporciona un método para reducir la dimensión, este método encuentra una representación de los datos en un espacio de menor dimensión. Idealmente para cada observación n en un espacio de dimensión p, hay dimensiones que no son muy interesantes, así PCA busca un número pequeño de dimensiones en las cuales todas estas dimensiones son interesantes, el concepto de interesante se mide como la cantidad de observaciones que varía a lo largo de cada dimensión. Cada dimensión encontrada por PCA es una combinación lineal de las p variables.

La primera componente principal de un conjunto de variables $X_1, X_2, \ldots, X_p$ es combinación lineal normalizada de las variables así:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Donde $\phi_{11}, \ldots, \phi_{p1}$ son los valores de la primer componente principal.

Dado un conjunto de datos X de tamaño $n \times p$ estamos interesados en la varianza de las componentes principales, así observamos la combinación lineal de los datos de la forma:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

Donde queremos que esta combinación tenga la varianza más grande sujeta a $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. En otras palabras, la primer componente principal debe de resolver el siguiente problema de optimización:

$$maximizar \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ sujeto a } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Para calcular la segunda componente principal, ésta debe de ser una combinación lineal de $X_1, \ldots, X_p$ que tenga la máxima varianza y que no este correlacionada con Z_1. Así la segunda componente principal toma la forma.

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip}$$

Para las siguientes componentes principales, se sigue el mismo proceso; se calcula la combinación lineal normalizada de los datos que tenga la máxima varianza sujeta a $\sum_{j=1}^{p} \phi_{ji}^2 = 1$ donde $i = 1, \ldots, n$ y que no este correlacionada con las anteriores componentes principales obtenidas.

Nos podemos preguntar cuanta información aportan las componentes principales del total de los datos, entonces nos preguntamos cuanta varianza tiene las componentes principales con el total de los datos. En general estamos interesados en saber la proporción de la varianza que aporta cada una de ellas.

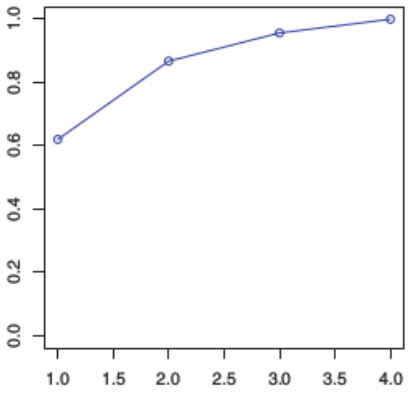


Figura 3.8: Se muestra la proporción de la varianza acumulada que aporta cada componente principal de tamaño cuatro con datos simulados. De [Hastie and Tibshirani(2016)]

En la figura 3.8 se puede concluir que las dos primeras componentes principales son las que más aportan, debido que la primer componente aporta aproximadamente 62,0 % de la varianza total, y la segunda aporta 13 %.

En general una matriz X de datos de tamaño $n \times p$ tiene $n-1$ ó p (se elige la que sea menor) distintas componentes principales, por lo que no estamos interesados en todas ellas, idealmente estamos interesados en unas pocas que nos proporcionen la información suficiente de los datos totales, para resolver este problema podemos escoger el número de componentes principales usando la proporción de la varianza acumulada que aportan, es decir quedarnos con las componentes que aporten hasta que cierto umbral que nosotros definamos sea superado, por ejemplo en la figura 3.8 el umbral podría ser 80 %, así nos quedaremos con las dos primeras componentes principales.

3.3. Apriori

Apriori es un algoritmo que sirve para la búsqueda de reglas de asociación en bases de datos que comparten elementos en común, por ejemplo tenemos una base datos en la cual está la información de los crímenes cometidos en cierta ciudad, por lo que uno se podría preguntar que elementos o que características comparten los crímenes de robo. Aprori consta de dos etapas las cuales son:

- Identificar todos las variables (también llamado item) que ocurren con una frecuencia por encima de un umbral que definimos.
- Convertir estos variables en reglas de asociación.

Siguiendo con el ejemplo anterior, en la base de crímenes nos quedamos con aquellos que solo hayan sido crimen de robo. En este caso, cada crimen de robo cubre cuatro características distintas $\{C_1\}, \{C_2\}, \{C_3\}, \{C_4\}$.

Crimen	Características
Crimen 1	$\{C_1, C_2, C_3, C_4\}$
Crimen 2	$\{C_1, C_2, C_4\}$
Crimen 3	$\{C_1, C_2\}$
Crimen 4	$\{C_2, C_3, C_4\}$
Crimen 5	$\{C_2, C_3\}$
Crimen 6	$\{C_3, C_4\}$
Crimen 7	$\{C_2, C_4\}$

Tabla 3.1: Tabla que muestra las características que tienen los crímenes de robos

Antes de entrar en los detalles del algoritmo, vamos a definir una serie de conceptos:

- Soporte: El soporte del item X es el número de características que contienen X dividido entre el total de características.

- Confianza: La confianza de una regla "Si X, entonces Y" se define como todos los items de X y de Y. La confianza se interpreta como la probabilidad de que una transacción que contiene los items de X, también contenga los items de Y.

Continuando con el ejemplo, podemos observar que $\{C_1\}$ aparece en 3 de los 7 crímenes, $\{C_2\}$ en 6 y las dos juntas en 3. El soporte del item $\{C_1\}$ es por lo tanto del 43 %, el del item $\{C_2\}$ del 86 % y del item unido $\{C_1, C_2\}$ del 43 %. De los 3 crímenes que incluyen a $\{C_1\}$, también incluyen a $\{C_2\}$, por lo tanto, la regla "los crímenes de robos que tienen la característica $\{C_1\}$ también tienen la característica $\{C_2\}$", se cumple, un 100 %. Esto significa que la confianza de la regla $\{C_1, C_2\}$ es del 100 %.

Encontrar items frecuentes (items con una frecuencia mayor o igual a un determinado umbral definido) no es un proceso trivial debido a la explosión combinatoria de posibilidades, sin embargo, una vez identificados, es relativamente directo generar reglas de asociación que presenten una confianza mínima. El algoritmo Apriori hace una búsqueda exhaustiva por niveles de complejidad (de menor a mayor tamaño de items). Para reducir el espacio de búsqueda aplica la norma de "si un item no es frecuente, ninguna de sus combinaciones (combinaciones de items de mayor tamaño que contengan al primero) puede ser frecuente". Visto de otra forma, si un conjunto es infrecuente, entonces, todos los conjuntos donde este último se encuentre, también son infrecuentes. Por ejemplo, si el items $\{C_1, C_2\}$ es infrecuente, entonces, $\{C_1, C_2, C_3\}$ y $\{C_1, C_2, C_4\}$ también son infrecuentes ya que todos ellos contienen $\{C_1, C_2\}$.

El algoritmo se inicia identificando los items individuales que aparecen en el total de transacciones con una frecuencia por encima de un umbral definido por el usuario. A continuación, se extienden los candidatos añadiendo un nuevo item y se eliminan aquellos que contienen un subconjunto infrecuente o que no alcanzan el soporte mínimo. Este proceso se repite hasta que el algoritmo no encuentra más ampliaciones exitosas de los items previos o cuando se alcanza un tamaño máximo.

Se procede a identificar los items frecuentes de la tabla 3.1 y, a partir de ellos, crear reglas de asociación. Para este problema se considera que un item es frecuente si aparece un mínimo de 3 transacciones, es decir, su soporte debe de ser igual o superior a $3/7 = 0.43$. Se inicia el algoritmo identificando todos los items individuales y calculando su soporte.

Item (K=1)	Ocurrencias	Soporte
$\{C_1\}$	3	0.43
$\{C_2\}$	6	0.86
$\{C_3\}$	4	0.57
$\{C_4\}$	5	0.71

Tabla 3.2: Tabla que muestra las ocurrencias y el soporte de los items de tamaño k=1 de la tabla 3.1

Todos los items de tamaño k = 1 tienen un soporte igual o superior al mínimo establecido, por lo que todos superan la fase de filtrado. Posteriormente se generan todos los posibles items de tamaño k = 2 que se pueden crear con los items que han superado el paso anterior y se calcula su soporte.

Item (K=2)	Ocurrencias	Soporte
$\{C_1, C_2\}$	3	0.43
$\{C_1, C_3\}$	1	0.14
$\{C_1, C_4\}$	2	0.29
$\{C_2, C_3\}$	3	0.43
$\{C_2, C_4\}$	4	0.57
$\{C_3, C_4\}$	3	0.43

Tabla 3.3: Tabla que muestra todas las ocurrencias y el soporte de los items de tamaño k=2 la tabla 3.1

Los items $\{C_1, C_2\}$, $\{C_2, C_3\}$, $\{C_2, C_4\}$ y $\{C_3, C_4\}$ superan el límite de soporte, por lo que son frecuentes. Los items $\{C_1, C_3\}$ y $\{C_1, C_4\}$ no superan el soporte mínimo por lo que se descartan. Además, cualquier item que los contenga también será descartado ya que no puede ser frecuente por el hecho de que contiene un subconjunto infrecuente. Por lo que el resultado del algoritmo sería:

Item (K=2)	Ocurrencias	Soporte
$\{C_1, C_2\}$	3	0.43
$\{C_2, C_3\}$	3	0.43
$\{C_2, C_4\}$	4	0.57
$\{C_3, C_4\}$	3	0.43

Tabla 3.4: Tabla que muestra las ocurrencias y el soporte de los items de tamaño k=2 y soporte 0.43 de la tabla 3.1

Se repite el proceso, esta vez creando items de tamaño k = 3.

Item (K=3)
$\{C_1, C_2, C_3\}$
$\{C_1, C_2, C_4\}$
$\{C_2, C_3, C_4\}$
$\{C_3, C_4, C_1\}$

Tabla 3.5: Tabla que muestra los items de tamaño K=3 de la tabla 3.1

Los items $\{C_1, C_2, C_3\}$, $\{C_1, C_2, C_4\}$ y $\{C_3, C_4, C_1\}$ contienen subconjuntos infrecuentes, por lo que son descartados. Para los restantes se calcula su soporte:

Item (K=3)	Ocurrencias	Soporte
$\{C_2, C_3, C_4\}$	2	0.29

Tabla 3.6: Tabla que muestra las ocurrencias y el soporte de los items de tamaño k=3 de la tabla 3.1

El item $\{C_2, C_3, C_4\}$ no supera el soporte mínimo por lo que se considera infrecuente. Al no haber ningún nuevo item frecuente, se detiene el algoritmo.

Como resultado de la búsqueda se han identificado los siguientes itemsets frecuentes:

Items frecuentes
$\{C_1, C_2\}$
$\{C_2, C_3\}$
$\{C_2, C_4\}$
$\{C_3, C_4\}$

Tabla 3.7: Tabla que muestra los items de tamaño K=3 de la tabla 3.1

El siguiente paso es crear las reglas de asociación a partir de cada uno de los items frecuentes. De nuevo, se produce una explosión combinatoria de posibles reglas ya que, de cada item frecuente, se generan tantas reglas como posibles particiones binarias. Supongamos que se desean únicamente reglas con una confianza igual o superior a 0.7, es decir, que la regla se cumpla un 70 % de las veces. Tal y como se describió anteriormente, la confianza de una regla se calcula como el soporte del item formado por todos los items que participan en la regla, dividido por el soporte del item formado por los items del antecedente.

Reglas	Confianza	Confianza
$\{C_1\} \Rightarrow \{C_2\}$	$soporte\{C_1, C_2\}/soporte\{C_1\}$	$0.43/0.43 = 1$
$\{C_2\} \Rightarrow \{C_1\}$	$soporte\{C_1, C_2\}/soporte\{C_2\}$	$0.43/0.86 = 0.5$
$\{C_2\} \Rightarrow \{C_3\}$	$soporte\{C_2, C_3\}/soporte\{C_2\}$	$0.43/0.86 = 0.5$
$\{C_3\} \Rightarrow \{C_2\}$	$soporte\{C_2, C_3\}/soporte\{C_3\}$	$0.43/0.57 = 0.75$
$\{C_2\} \Rightarrow \{C_4\}$	$soporte\{C_2, C_4\}/soporte\{C_2\}$	$0.43/0.86 = 0.5$
$\{C_4\} \Rightarrow \{C_2\}$	$soporte\{C_2, C_4\}/soporte\{C_4\}$	$0.43/0.71 = 0.6$
$\{C_3\} \Rightarrow \{C_4\}$	$soporte\{C_3, C_4\}/soporte\{C_3\}$	$0.43/0.57 = 0.75$
$\{C_4\} \Rightarrow \{C_3\}$	$soporte\{C_3, C_4\}/soporte\{C_4\}$	$0.43/0.71 = 0.6$

Tabla 3.8: Tabla que muestra las reglas de asociación de la tabla 3.1

De todas las posibles reglas, únicamente $\{C_3\} \Rightarrow \{C_2\}$ y $\{C_3\} \Rightarrow \{C_4\}$ superan el límite de confianza.

La principal desventaja de algoritmo Apriori es el número de veces que se tienen que escanear los datos en busca de los items frecuentes, en concreto, el algoritmo escanea todas las transacciones un total de $k_{max} + 1$, donde k_{max} es el tamaño máximo de item permitido. Esto hace que el algoritmo Apriori no pueda aplicarse en situaciones con millones de registros.

Capítulo 4

Desempeño académico e infraestructura de las escuelas

Para el presente trabajo se tomaron las bases de datos de la prueba ENLANCE[1] del año 2013, y la base de CEMABE[2] del mismo año, ambas constan con información de las escuelas a nivel básico en México del desempeño académico y de la infraestructura respectivamente.

4.1. Prueba Enlace

Parte de la información se tomó de [DGEP(2013)]. El programa educativo ENLACE (Evaluación Nacional del Logro Académico en Centros Escolares) inició casi al concluir el ciclo escolar 2005-2006, a fin de evaluar el logro académico de todos los alumnos de tercero a sexto grados de educación primaria y del tercer grado de educación secundaria en las asignaturas de Español y Matemáticas (con base en los planes y programas de estudio aprobados por la Secretaría de Educación Pública y vigentes en el momento). Actualmente, ENLACE cubre de tercero a sexto grados de educación primaria y todos los grados de educación secundaria y explora cada año (además de las asignaturas referidas) una más, la cual varía para cada ciclo escolar y se repite cada cuatro años.

La prueba ENLACE es un instrumento estandarizado, objetivo, de alcance nacional, diseñado para que los docentes, directivos, autoridades educativas, investigadores y escolares de todo el país, dispongan de una medida válida, objetiva y confiable, del estado actual del logro académico de los estudiantes de educación básica. En este sentido, se cumple la expectativa de que, con el paso del tiempo,

[1]Las bases se tomaron de: http://enlace.sep.gob.mx/ba/db/escuelas.html
[2]Las bases se tomaron: http://www.mejoratuescuela.org/bases

ENLACE se constituya (a partir de sus resultados) en una referencia válida y confiable de la evolución del avance en el desempeño escolar, de la concreción de los esfuerzos de todo el sistema escolar en los resultados escolares, tomando en cuenta diferentes niveles de agregación: estatal, municipal, local, escolar, grupal e individual.

El propósito primordial de ENLACE es recopilar datos y producir información respecto del logro académico de cada alumno de las escuelas de educación primaria y secundaria del país. Con el procesamiento de los resultados se cuenta con información específica de la población objetivo para: (1) identificar áreas donde hay progreso, (2) reconocer donde hay deficiencias y, por tanto, se erigen como áreas de oportunidad para diseñar mediaciones pedagógicas a realizar en clase por los docentes, (3) intercambiar opiniones de las que emanen acciones donde intervengan los padres de familia para incidir en el aprendizaje y el desarrollo de sus hijos, (4) socializar el trabajo de la escuela y (5) fortalecer la idea de comunidad escolar y su participación en los procesos formativos de los estudiantes.

Los instrumentos de evaluación utilizados en el programa ENLACE en educación básica tienen como principales objetivos:

a) Medir el logro académico en Español y Matemáticas (y las competencias de otro ámbito del conocimiento, diferente cada año pero que se repite cíclicamente) de todos los alumnos de los grados educativos considerados.

b) Establecer criterios y estándares de calidad aceptados en todo el país, como una base de referencia. No se trata de conocimientos o habilidades mínimos, sino los comunes o críticos aceptables para todo el país.

c) Obtener y entregar resultados de todos los alumnos y todas las escuelas.

4.1.1. Características de la prueba

Siguiendo con los estándares nacionales e internacionales, las características básicas que posee ENLACE son:

Estandarización Se refiere a que el diseño, la administración y la calificación de la prueba se hacen en condiciones iguales para todos los examinados, con atención al estrato socioeconómico, tipo de escuela u otras características distintivas de la población evaluada. ENLACE es una prueba estandarizada para poblaciones ubicadas en un cierto grado escolar; no se estandariza por edad, estrato socioeconómico u otro atributo poblacional.

Objetividad Se refiere a que la prueba cuenta con una metodología de diseño que garantiza su independencia respecto de los sujetos que son medidos, evitando cualquier influencia injustificada de parte de evaluadores ante grupos particulares de sujetos, con un esquema de calificación preciso y preestablecido, común para todos los estudiantes. Gracias a la objetividad, ENLACE permite

realizar comparaciones contra referencias nacionales y contra criterios externos de desempeño.

Enfoque Se refiere a que la prueba está centrada en el rendimiento (logro académico) y trata de reflejar el resultado del trabajo escolar, a diferencia de otras pruebas que miden destrezas innatas.

Comparabilidad Se refiere a que ENLACE cuenta con una metodología de comparación de resultados de años sucesivos con base en una escala estandarizada por medio de aplicaciones piloto realizadas en condiciones de muestreo poblacional controlado, utilizando una escala subyacente basada en la Teoría de la Respuesta al Ítem (TRI). Para el diseño de ENLACE (en su versión 2013) se consideran las áreas de Español, Matemáticas y Formación Cívica y Ética, con lo cual se tienen tres pruebas objetivas a ser aplicadas durante dos días. Al terminar la aplicación, los estudiantes pueden conservar las pruebas lo cual produce una gran volatilidad en las pruebas, ya que no se reutilizan los reactivos en aplicaciones posteriores. Para el diseño se aplica una muestra controlada que sirve para determinar la confiabilidad y la precisión de los resultados que se obtienen. El propósito de esta muestra controlada se fundamenta en la necesidad de contar con reactivos calibrados para ser empleados en la aplicación del año siguiente, esto es, disponer de la escala de referencia para asignar puntajes a los estudiantes en el siguiente año con referencia a la métrica del año anterior.

4.1.2. Definición de la escala de la prueba ENLACE y modelo de calificación

Con el propósito de mejorar la interpretación de los resultados, la escala con la que se define ENLACE considera valores que van de 200 a 800 puntos, con un puntaje medio de 500 puntos y una desviación estándar de 100 puntos. Dentro de esta escala se establecen los niveles de logro por cada grado y asignatura. Una escala específica, como ésta que va de 200 a 800 puntos, evita la interpretación tradicional que se acostumbra al tomar el 5 como reprobación y el 6 como aprobación, por ejemplo. Para ENLACE no se tiene un punto de corte de aprobación, debido a que el interés es la medida de los conocimientos y las habilidades de cada alumno; indica capacidades específicas que pueden detallarse a lo largo de la escala.

La escala de ENLACE cuenta con dos elementos de referencia: una escala objetiva para reportar los resultados y una escala subyacente respecto de los constructos. Con la escala se cuenta con una métrica sobre la cual se tiene una mejor interpretación de las competencias en las cuales se presentan los desempeños de los estudiantes, con la máxima precisión posible. La amplitud de la escala (200-800) permite reportar resultados con una mayor precisión que la escala 0-10. Considerando que, dependiendo del grado escolar, se tiene un error estándar de medida de diseño de un 4 ó 5 %, aproximadamente, que es muy similar a 24 ó

30 puntos en la escala de 200 a 800 (cada punto es 1/600 de la escala, por lo que 1 % es equivalente a 6 puntos de la escala). La segunda faceta de la escala es subyacente, con la cual se puede asignar a los alumnos a niveles de logro y potenciar la comparabilidad entre años distintos. Esta escala subyacente está basada en la Teoría de Respuesta al Ítem; en ella la puntuación no depende sólo del número de respuestas correctas, sino de cuáles reactivos se respondieron correctamente. Con esta escala subyacente se puede afirmar en qué nivel y subnivel de logro se encuentra el alumno y, además, cómo es su rendimiento con respecto a los demás alumnos del país.

El proceso de calificación con el cual se calculan los puntajes de los alumnos se divide en dos fases: la primera sirve para observar estadísticamente los reactivos con el modelo clásico (dificultad como proporción de aciertos y la correlación punto-biserial como aproximación de la discriminación de los reactivos); en una segunda fase se calibran los reactivos y se califica a los sujetos por medio del modelo de tres parámetros de la Teoría de la Respuesta al Ítem (TRI), que considera la adivinación, la dificultad y la discriminación. El modelo permite asignar a cada alumno un valor en puntaje no sólo por la cantidad de respuestas correctas sino por cuáles reactivos contestó acertadamente. El valor obtenido en escala logarítmica se transforma a una forma estandarizada, con media en 500 y desviación estándar de 100 para cada grado- asignatura. La escala se establece para cada grado y asignatura; por lo tanto, resulta incorrecto hacer comparaciones de puntajes entre niveles, asignaturas y grados diferentes.

La dimensión implícita de ENLACE depende de la dificultad de los reactivos, con la cual se establecen tres niveles de dificultad para los reactivos y cuatro niveles de logro para los sujetos. El procedimiento para establecer estos niveles se esquematiza a continuación:

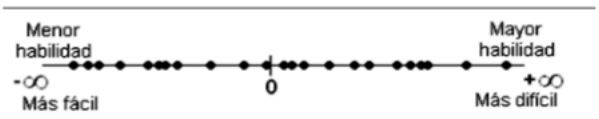
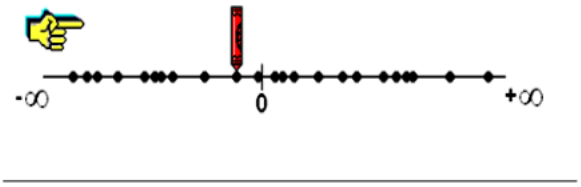
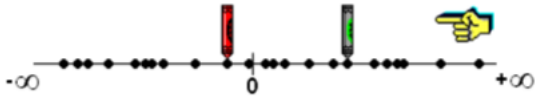
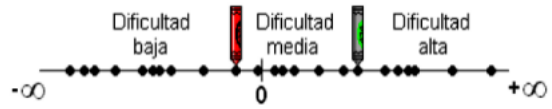
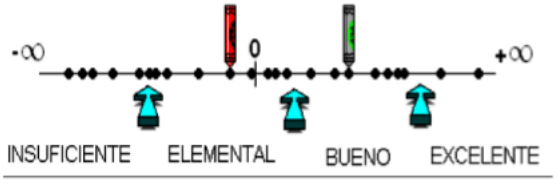
Paso 1. Ubicar los reactivos en la escala en función de su medida, ordenados del más fácil al más difícil.	
Paso 2. El primer punto de corte parte desde el reactivo más fácil de la prueba hasta llegar a la posición del reactivo referido a un conocimiento o habilidad que representa un mínimo a dominar por un alumno del grado.	
Paso 3. El segundo punto de corte parte del reactivo más difícil de la prueba hasta llegar a la posición del reactivo por arriba del cual se tienen conocimientos o habilidades clasificadas como sobresalientes.	
Paso 4. Los dos puntos de corte obtenidos establecen los tres niveles de dificultad: baja, media y alta.	
Paso 5. A partir de los tres niveles, se determina el punto medio de cada nivel. Con esto se tienen cuatro niveles de logro (insuficiente, elemental, bueno y excelente).	

Figura 4.1: Niveles de la prueba ENLANCE de [DGEP(2013)].

Los alumnos en el nivel insuficiente responden menos del 50 % de los reactivos de dificultad baja y los alumnos en el nivel excelente responden al menos el 50 % de los reactivos de dificultad alta. Una vez organizados los reactivos por niveles de dificultad, se especifica el constructo o contenido de las competencias, a partir del análisis de las comunalidades de los reactivos ubicados por debajo del primer punto de corte. Dentro de cada nivel se establecen tres subniveles con características similares. Los puntos de corte para los niveles de logro no son los mismos en todos los casos, ya que se definen para cada grado y asigna-

tura. Se muestra, aquí, a manera de ejemplo, las tablas donde puede apreciarse claramente que las escalas son específicas para cada caso.

Las pruebas que se aplican en ENLACE son conceptuadas y elaboradas de acuerdo con un proceso de planeación que se inserta dentro de una metodología de diseño de la evaluación cumpliendo con estándares técnico-pedagógicos internacionales que aseguren la pertinencia de sus resultados y redunden en propuestas eficaces de mejora continua.

TERCERO DE PRIMARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	416,427038	416,427038	560,301233	560.301233	691,593427	691,593427
MATEMÁTICAS	419,633171	419,633171	565,623440	565,623440	674,076003	674,076003
FCE	423,979726	423,979726	512,129618	512,129618	655,389863	655,389863

CUARTO DE PRIMARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	414,729226	414,729226	573,200083	573,200083	699,021147	699,021147
MATEMÁTICAS	415,933056	415,933056	587,869326	587,869326	710,616901	710,616901
FCE	418,534241	418,534241	528,206526	528,206526	658,912122	658,912122

QUINTO DE PRIMARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	410,926869	410,926869	585,292037	585,292037	722,478577	722,478577
MATEMÁTICAS	416,880173	416,880173	586,798008	586,798008	714,307586	714,307586
FCE	417,381310	417,381310	514,977861	514,977861	666,628153	666,628153

SEXTO DE PRIMARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	413,843242	413,843242	581,624259	581,624259	714,008194	714,008194
MATEMÁTICAS	412,609821	412,609821	608,127787	608,127787	735,704144	735,704144
FCE	401,948134	401,948134	509,292507	509,292507	668,474004	668,474004

PRIMERO DE SECUNDARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	446,318282	446,318282	593,189536	593,189536	735,694263	735,694263
MATEMÁTICAS	507,272622	507,272622	634,854980	634,854980	737,262289	737,262289

SEGUNDO DE SECUNDARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	445,084310	445,084310	592,411914	592,411914	735,359421	735,359421
MATEMÁTICAS	505,399184	505,399184	634,056977	634,056977	737,327513	737,327513
FCE	363,804586	363,804586	489,655082	489,655082	636,765587	636,765587

TERCERO DE SECUNDARIA

ASIGNATURA	NIVELES DE LOGRO					
	INSUFICIENTE	ELEMENTAL		BUENO		EXCELENTE
	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A	MENOR O IGUAL	MAYOR A
ESPAÑOL	462,942346	462,942346	608,224039	608,224039	749,186463	749,186463
MATEMÁTICAS	525,993007	525,993007	657,032576	657,032576	762,214906	762,214906
FCE	376,131530	376,131530	481,559040	481,559040	654,416771	654,416771

Tabla 4.1: Niveles de logro de [DGEP(2013)].

4.1.3. Descripción de la base

ENLANCE viene dividido por entidad federativa y a su vez cada entidad está separada por primaria y secundaria, así se tiene un total de 66 archivos (32 entidades federativas y una más que proviene de escuelas nacionales), en total las bases para cada entidad federativa tienen un total de 62 columnas, las cuales constan de la información por escuela del turno, del nombre, de la ubicación, tipo de escuela (público o general), el grado de marginación, además de los resultados promedio por año escolar, en total según datos de ENLANCE, se tiene en total el 95 % aproximadamente (120,000) de las escuelas del nivel básico en México.

ENT	NOMBRE DE LA ENTIDAD	CLAVE DE LA ESCUELA	TURNO	NOMBRE DE LA ESCUELA	TIPO DE ESCUELA	CLAVE MUN.	NOMBRE DEL MUNICIPIO	CLAVE LOC.
01	AGUASCALIENTES	01DPR0001V	1	RODRIGO RINCON GALLAR	GENERAL	007	RINCON DE ROMOS	0001
01	AGUASCALIENTES	01DPR0002U	1	ANTONIO VENTURA MEDI	GENERAL	008	SAN JOSE DE GRACIA	0001
01	AGUASCALIENTES	01DPR0003T	2	BENIGNO CHAVEZ	GENERAL	005	JESUS MARIA	0001
01	AGUASCALIENTES	01DPR0004S	2	MOISES SAENZ	GENERAL	010	EL LLANO	0001
01	AGUASCALIENTES	01DPR0005R	1	15 DE SEPTIEMBRE	GENERAL	001	AGUASCALIENTES	0001
01	AGUASCALIENTES	01DPR0006Q	1	TIERRA SOLIDARIA	GENERAL	001	AGUASCALIENTES	0001
01	AGUASCALIENTES	01DPR0007P	1	MIGUEL ANGEL BARBERE	GENERAL	001	AGUASCALIENTES	0001

MEDIAS											
TERCER GRADO			CUARTO GRADO			QUINTO GRADO			SEXTO GRADO		
ESPAÑOL	MATEMÁTICAS	FORMA.CÍV. Y ÉTICA	ESPAÑOL	MATEMÁTICAS	FORMA.CÍV. Y ÉTICA	ESPAÑOL	MATEMÁTICAS	FORMA.CÍV. Y ÉTICA	ESPAÑOL	MATEMÁTICAS	FORMA.CÍV. Y ÉTICA
439.44	433	415.88	456.56	530.52	434.01	465.46	472.08	426.11	458.46	529.57	433.78
533.35	536.41	515.22	485.45	500.64	486.86	429.29	417.15	411.62	484.9	499.85	440.76
512.43	502.4	478.67	473.83	471.05	463.92	498.41	504.99	472.8	501.48	504.32	477.39
429.75	418.54	410.89	461.15	464.39	470.36	412.82	443.89	419.27	414.48	463.94	400.94
594.73	589.22	552.93	549.61	563.78	528.27	596.05	589.31	565.61	601.61	607.4	570.76
591.68	611.2	558.81	620.34	646.38	685.83	539.24	586.25	521.09	586.75	580.81	556.52
515.83	521.29	493.93	463.26	464.4	459.51	607.1	615.87	559.07	567.66	581.26	561.52

Tabla 4.2: Ejemplo de la información que contiene la base de datos de la prueba ENLANCE

La tabla 4.2, muestra un ejemplo de la bases da datos de ENLANCE, cada fila corresponde a una escuela y en cada columna hay información de ésta.

4.2. Cemabe

Parte de la información se tomó de [INEGI(2014)]. La Secretaría de Educación Pública (SEP) y el Instituto Nacional de Estadística y Geografía (INEGI) presentaron los resultados del Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (CEMABE), que se llevó a cabo del 26 de septiembre de 2013 al 29 de noviembre del mismo año. El CEMABE se realizó para conocer estadísticamente el sistema educativo nacional mediante la ubicación geográfica de todos los planteles, el conocimiento de la situación de la infraestructura instalada, equipamiento y mobiliario escolar, el registro de todos y cada uno de los docentes, personal administrativo y supervisores. Asimismo, se integró un registro nacional de alumnos, centrando atención en quienes tienen alguna discapacidad o necesidades especiales de aprendizaje. En total se censaron 236,973 centros de trabajo, donde se contemplan escuelas, bibliotecas, centros de maestros, oficinas administrativas y de supervisión, entre otros. Del total de centros de trabajo censados, 207,682 son escuelas de educación básica y especial, de las cuales el 86.4 % son de carácter público y 13.6 % son privadas. Por nivel escolar, la distribución de planteles fue de 40.1 % preescolar, 42.5 % primaria, 16.7 % secundaria, y 0.7 % centros de atención múltiple.

El total de alumnos en dichos planteles fue de 23 millones 562 mil 183, de los cuales el 18.3 % pertenecen a nivel preescolar, 55.8 % a primaria, 25.6 % a secundaria y el 0.3 % a centros de atención múltiple. Por otro lado, el personal que labora en los centros de trabajo alcanzó la cifra de 1 millón 949 mil 105. De ellos, 88.1 % desempeña funciones en escuelas de educación básica y especial, 2 % lo hace en centros de trabajo de apoyo a la educación especial y 9.9 % en otro tipo de centros de trabajo. El personal que desempeña función de maestro

frente a grupo fue de 978 mil 118.

Respecto a los servicios básicos e infraestructura en los centros de trabajo los resultados son alarmantes y desiguales: mientras que en las escuelas públicas tan sólo 51.6 % cuentan con drenaje, 69 % con disponibilidad de agua potable, 87.2 % con sanitarios y 88.8 % con energía eléctrica, en las escuelas privadas casi cumplen al 100 % con la demanda de estos servicios básicos. La SEP y el INEGI afirman que los datos y elementos del CEMABE permitirán emprender acciones concretas y eficaces por parte de todos los agentes educativos para elevar la calidad de la educación en nuestro país y garantizar el pleno acceso a este derecho de las y los niños.

El propósito fundamental del CEMABE fue captar las características específicas de las escuelas, maestros y alumnos de instituciones públicas y privadas de educación básica del sistema educativo escolarizado y especial, con el propósito de proveer información al Sistema de Información y Gestión Educativa del país, además de:

- Disponer de la ubicación geográfica en el Marco Geoestadístico del INEGI, de todos los inmuebles donde se imparte educación de nivel básico y de educación especial, así como de aquéllos en los que se atienden asuntos de tipo administrativo.
- Conocer la situación de la infraestructura instalada, los servicios, el equipamiento y mobiliario escolar de cada inmueble educativo, así como el uso de los espacios disponibles, con el fin de determinar las condiciones en las que se imparte la educación básica y especial.
- Tener un registro de los datos generales de todo el personal que labora en los centros de trabajo, así como las características de sus plazas, la escolaridad alcanzada, las funciones que realizan, la capacitación recibida y los programas en que participan; en el caso de los docentes frente a grupo, se identifica el grado, grupo y materias que tienen a cargo y las horas de clase que imparte a la semana.
- Contar con un registro de cada uno de los alumnos según su grado y nivel educativo, así como sus características sociodemográficas generales.

4.2.1. Descripción de la base

Como se mencionó anteriormente, el CEMABE captó información de las características de los inmuebles, de los centros de trabajo, del personal y de los alumnos; sin embargo en cumplimiento al artículo 38 de la Ley del Sistema Nacional de Información Estadística y Geográfica (SNIEG) para garantizar la privacidad y reserva de los datos personales, únicamente se publican las tablas de datos correspondientes a inmuebles y centros de trabajo que imparten educación básica o especial o son de apoyo a la educación especial y para complementar

la información, se incluyen las principales características del personal y de los alumnos agregadas por centro de trabajo. En función de la naturaleza propia de las tablas, la separación de inmuebles y escuelas se debe a que en un mismo inmueble pueden operar diferentes centros de trabajo, particularmente cuando se abre un turno adicional o cuando se prestan servicios educativos de distinto nivel. Razón por la cual de origen se diseñaron instrumentos de captación para levantar sus características de manera independiente.

CEMABE consta de información de bibliotecas, centros de maestros, etc; como a nosotros únicamente nos interesa la educación básica, nos quedamos con esa parte de la información. Las bases a diferencia de ENLANCE constan de sólo dos archivos uno para primaria y otro para secundaria. Según datos de CEMABE esta información consta del 98 % aproximadamente (140,000) de las escuelas; la información viene dada por escuela, por lo que la información consta de un total 266 columnas sobre la infraestructura de la escuela; si la escuela recibe apoyo federal, si tiene patio, a que hora abre, a que hora cierra, etc.

ID_INM	CLAVE_CT	ENT	NOM_ENT	MUN	NOM_MUN	LOC	NOM_LOC
664	01DPR0698R2	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
644	01PES0046P1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
513	01DPR0647K1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
477	01DPR0073O2	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
650	01PJN0124T1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
636	01DPR0127B2	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
472	01FUA0041F1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
598	01DES0007I1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes
741	01PJN0134Z1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes

P231	P232	P233	P234	P235	P236	P237	P238	P239	P240	P241	P242	P243	P244	P245	P246	P247	P248	P249	P250	P251	P252	P253	P254	P255	P256
1		12	1		12	1	1	999	2		0		2	2	2		2	2	2	2	2	2	2	2	2
1		0	1		0	8	0		2		0				2								2		2
1		4	1		4	4	2		2		0			2	2		2	2	2	2	2	2	1	2	2
1		0	1		0	0		2	2		0		1	2	2		2	2	2	2	2	2	2	2	2
1		0	1		0	0			2		0				2								2		2
1		6	1		4	0		2	2		0		1	2	2		2	2	2	2	2	2	2	2	2
						0			2		0														
1		18	1		0	7	0	0	2		0		2	2	1		1	2	2	2	2		1	2	2
1		0	1		1	1	0		2		0				2								1		1

Tabla 4.3: Ejemplo de la información que contiene la base de datos de CEMABE

La tabla 4.3, muestra un ejemplo de la base de CEMABE, cada fila corresponde a una escuela y en cada columna hay información de ésta, debido al gran número de columnas se hizo un diccionario que describe; P231 de la tabla 4.3 es ¿Todas las aulas para impartir clase cuentan con un escritorio o mesa para el maestro?, etc. Por otro lado la base de datos se encuentra a nivel de $CLAVE\_CT$, campo que identifica a cada una de las escuela de manera única, por lo que cada registro representa a una escuela.

Capítulo 5

Metodología aplicada y resultados

5.1. Descripción general

En este capitulo se presentara la metodología que llevamos acabo para clasificar las escuelas de nivel básico en México. Se describirá como se creo la base datos final, como a partir de ésta se hizo la clasificación en tres *clusters* y por último, como se determinó, que representaba cada *cluster*.

5.1.1. ETL

El primer paso fue el proceso de ETL, como observamos del capitulo 4 tenemos la información del nivel educativo (ENLANCE) y de la infraestructura (CEMABE) de las escuelas de nivel básico en México, esta información viene organizada por escuela, ejemplo:

ENT	NOMBRE DE LA ENTIDAD	CLAVE DE LA ESCUELA	TURNO	NOMBRE DE LA ESCUELA	TIPO DE ESCUELA
01	AGUASCALIENTES	01DES0001O	1	LIC. BENITO JUAREZ	GENERAL
01	AGUASCALIENTES	01DES0001O	2	LIC. BENITO JUAREZ	GENERAL
01	AGUASCALIENTES	01DES0002N	1	MOISES SAENZ	GENERAL

Tabla 5.1: Ejemplo de la información que contiene la base de datos de la prueba ENLANCE

ID_INM	CLAVE_CT	ENT	NOM_ENT	MUN
000664	01DPR0698R2	01	Aguascalientes	001
000644	01PES0046P1	01	Aguascalientes	001
000513	01DPR0647K1	01	Aguascalientes	001

Tabla 5.2: Ejemplo de la información que contiene la base de datos de CEMABE

En este caso tenemos la información para primaria de la prueba ENLANCE

y CEMABE, como podemos observar de la tabla 5.2 ambas bases cuentan con un elemento en común el cual es "CLAVE DE LA ESCUELA" y "CLAVE _ CT". Por lo que el primer paso fue unirlas, cabe mencionar que ENLANCE viene dividido por entidad federativa y a su vez cada entidad está separada por primaria y secundaria, así se tiene un total de 66 archivos (32 entidades federativas y una más que proviene de escuelas nacionales), por otro lado CEMABE tiene la información en sólo dos archivos (primaria y secundaria). Para unir estas bases se uso el identificador de "CLAVE _ CT"por lo que el resultado quedaría:

CLAVE_CT	ID_INM	ENT	NOM_ENT	MUN	LOC	AGEB	MZA	NIVEL	MODALIDAD	TURNO	CONTROL
01DES0001O	426	1	Aguascalientes	1	1	619	21	4	1	2	1
01DES0002N	699	1	Aguascalientes	1	239	419A	30	4	1	1	1
01DES0003M	1138	1	Aguascalientes	7	1	59	5	4	1	2	1

Tabla 5.3: Ejemplo de unir ENLANCE y CEMABE

Al hacer este proceso, como mostramos en el capitulo 4, no se tiene el 100 % de las escuelas, por lo que al juntarlas había escuelas que no estaban en la base de ENLANCE o en la de CEMABE por lo que sólo se juntaron aquellas que estaban en ambas bases.

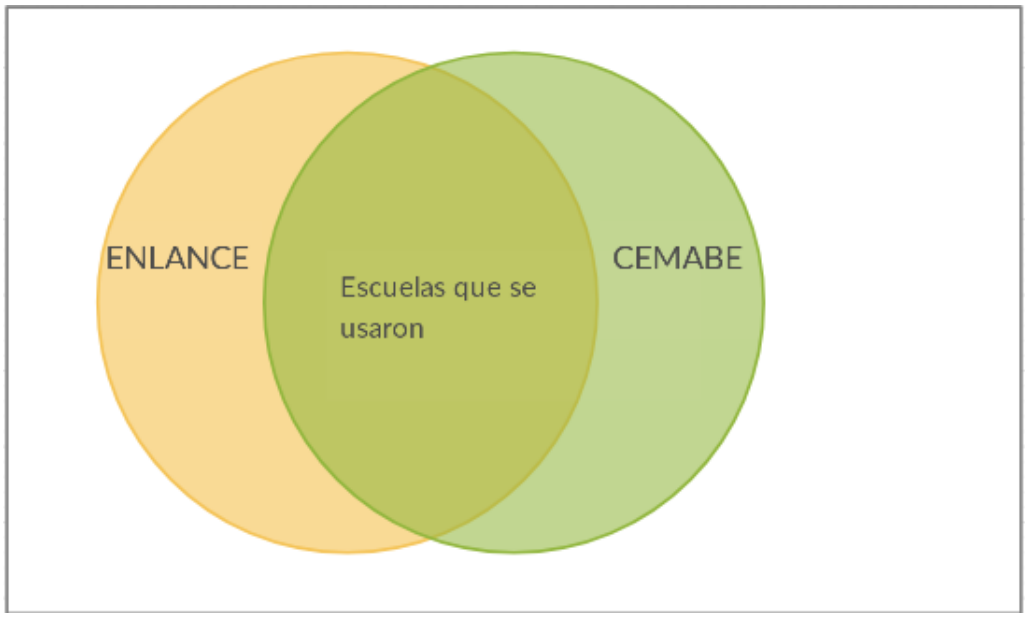


Figura 5.1: Figura que muestra las escuelas que nos interesan en este trabajo

Una vez juntadas se procedió a limpiar la base, primero se quitaron datos repetidos usando "CLAVE _ CT" como indicador principal, al juntarlas y al quitar repetidas el número total de escuelas que se quedó es de aproximadante 90,000, posteriormente se quito la información redundante y aquella que no sirviera para aplicar los algoritmos de clasificación, ejemplo:

Nombre Columna
NOM_ENT
MUN
LOC
AGEB
MZA

Tabla 5.4: Muestra de las columnas que se quitaron de la unión de ENLANCE y CEMABE

La tabla 5.4 es una muestra de todas las columnas que se quitaron y el criterio para esto fue el siguiente:

- ENLANCE: Se quitaron aquellas columnas que se compartían con CEMABE como el turno, entidad, etc.

- CEMABE: Se quitaron aquellas columnas que son redundantes como las de la figura 5.4, esto se debe a que estas columnas hacen alusión a la localización de la escuela, pero ésta está ligada a la entidad federativa por lo que al quedarnos unicamente con la columna de ENT es suficiente y aquellas que a consideración no aportan nada a la clasificación como por ejemplo la columna que marcaba si la escuela tenía arenero o no, entre otras.

La eliminación de las columnas es más un arte, debido a que la intuición y la experiencia es la que nos marca que columnas son más importantes y que columnas pueden causar ruido [Gareth James(2013)]. Por lo que el número de columnas es de 183 para secundarias y 205 para primarias. Entre las columnas que se eliminaron están las siguientes.

Nombre Columna	Descripción
Municipio, localidad, etc	Como es una clasificación a nivel nacional se ignora las variables de ubicación
P194	Uso de arenero por el personal o alumnos, no se menciona que es un arenero
P238	Número de proyectores o cañones que no funcionan, se puede sesgar por las escuelas que no tienen
P302	Personal masculino censado, se tomo en cuenta todo el personal sin distinción

Tabla 5.5: Ejemplo de columnas omitidas.

5.1.1.1. Transformación y limpieza

Una vez teniendo la unión de ENLANCE y CEMABE, se procedió a transformarla, notemos de la tabla 4.3 que P244 representa ¿El centro de trabajo participa en el programa de enciclomedia?, donde los valores de esta preguntan son: 1 si participa, 2 si no participa, 9 no especificó y nulo si no hay información, notemos que esta columna a pesar de tener números en sus registros es una columna con datos cualitativos, por lo que se tomó la decisión de transformar los datos a numérico, quedando únicamente 1 si la escuela participa y 0 para los otros casos. Esto se hizo para todas las columnas con datos cualitativos, debido a que todas son preguntas de sí o no, por ejemplo ¿Todos los alumnos tienen silla para sentarse?, ¿La escuela tiene internet? [de Jonge and van der Loo(2013)].

El tratamiento para las variables numéricas es; calcular los valores atípicos usando el algoritmo de Tukey 2.4, estos valores se transformaron a la media de los datos, esto se debe a que es la solución menos compleja y que no es sesgada [Vargas(2015)], se calcularon estos valores ya que había columnas que no tenían sentido numérico, por ejemplo, la columna de ¿Cúantos alumnos hay en la escuela? había datos con hasta 20,000 alumnos.

Se crearon dos columnas con base en la información que se tenia, las cuales fueron, CAPACIDAD la cual consiste de la capacidad que tiene la escuela, se calculó con el número de alumnos que tiene inscritos la escuela entre el número total de alumnos que soporta la escuela y HORAS que representa las horas que trabaja la escuela, se calculó con la hora a la que abre y cierra la escuela.

Por último se transformó las variables numéricas, a una distribución normal con media 0 y desviación estándar 1, esto se debe a que hay variables numéricas que son más grandes que otras, por ejemplo: el número de sillas que hay en la escuela; esta variable puede alcanzar números que van en promedio de 50 hasta 1000 sillas, comparada con la columna de HORAS, que en promedio va de 4 a 10 horas, por lo que, si se dejan los datos numéricos sin transformarlos, los de mayor rango tienen un mayor peso a la hora de clasificar comparados con los de un menor rango [Gareth James(2013)]. Así al final de estos pasos tenemos la base de datos final, la cual es la que usamos para aplicar los algoritmos de clasificación.

5.1.2. Obteniendo K

Del capitulo 3 observamos que para poder hacer una clasificación no supervisada usando el algoritmo $K - medias$ se necesita de antemano el número K, donde éste determina el número de *clusters* que tendrá la clasificación, así por ejemplo con $K = 2$ habrá dos clasificaciones, recordamos que determinar K no es trivial. Así para nuestro problema primero usamos el algoritmo jerárquico.

5.1.2.1. Algoritmo jerárquico

De la base de datos final, notemos que son dos bases (primaria y secundaria), a éstas les aplicamos PCA en donde el número de componentes se determinó a un 85 % del total de los datos, dando así aproximadamente 25 variables para primaria y secundaria, recordamos que PCA reduce la dimensionalidad de los datos, por ejemplo para nuestro caso pasamos de una base con 205 columnas a 25 para primaria. Una vez aplicando PCA usamos el algoritmo jerárquico visto en el capitulo 3, recordando que éste nos ayuda a determinar K, dando como resultado:

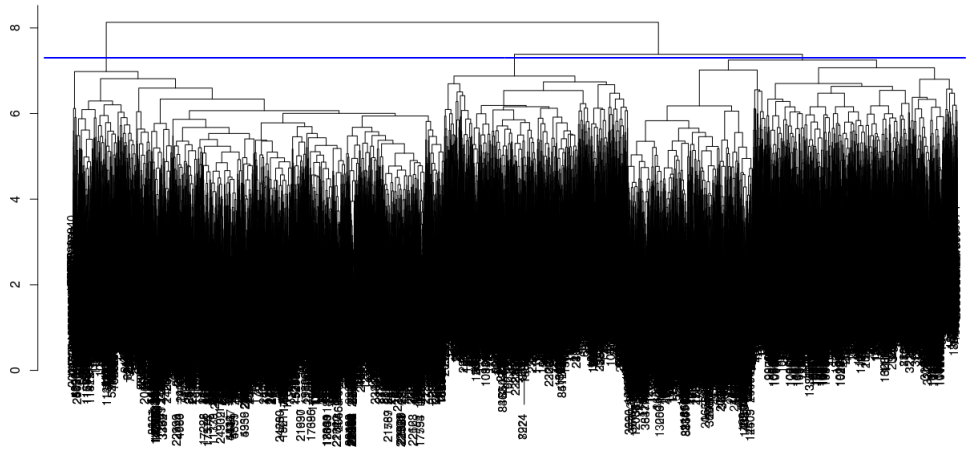


Figura 5.2: Dendrograma de la base final de primaria con un corte de tres toques.

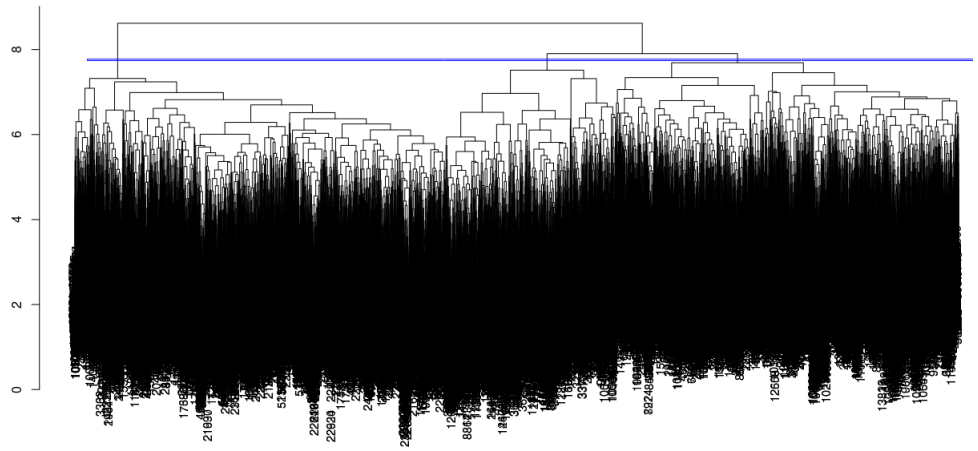


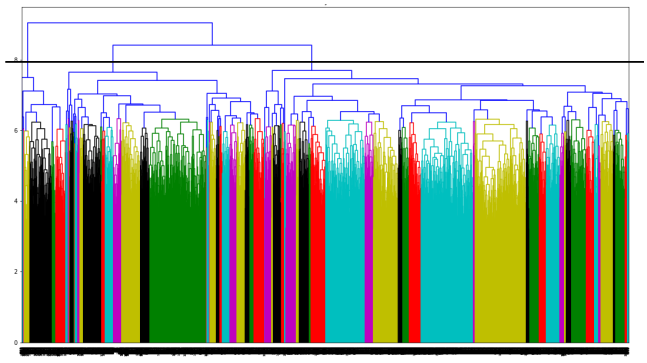
Figura 5.3: Dendrograma de la base final de secundaria con un corte de tres toques.

Como podemos observar de la figura 5.2 y 5.3, el corte que se mantiene en ambos árboles (existe en ambos) y que además cada toque que se hace en el árbol contiene la mayoría de la información, es de 3, aunque puede haber cortes por ejemplo de 4 o 5 toques que mantiene las propiedades anteriormente mencionadas, el corte de 3 toques tiene una ventaja, la cual es la fácil interpretación de las escuelas, recordamos que cada hoja del árbol es una escuela y en la rama en donde se corta el árbol todas las escuelas que estén en esa rama se clasifican de la misma manera, entonces al haber 3 clasificaciones podemos concluir que hay tres niveles, los cuales pueden ser malo, regular y bueno, o rojo, amarillo y verde, etc, en cambio la interpretación que representa cada clasificación se complica conforme el número de toques al árbol crece. Por lo anterior se podría escoger $K = 3$.

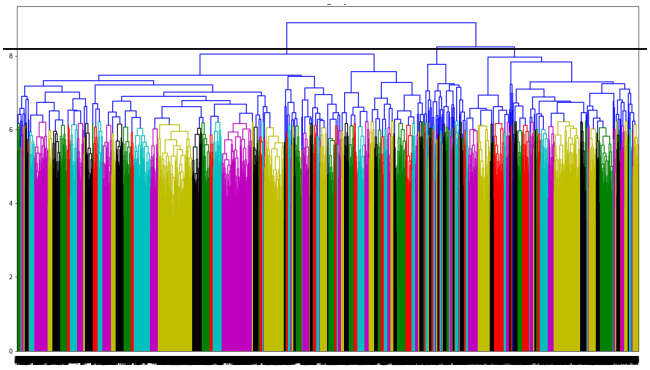
Posteriormente continuamos con el análisis para elegir K; el siguiente paso fue ver si se mantenía $K = 3$ en las diferentes regiones de México, usando como referencia la figura 5.4.

Figura 5.4: Regiones de México

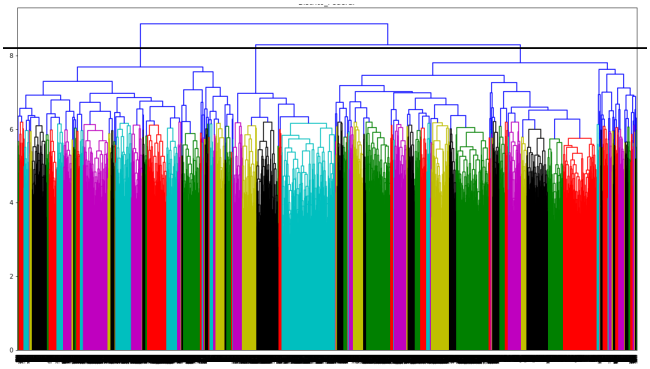
Así la base de datos final se separó por región, aplicamos PCA y el algoritmo jerárquico (como anteriormente describimos) a cada una de las bases formadas por las diferentes regiones, dando como resultado:



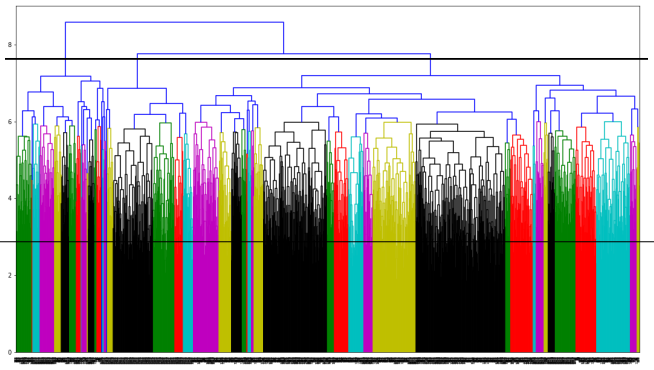
(a) Bajío



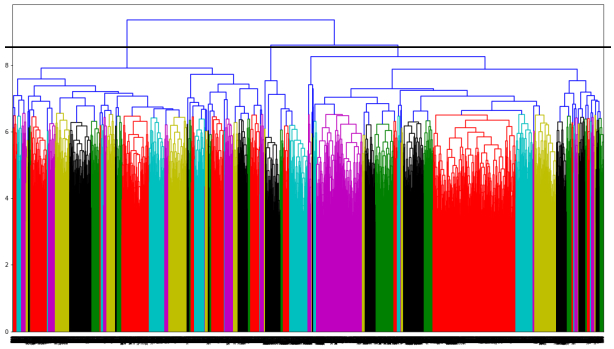
(b) Centro y golfo



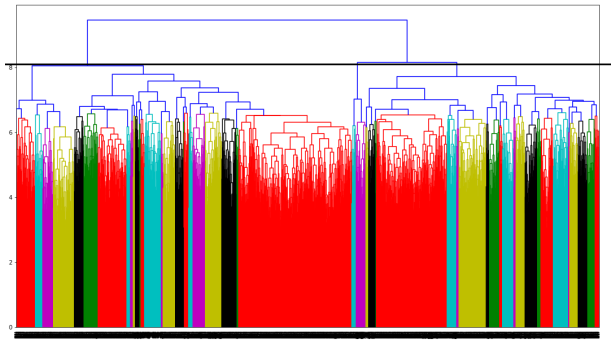
(c) Distrito Federal



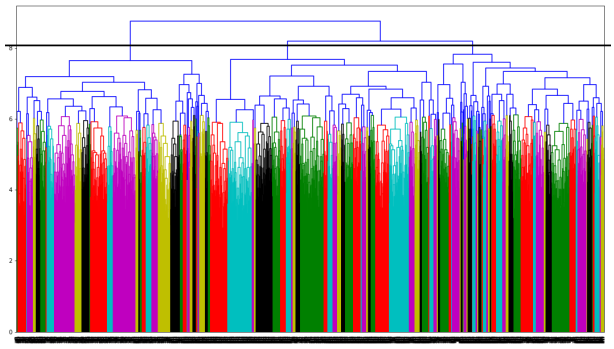
(d) Noroeste



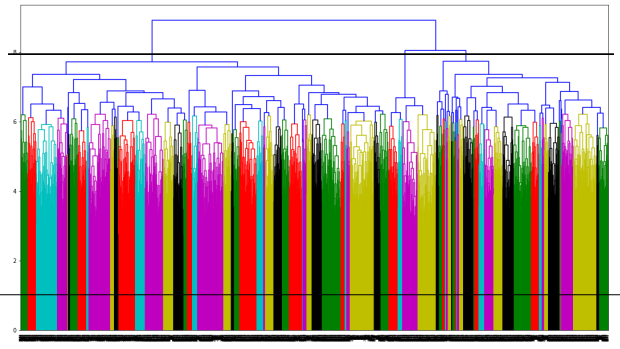
(e) Norte



(f) Occidente



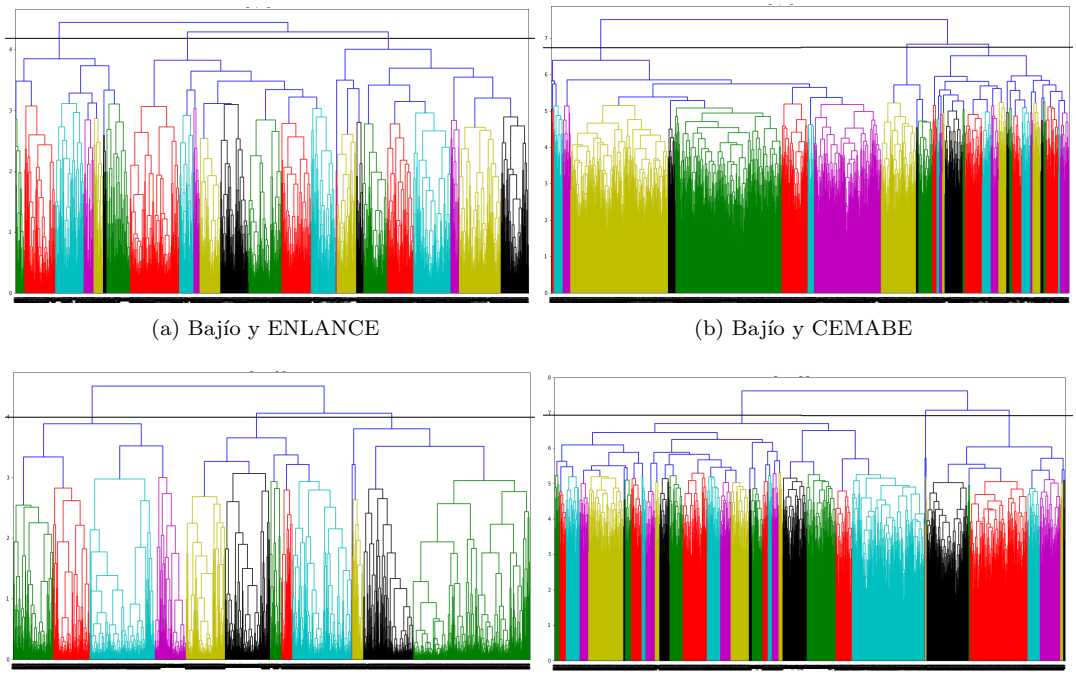
(g) Pacífico



(h) Sureste

Figura 5.5: Dendogramas de las regiones de México aplicada a secundarias

Como podemos observar de la figura 5.5, en todos los árboles se mantiene un corte de tres toques, por lo que $K = 3$ sigue siendo una buena elección. En el apéndice A están las regiones para primaria. Además de esto se hizo la clasificación por regiones y separando las bases de datos, en otras palabras clasificamos por regiones usando por separado la base de datos ENLANCE y CEMABE, ejemplo:



(a) Bajío y ENLANCE

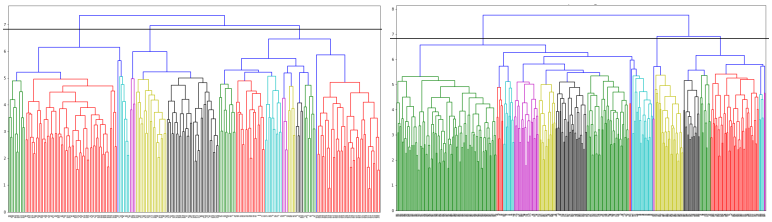
(b) Bajío y CEMABE

(c) Distrito Federal y ENLANCE

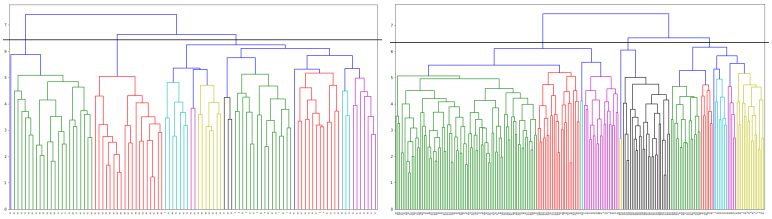
(d) Distrito Federal y CEMABE

Figura 5.6: Dendrogramas de las regiones de México aplicada a secundarias usando una sola base de datos.

De la figura 5.6, observamos que $K = 3$ se mantiene como una buena opción. En el apéndice A se muestra el resto de figuras. Continuando con el análisis la base de datos final se separó por los 32 estados de la república mexicana. En el apéndice A se muestra el resto de las figuras. Como podemos observar $K = 3$ fue constante a lo largo de los experimentos que hicimos, así $K = 3$ fue la elección que se hizo.



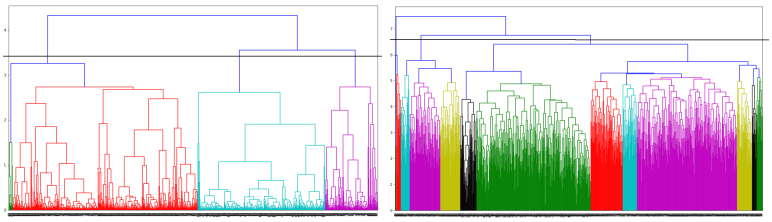
(a) Aguascalientes (b) Baja California Norte



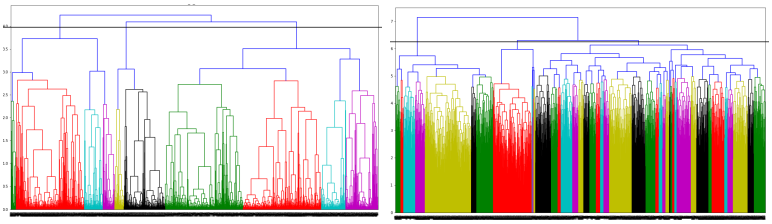
(c) Baja California Sur (d) Campeche

Figura 5.7: Dendrogramas de los estados de México para secundarias.

En el apéndice A se muestra el resto de figuras. Por último cada estado se separó por ENLANCE y CEMABE dando como resultado.



(a) Ciudad de México y ENLANCE (b) Ciudad de México y CEMABE



(c) México y ENLANCE (d) México y CEMABE

Figura 5.8: Dendrogramas de los estados de México para secundarias separados por ENLANCE y CEMABE.

.

5.1.3. Clasificando

Una vez que se decidió el número de *clusters*, procedimos a clasificar las escuelas. Primero aplicamos PCA a un 85 % del total de los datos, posteriormente usamos el algoritmo de $K-medias$ con $K = 3$, para primaria y secundaria. En la figura 5.9 y 5.10 se observa la clasificación de los datos para primaria y secundaria respectivamente, los ejes son las dos primeras componentes de PCA.

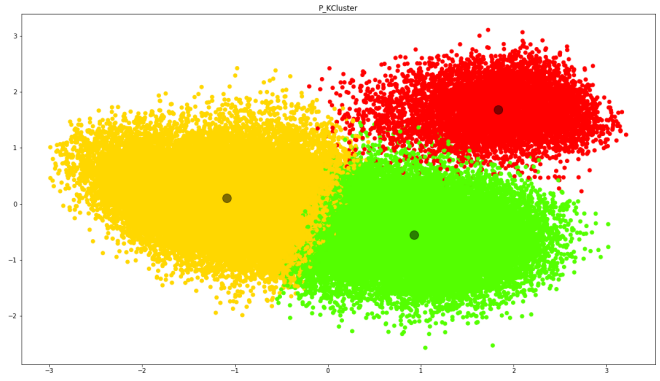


Figura 5.9: Clasificación de primarias usando las dos primeras componentes de PCA como ejes.

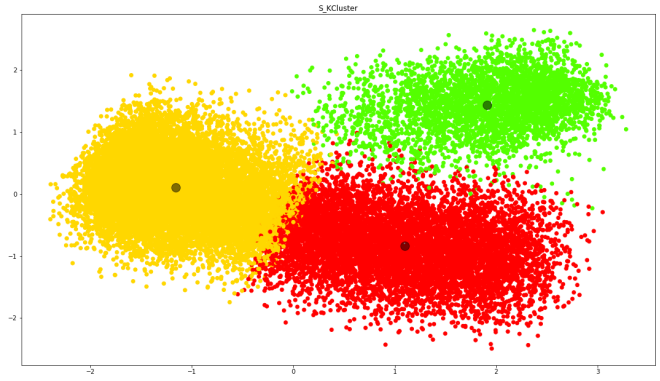


Figura 5.10: Clasificación de secundarias usando las dos primeras componentes de PCA como ejes.

Por lo que al final de este proceso a cada escuela se le asignó un número

entre 1 y 3, además clasificamos únicamente usando ENLANCE y CEMABE dando como resultado:

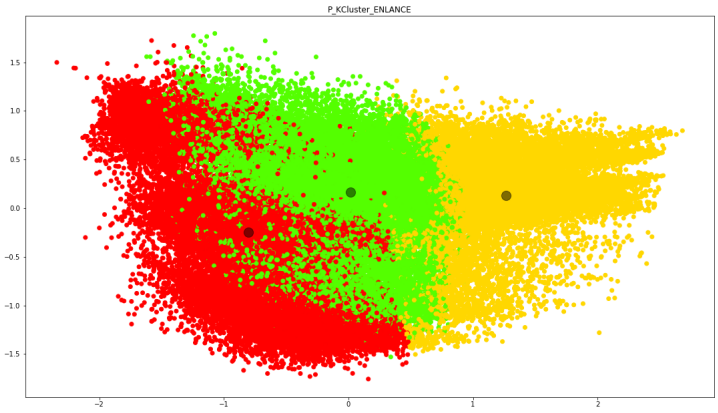


Figura 5.11: Clasificación de primarias con ENLANCE usando las dos primeras componentes de PCA como ejes.

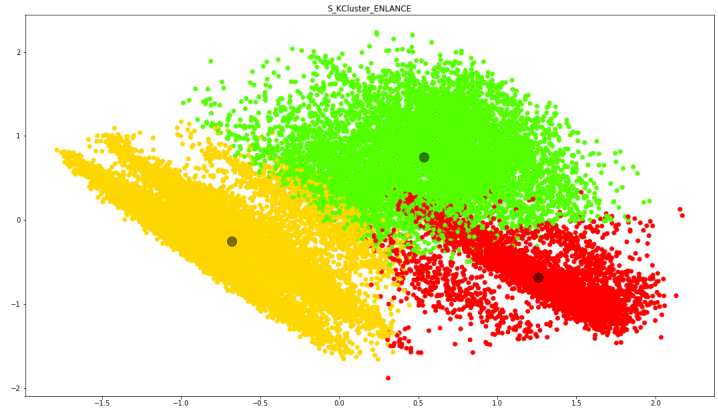


Figura 5.12: Clasificación de secundarias con ENLANCE usando las dos primeras componentes de PCA como ejes.

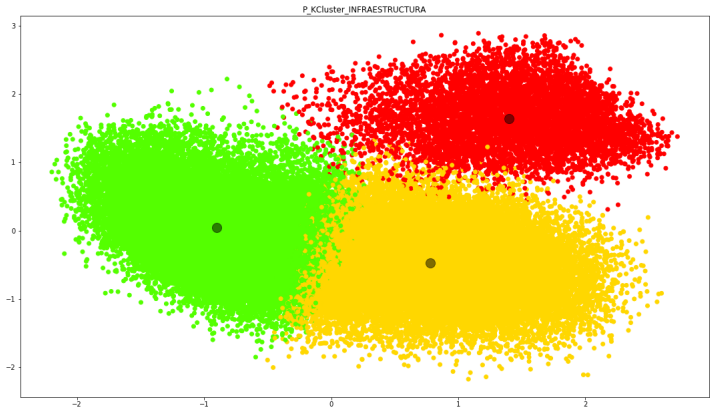


Figura 5.13: Clasificación de primarias con INFRAESTRUCTURA usando las dos primeras componentes de PCA como ejes.

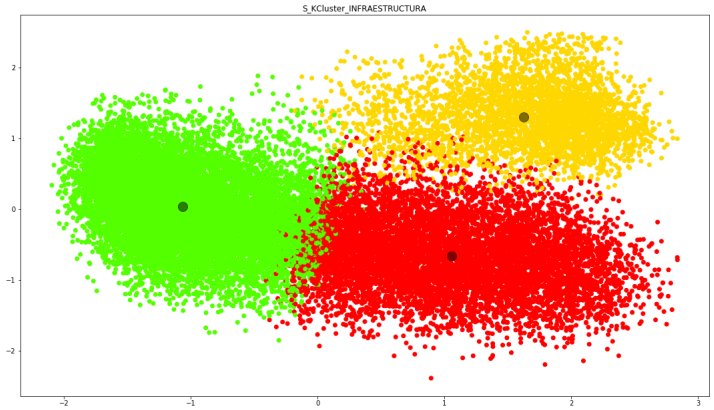


Figura 5.14: Clasificación de secundarias con INFRAESTRUCTURA usando las dos primeras componentes de PCA como ejes.

Notemos que las figuras mostradas son sólo una ayuda, debido a que el espacio de las bases de datos al aplicar PCA es de 25 variables, lo cual es imposible observar la gráfica para el humano, además notemos que hay $\binom{25}{2}$ imágenes posibles, por otro lado la clasificación no muestra que representa cada *cluster*, por lo que el color de las figuras solamente es representativo.

5.1.4. Determinando las clases

Como el algoritmo de $K - medias$ agrupó aquellas escuelas que son similares, necesitamos tener una mejor descripción de los *clusters*, ya que el algoritmo no muestra nada de las características que comparten éstos, así hicimos un estudio para poder determinar que representaban, recordemos que al escoger $K = 3$ hay tres clases; para este trabajo tomamos las clases como nivel 1, 2 y 3, siendo el nivel 1 bajo, 2 el medio y 3 alto.

Para determinar las características que hay en común de los *clusters* ocupamos el algoritmo Apriori, como éste sólo funciona con datos cuantitativos, transformamos los datos cualitativos de la forma que vimos en el Capitulo 2, para ello definimos los intervalos de la siguiente manera $I_0 = (-\infty, \mu - 2 * \sigma]$ $I_1 = [\mu - 2 * \sigma, \mu - \sigma]$ $I_2 = [\mu - \sigma, \mu]$ $I_3 = [\mu, \mu + \sigma]$ $I_4 = [\mu + \sigma, \mu + 2 * \sigma]$ y $I_5 = [\mu + 2 * \sigma, \infty)$ para cada columna que contuviera datos cualitativos. Así al final de este procedimiento aplicamos Apriori para cada *cluster*, dando como resultado:

Item	Soporte
1: { MEDIA-1-MATEMATICAS I_2 }	55.59 %
2: { LOGRO-MATEMATICAS-3-CERO I_3 }	56.01 %
3: { LOGRO-MATEMATICAS-2-CERO I_3 }	56.68 %
4: { LOGRO-MATEMATICAS-1-CERO I_3 }	57.14 %
5: { MEDIA-2-MATEMATICAS I_2 }	59.87 %
6: { ESCUELA GENERAL }	61.13 %
7: { LOGRO-ESPANOL-1-UNO I_3 }	63.66 %
8: { P269 1 }	64.84 %
9: { P207 1 }	67.69 %
10: { P260 0 }	77.06 %
11: { P268 1 }	91.46 %

Tabla 5.6: Tabla que muestra los resultados de Apriori para el *cluster* 1 de secundarias.

El resultado del algoritmo consta de un total de 152 items, en la tabla 5.6 está una muestra de los items, como podemos observar el *cluster* 1 está conformado por un total de 61.13 % de escuelas generales, por otro lado P207 1 (los alumnos tienen acceso a computadoras), recordando que 1 es sí y 0 es no, tiene un soporte de 67.69 %, P260 0 (la escuela no pertenece al programa nacional de inglés) tiene un soporte de 77.06 %, P269 (los alumnos tienen acceso a Internet) tiene un soporte de 64.84 % y P268 (la escuela tiene acceso a Internet) tiene un soporte de 91.46 %. Por otro lado viendo los items de la prueba ENLANCE, el item 1 que representa la media de matemáticas del primer año está en el intervalo I_2, en otras palabras el 55.59 % de las escuelas está entre una desviación estándar menos de la media y la media a nivel nacional para la prueba de matemáticas del primer año de secundaria, del item 5 el 59.87 % de las escuelas está entre una desviación estándar menos de la media y la media a nivel na-

cional para la prueba de matemáticas del segundo año de secundaria. Los item 2,3,4 que representan el logro de nivel cero obtenido en la prueba ENLANCE de matemáticas para el año 3,2 y 1 respectivamente, recordemos el logro cero según ENLANCE representa insuficiente, por lo que estos items representan que su grado de insuficiencia en matemáticas es mayor a la media para todos los años, para el item 7 el logro uno que representa elemental es mayor a la media.

Continuamos con el *cluster* 2:

Item	Soporte
1: { MEDIA-2-ESPAÑOL I_2 }	50.19 %
2: { MEDIA-1-MATEMATICAS I_2 }	50.86 %
3: { MEDIA-3-ESPAÑOL I_2 }	51.25 %
4: { GRADO-MARGINACION-ALTO }	52.12 %
5: { HORAS I_2 }	71.40 %
6: { P268 0 }	72.64 %
7: { P207 0 }	76.99 %
8: { P269 0 }	79.09 %
9: { TELESECUNDARIA }	93.02 %
10: { P260 0 }	97.74 %

Tabla 5.7: Tabla que muestra los resultados de Apriori para el *cluster* 2 de secundarias.

EL resultado del algoritmo consta de un total de 158 items, en la tabla 5.7 está una muestra de los items, como podemos observar el *cluster* 2 está conformado por un total de 93.02 % de escuelas telesecundarias, del item 10 el 97.74 % no pertenece al programa nacional de inglés, el item 6 (la escuela no tiene acceso a Internet) tiene un soporte de 72.64 %, el item 8 (los alumnos no tienen acceso a Internet) tiene un soporte de 79.09 %, del item 207 (los alumnos no tienen acceso a computadoras) tiene un soporte de 76.99 %. Por otro lado los items 1,2,3 están por debajo de la media nacional para sus respectivos campos, notemos además del item 4 que el 52.12 % de las secundarias tienen un grado de marginación alto, donde según el Consejo Nacional de Población (CONAPO) el índice de marginación es una medida-resumen que permite diferenciar localidades del país según el impacto global de las carencias que padece la población como resultado de la falta de acceso a la educación, la residencia en viviendas inadecuadas y la carencia de bienes, por lo que entre mayor sea el grado de marginación mayor es la carencia y por último del item 5 el 71.40 % de las escuelas están abiertas menos de la media nacional, por lo tanto los alumnos están entre 5 y 7 horas en la escuela.

Continuamos con el *cluster* 3:

Item	Soporte
1: { MEDIA-1-MATEMATICAS I_3 }	53.91 %
2: { MEDIA-3-MATEMATICAS I_3 }	55.52 %
3: { LOGRO-ESPANOL-3-CERO I_1 }	60.88 %
4: { P260 0 }	78.91 %
5: { P269 1 }	82.33 %
6: { LOGRO-FORMACION-3-CERO I_2 }	86.62 %
7: { GRADO-MARGINACION-MUY-BAJO }	87.55 %
8: { P207 1 }	94.72 %
9: { PARTICULAR }	98.46 %
10: { P268 1 }	98.93 %

Tabla 5.8: Tabla que muestra los resultados de Apriori para el *cluster* 3 de secundarias

EL resultado del algoritmo consta de un total de 147 items, en la tabla 5.8 está una muestra de los items, como podemos observar el *cluster* 3 está conformado por un total de 98.46 % de escuelas particulares, del item 4 el 78.91 % no pertenece al programa nacional de inglés, el item 10 (la escuela tiene acceso a Internet) tiene un soporte de 98.93 %, el item 5 (los alumnos tienen acceso a Internet) tiene un soporte de 82.33 %, del item 8 (los alumnos tienen acceso a computadoras) tiene un soporte de 94.72 %. Por otro lado los items 1,2 están encima de la media nacional para sus respectivos campos, notemos además del item 3 que el 60.88 % su logro en español de nivel cero está dos desviaciones debajo de la media.

Entonces al hacer la comparación de los *clusters* tenemos:

Cluster 1 (Nivel 2)	Cluster 2 (Nivel 1)	Cluster 3 (Nivel 3)
El logro de matemáticas de nivel cero para todos los años está por encima de la media nacional.	El 71.40% de las escuelas tienen Un horario menor a la media nacional	El logro de español y formación para el año 3 de nivel cero está por encima de la media nacional.
La media de matemáticas del año 1 y 2 está por debajo de la media nacional.	La media de español para el año 2 y 3 está por debajo de la media nacional así como la media de matemáticas para el año 1.	La media de matemáticas para el 1 y 3 esa por encima de la media
El 61.13% son escuelas públicas	El 93.02% son telesecundarias.	El 98.46% son escuelas privadas.
El 91.46% de las escuelas tiene acceso a internet.	El 72.64% de las escuelas no tiene acceso a Internet.	El 98.93% de las escuelas tienen acceso a Internet.
El 64.84% de las escuelas da acceso de Internet a los alumnos.	El 79.09% de las escuelas no da acceso de Internet a los alumnos.	El 82.33% de las escuelas da acceso de Internet a los alumnos
El 77.06% de las escuelas no pertenece al programa nacional de inglés.	El 97.74% de las escuelas no pertenece al programa nacional de inglés.	El 78.91% de las escuelas no pertenece al programa nacional de inglés.
El 67.69% de las escuelas tienen computadoras para los alumnos.	El 76.99% de las escuelas no tienen computadoras para los alumnos.	El 82.33% de las escuelas tienen computadoras para los alumnos.

Tabla 5.9: Comparación de los *clusters* de las secundarias

De la tabla 5.9 tomamos los *cluster* 1, *cluster* 2 y *cluster* 3 como nivel 2, 1 y 3 respectivamente. Volvemos a graficar la figura 5.10, con los colores rojo, amarillo y verde para los niveles 1, 2 y 3 respectivamente.

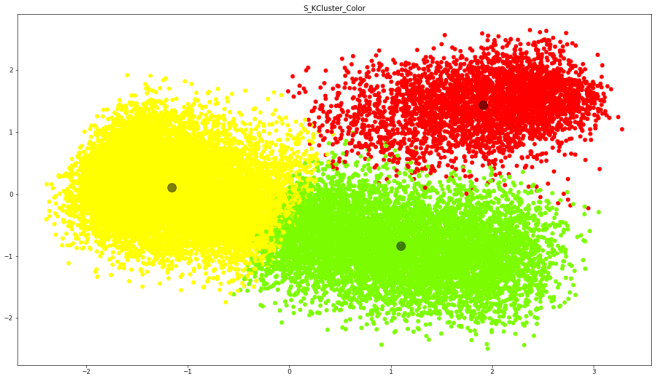


Figura 5.15: Clasificación de secundarias usando las dos primeras componentes de PCA como ejes y aplicando su respectivo color.

Al aplicar el mismo análisis que mostramos con anterioridad a primarias, obtenemos la siguiente comparación:

Cluster 1 (Nivel 1)	Cluster 2 (Nivel 3)	Cluster 3 (Nivel 2)
El 50.56% de las escuelas tiene un grado de marginación alto.	El logro de matemáticas de nivel cero para todos los años está por debajo de la media nacional.	El logro de matemáticas de nivel cero para todos los años esta por debajo de la media nacional.
La media de matemáticas del año 5 y 6 está por debajo de la media nacional.	La media de matemáticas para el año 3, 5 y 6 está por encima de la media nacional	La media de todas las meterias de todos lo años está por encima de la media nacional.
El 85.72% son escuelas públicas	El 99.20% son escuelas privadas.	El 99.62% son escuelas públicas.
El 85.62% de las escuelas no tiene acceso a internet.	El 97.09% de las escuelas tiene acceso a Internet.	El 84.64% de las escuelas tienen acceso a Internet.
El 92.09% de las escuelas no da acceso de Internet a los alumnos.	El 71.73% de las escuelas da acceso de Internet a los alumnos.	El 54.15% de las escuelas no da acceso de Internet a los alumnos
El 95.77% de las escuelas no pertenece al programa nacional de inglés.	El 80.97% de las escuelas no pertenece al programa nacional de inglés.	El 70.23% de las escuelas no pertenece al programa nacional de inglés.
El 92.13% de las escuelas no tienen computadoras para los alumnos.	El 89.39% de las escuelas tienen computadoras para los alumnos.	El 58.69% de las escuelas no tienen computadoras para los alumnos.

Tabla 5.10: Comparación de los *clusters* de las primarias

De la tabla 5.10 tomamos los *cluster* 1, *cluster* 2 y *cluster* 3 como nivel 1, 3 y 2 respectivamente. Volvemos a graficar la figura 5.9, con los colores rojo, amarillo y verde para los niveles 1, 2 y 3 respectivamente.

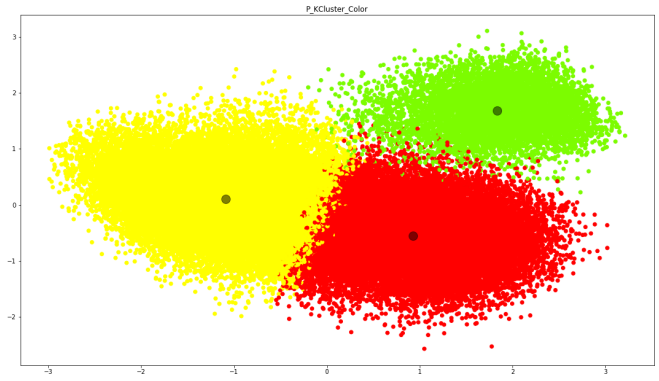


Figura 5.16: Clasificación de primaria usando las dos primeras componentes de PCA como ejes y aplicando su respectivo color.

Por último se hizo el mismo análisis para las figuras 5.11, 5.13, 5.12 y 5.14.

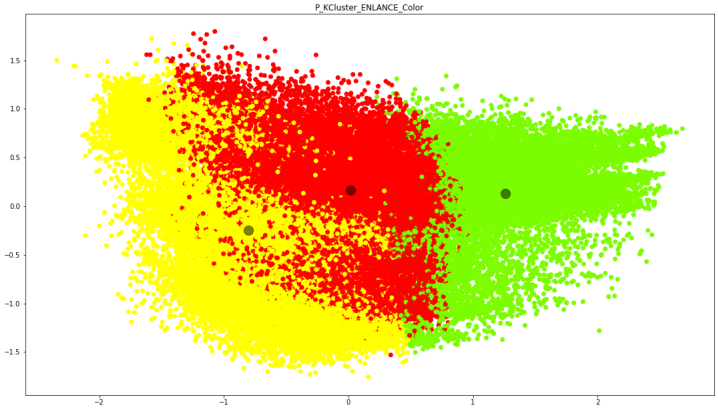


Figura 5.17: Clasificación de primarias con ENLANCE usando las dos primeras componentes de PCA como ejes, donde los *clusters* 1, 2 y 3 se reasignaron como nivel 3, 1 y 2 respectivamente.

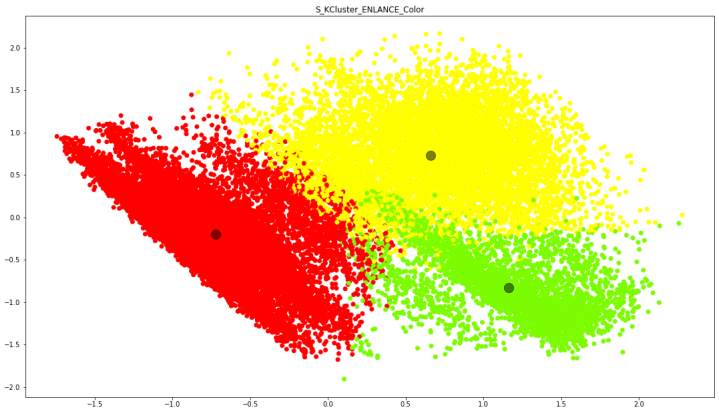


Figura 5.18: Clasificación de secundarias con ENLANCE usando las dos primeras componentes de PCA como ejes, donde los *clusters* 1, 2 y 3 se reasignaron como nivel 3, 1 y 2 respectivamente.

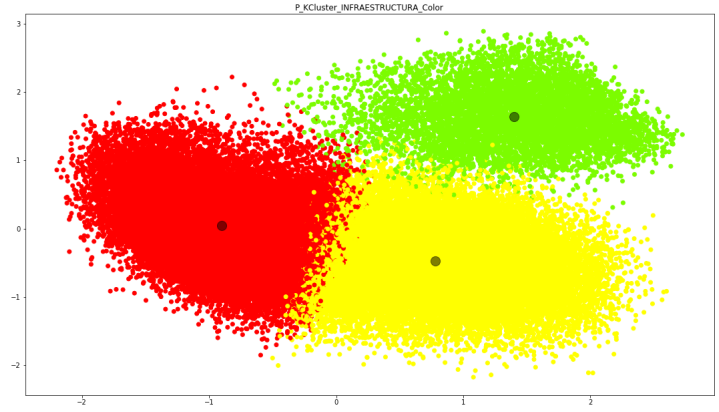


Figura 5.19: Clasificación de primarias con INFRAESTRUCTURA usando las dos primeras componentes de PCA como ejes, donde los *clusters* 1, 2 y 3 se reasignaron como nivel 3, 2 y 1 respectivamente.

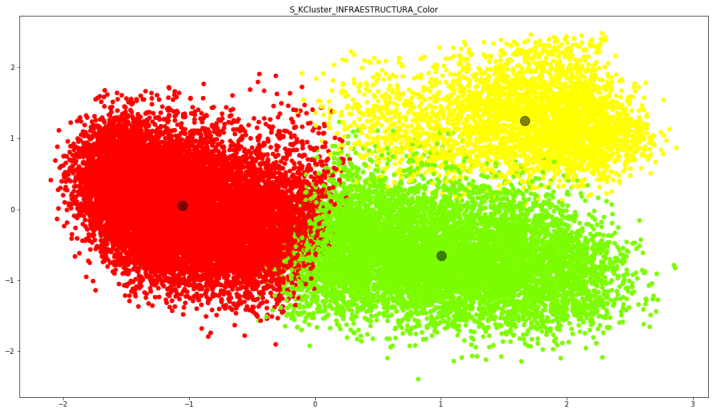


Figura 5.20: Clasificación de secundarias con INFRAESTRUCTURA usando las dos primeras componentes de PCA como ejes, donde los *clusters* 1, 2 y 3 se reasignaron como nivel 2, 3 y 1 respectivamente.

Capítulo 6

Conclusiones

A lo largo de este trabajo mostramos, los pasos necesarios que se llevaron acabo para poder determinar la calidad educativa de las escuelas de nivel básico en México; notemos que los algoritmos y todas las conclusiones se hicieron con la unión de la bases de ENLANCE y CEMABE (*datamart*) y como a partir de esta unión se hizo la transformación y limpieza de los datos, por lo que si una nueva escuela se quisiera clasificar tendría, que tener el mismo formato que el *datamart*.

Notemos que todo el trabajo cambiaría si se hiciera con un K diferente, la elección de $K = 3$ fue con base a los dendrogramas mostrados, aunque $K = 4$ e incluso $K = 5$ podrían ser una buena opción (debido a que éstas también fueron constantes en todos los dendrogramas) aunque la interpretación se haría más complicada, ya que por ejemplo con $K = 3$ en general las escuelas se separaron en escuelas publicas con grado de marginación alto, en escuelas publicas y en escuelas privadas, al aumentar K la separación seria más fina por lo que se podría perder información, notemos además que al aumentar K cada escuela se estaría clasificando en un sólo *cluster*, por lo que para un valor de K muy grande cada escuela sería su propio *cluster*.

Se tomaron tres clasificaciones las cuales fueron: la clasificación donde se unió la información de la prueba ENLANCE y CEMABE, donde ésta es la más importante debido a que en esta clasificación se tomaron los parámetros del desempeño académico de los alumnos y la infraestructura de la escuela, la segunda y la tercer clasificación fueron usando únicamente la información ENLANCE y CEMABE respectivamente; ahora notemos que la elección de los niveles que hicimos fue con base a las características en común que compartían los *clusters*, así el nivel 1 fue el nivel más bajo, debido a que fue el *cluster* que en promedio tuvo menor desempeño en la prueba ENLANCE y su infraestructura era más carente, además de que tanto para primaria y secundaria más del 50 % de las escuelas tenían un grado de marginación alto, ahora haciendo la comparación del nivel 2 con el 3, tenemos que en la prueba ENLANCE se comportaron de manera similar, en donde se mostró la diferencia fue en la infraestructura, esto

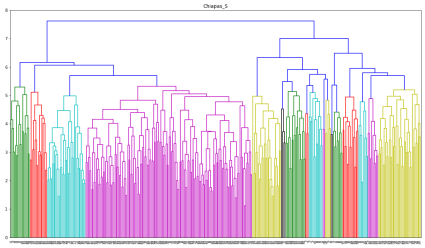
se debe a que un *cluster* tanto para primaria y secundaria estaba conformado por su mayoría de escuelas privadas. Al ver esto observamos que en general hay similitud entre las tres clasificaciones, es decir que para unas escuelas sus tres clasificaciones Unión, ENLANCE y CEMABE por ejemplo son de la forma nivel 2, nivel 2 y nivel 2 respectivamente, esto se puede deber a que las bases tienen la suficiente correlación, es decir una escuela que la clasificación de CEMABE haya sido clasificada como nivel 1, no puede obtener un nivel 3 en ENLANCE, debido a que la infraestructura son variables de importancia para que los alumnos tengan un mayor desempeño, por ejemplo una escuela con grado de marginación alto, sin Internet y siendo telesecundaria es muy improbable que obtenga un nivel 3 en la prueba ENLANCE.

El presente trabajo puede funcionar como un método de consulta, en donde por ejemplo se pondría información de la escuela, tipo de escuela, grado de marginación, los tres niveles de clasificación que se obtuvieron, el turno, etc, entonces cualquier persona interesada puede conocer la información de las escuelas que sean de su interés. Además puede servir para la toma de decisiones, para programas como escuela DIGNA, ya que se muestra las escuelas que necesitan más atención. Por otro lado si quisiéramos clasificar nuevas escuelas, entonces podríamos hacer un proceso de clasificación supervisada tomando como referencia el presente trabajo.

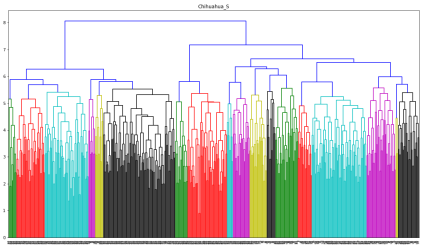
Hicimos distinción entre primarias y secundarias, ésto se hizo así ya que no son iguales, tanto los lineamentos para la prueba ENLANCE y CEMABE son distintos, por lo que hacer una clasificación en donde estén juntas seria un error. Como posible mejora a la clasificación, podríamos obtener información del nivel socioeconomico de la zona en donde están las escuelas o información referente a la economía de la zona, ya que la intuición nos dice que una escuela en una de la zona más exclusiva de México es mejor que una en la zona más marginales.

Apéndice A

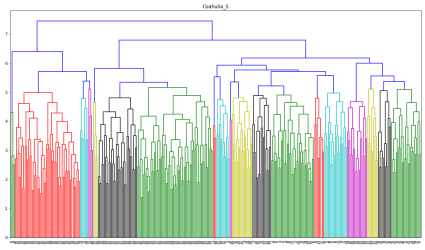
Dendrogramas



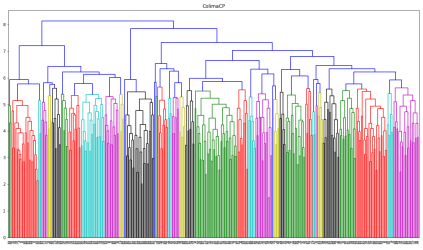
Chiapas



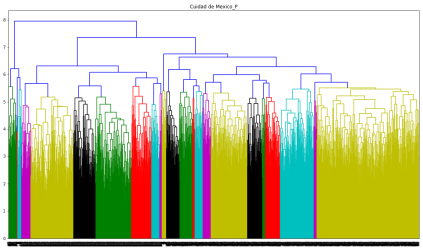
Chihuahua



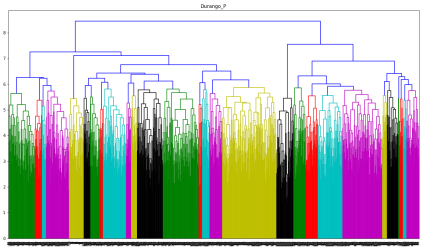
Coahulia



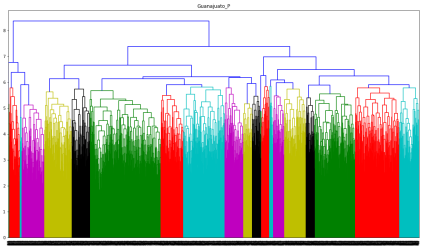
Colima



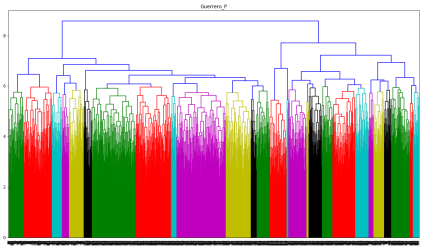
Ciudad de México



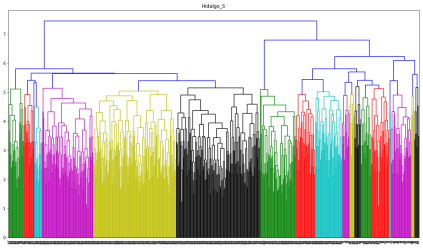
Durango



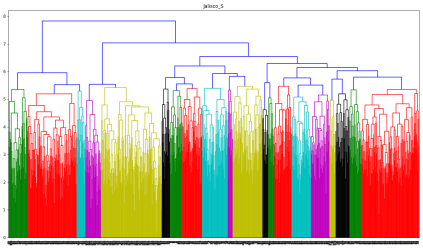
Guanajuato



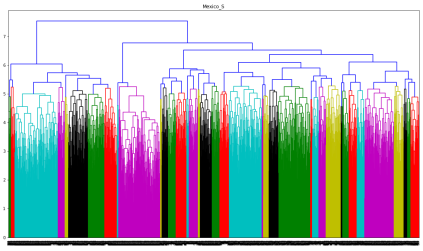
Guerrero



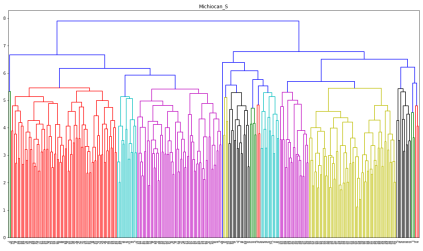
Hidalgo



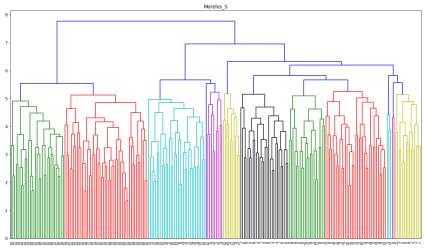
Jalisco



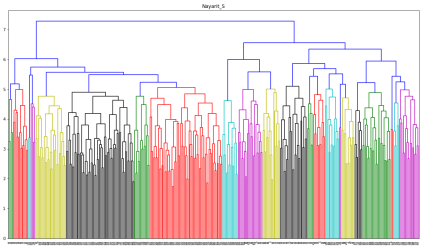
México



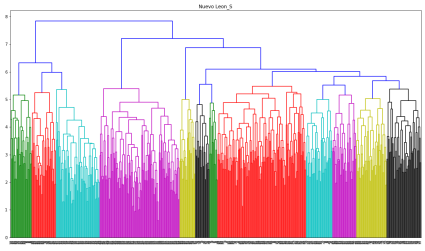
Michoacán



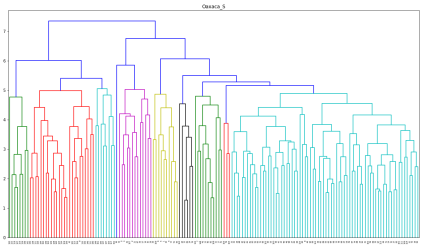
Morelos



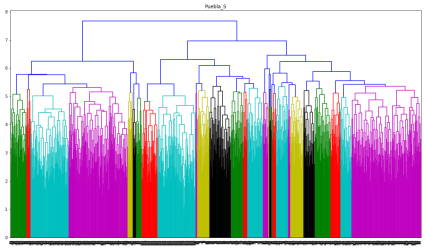
Nayarit



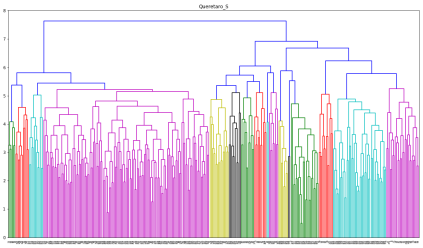
Nuevo León



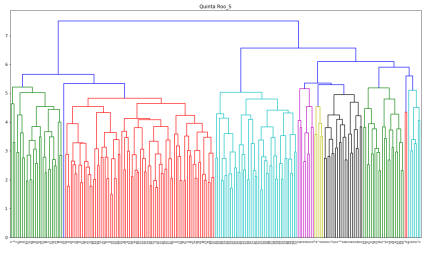
Oaxaca



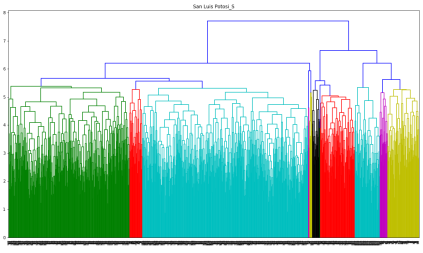
Puebla



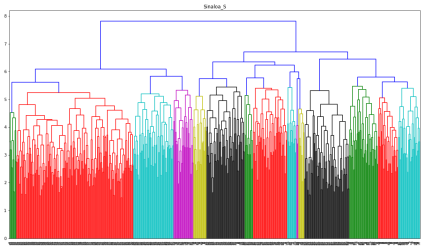
Querétaro



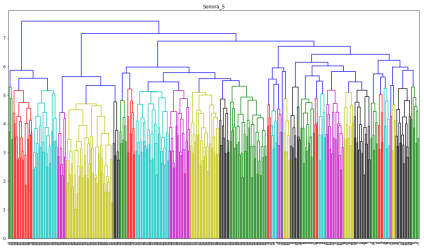
Quintana Roo



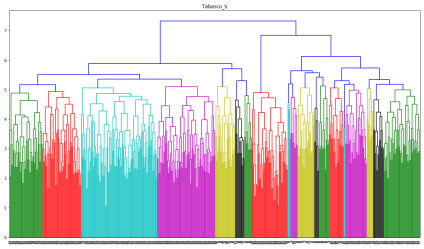
San Luis Potosí



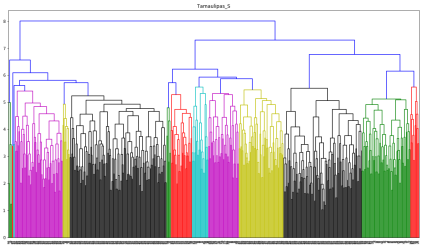
Sinaloa



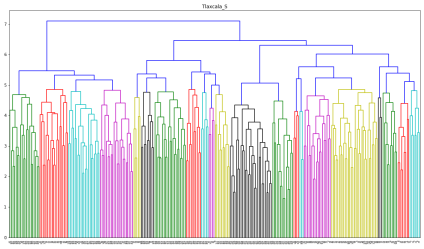
Sonora



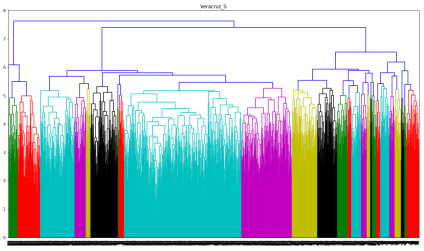
Tabasco



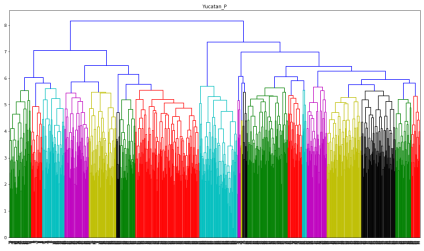
Tamaulipas



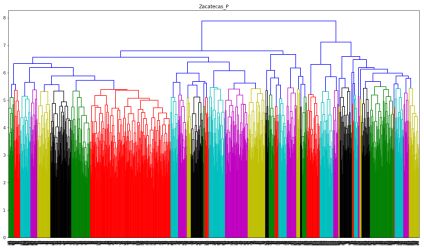
Tlaxcala



Veracruz

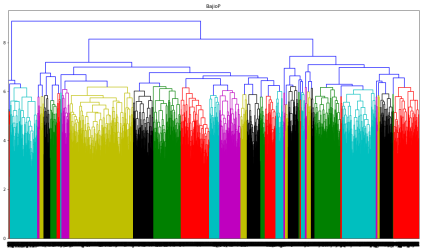


Yucatán

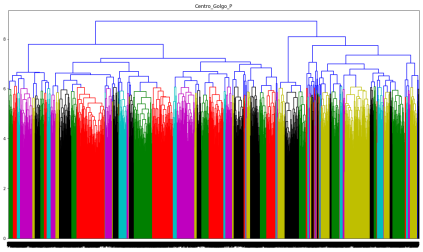


Zacatecas

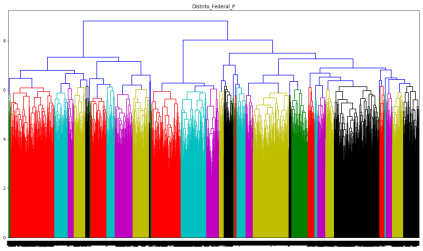
Figura A.1: Dendrogramas de la república mexicana aplicadas a secundarias.



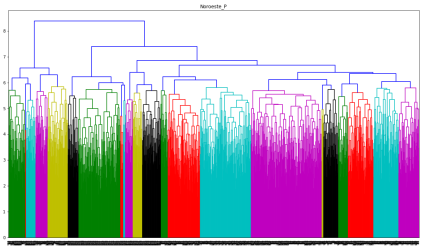
Bajío



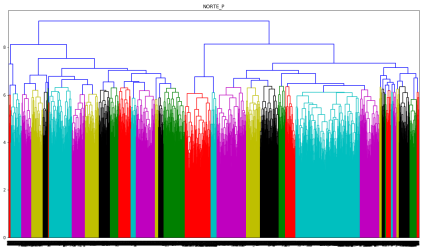
Centro y golfo



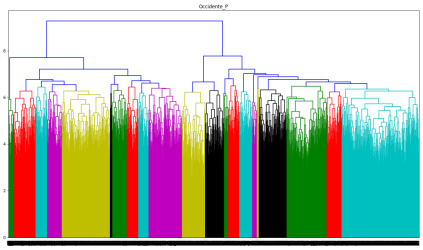
Distrito Federal



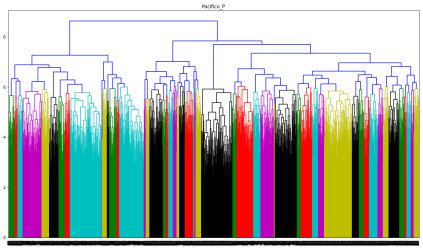
Noroeste



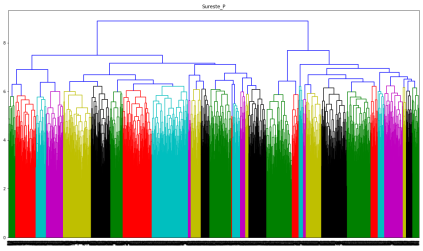
Norte



Occidente

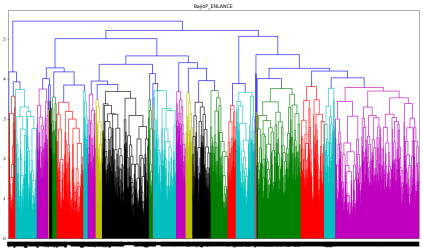


Pacífico

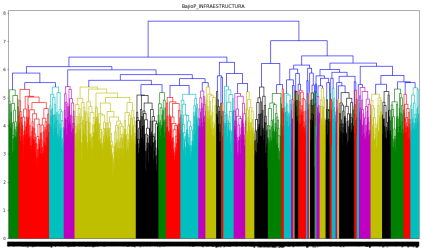


Sureste

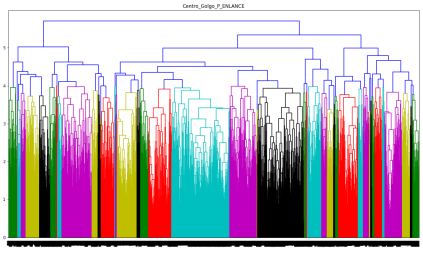
Figura A.2: Dendrogramas de las regiones de México aplicadas a primarias.



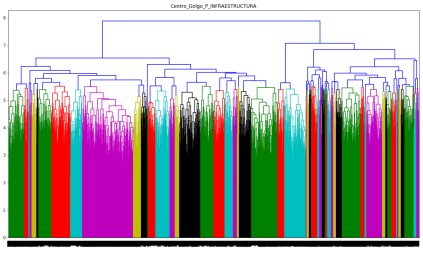
Bajío y ENLANCE



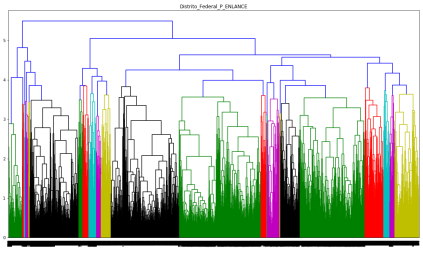
Bajío y CEMABE



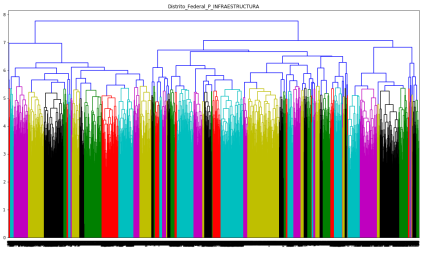
Centro golfo y ENLANCE



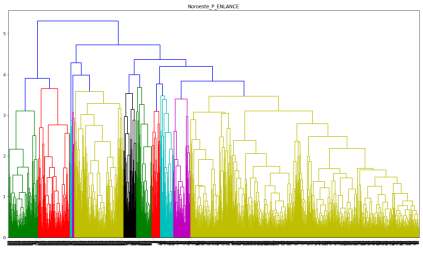
Centro golfo y CEMABE



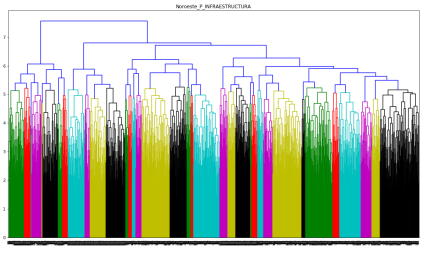
Distrito Federal y ENLANCE



Distrito Federal y CEMABE

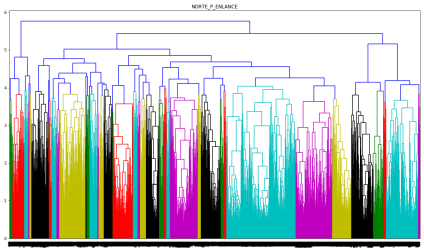


Noroeste y ENLANCE

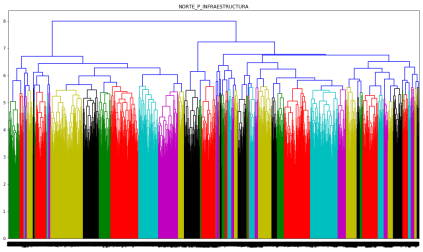


Noroeste y CEMABE

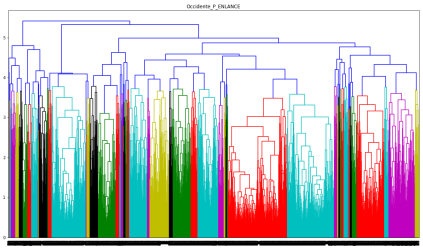
Figura A.3: Dendrogramas de las regiones de México aplicadas a primarias separadas por ENLANCE y CEMABE.



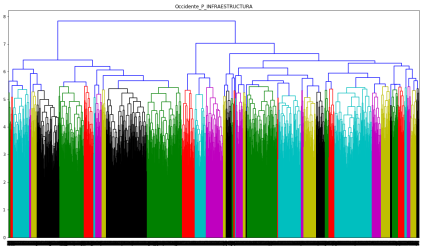
Norte y ENLANCE



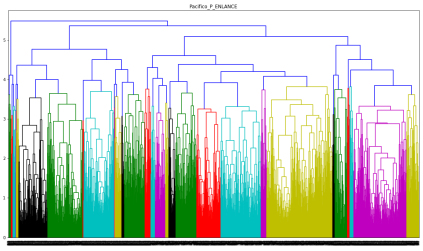
Bajío y CEMABE



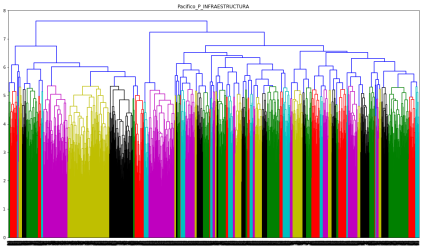
Occidente y ENLANCE



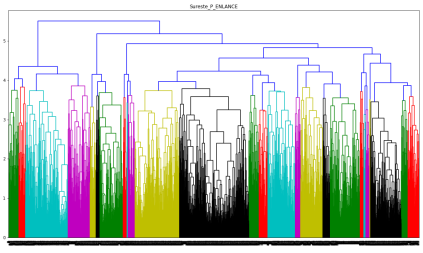
Occidente y CEMABE



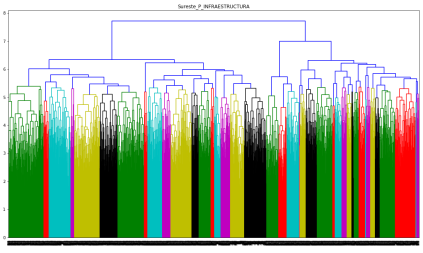
Pacifico y ENLANCE



Pacifico y CEMABE

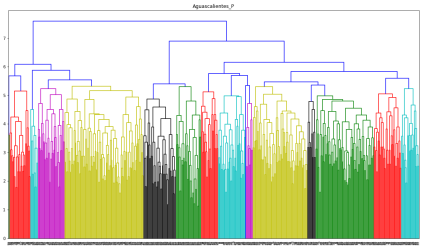


Sureste y ENLANCE

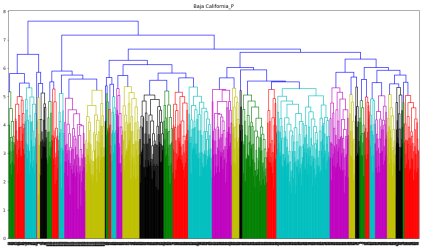


Sureste y CEMABE

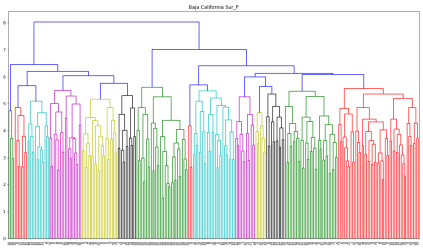
Figura A.4: Dendrogramas de las regiones de México aplicadas a primarias separadas por ENLANCE y CEMABE.



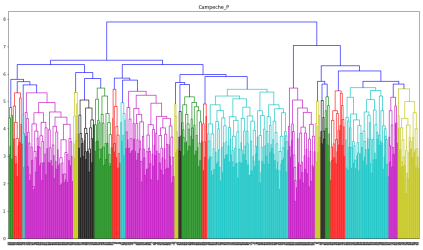
Aguascalientes



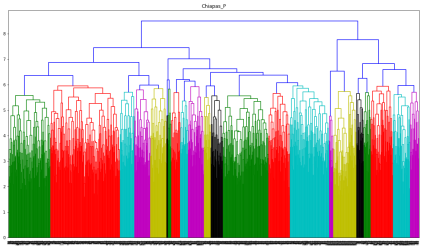
Baja California Norte



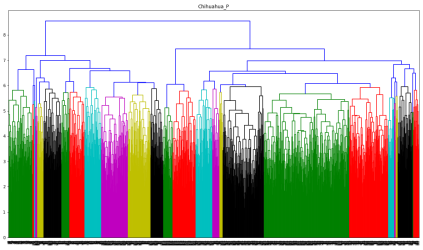
Baja California Sur



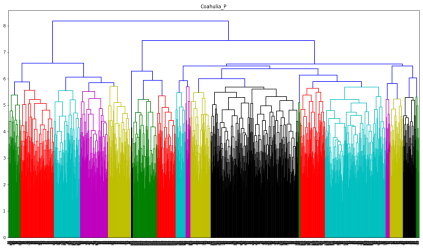
Campeche



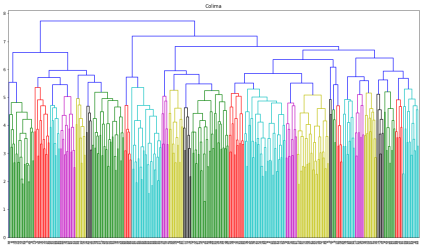
Chiapas



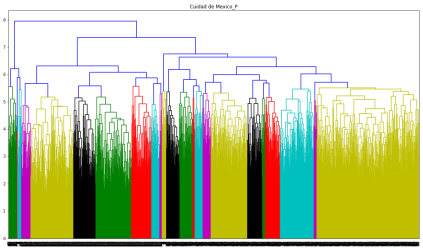
Chihuahua



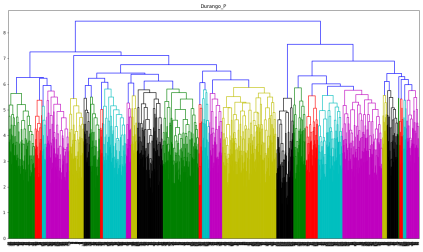
Coahulia



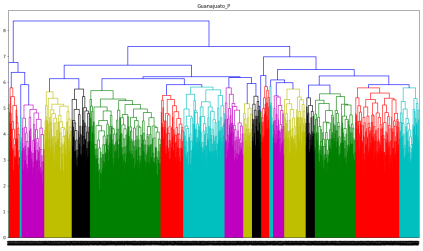
Colima



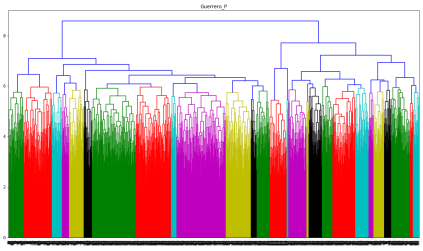
Ciudad de México



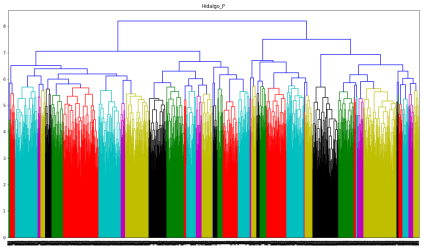
Durango



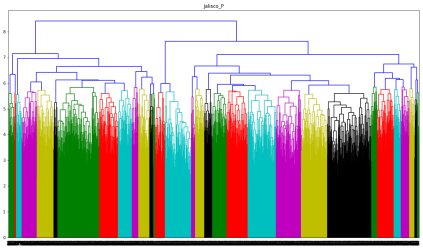
Guanajuato



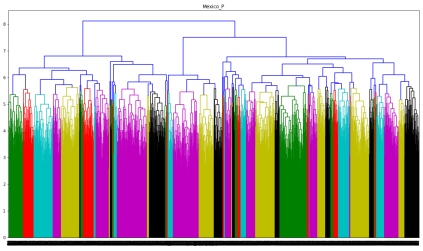
Guerrero



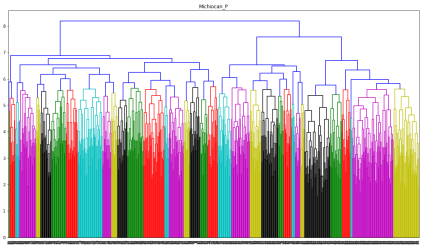
Hidalgo



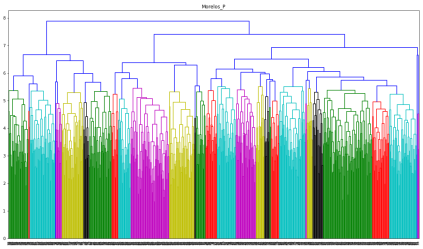
Jalisco



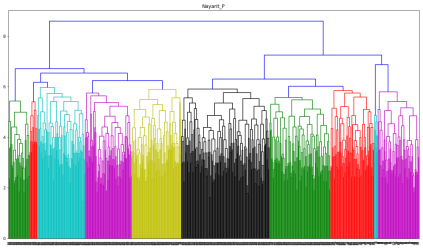
México



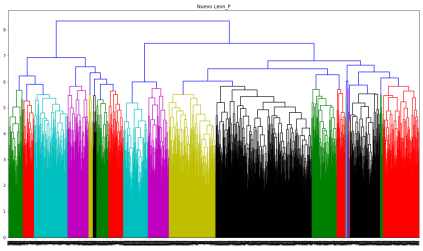
Michoacán



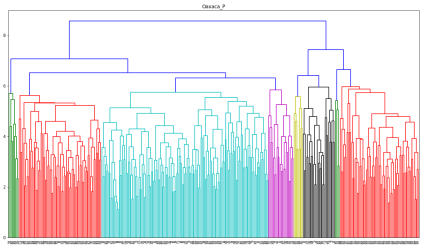
Morelos



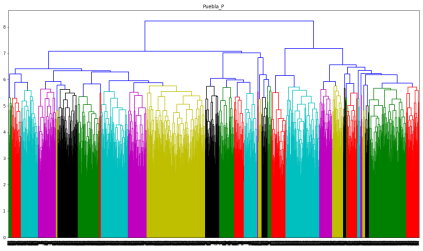
Nayarit



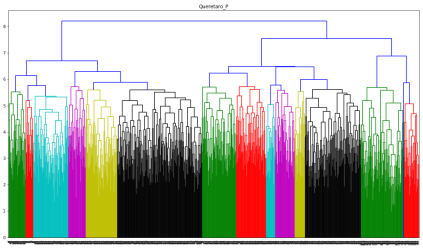
Nuevo León



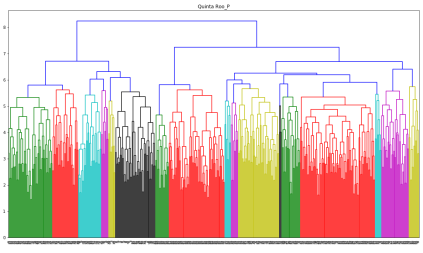
Oaxaca



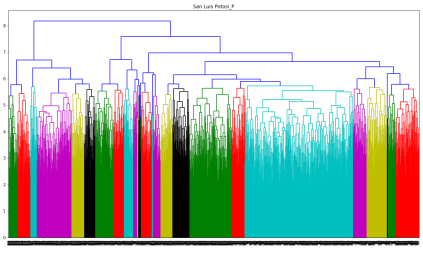
Puebla



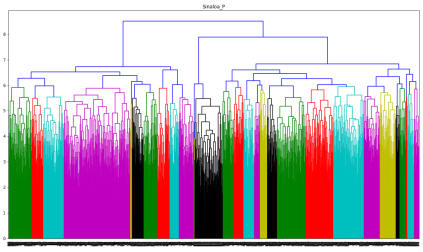
Querétaro



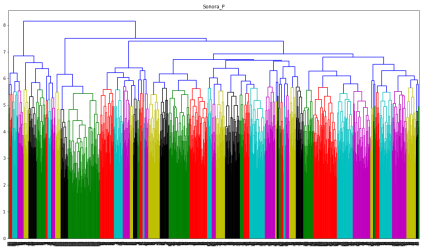
Quintana Roo



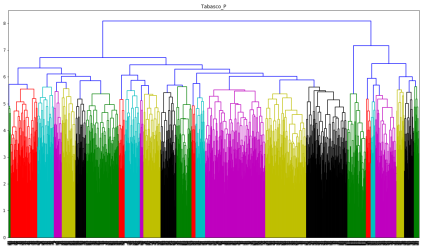
San Luis Potosí



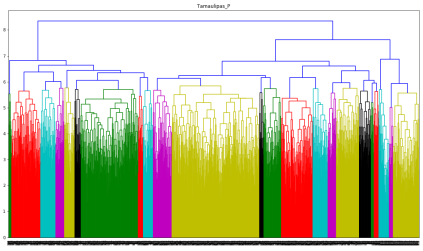
Sinaloa



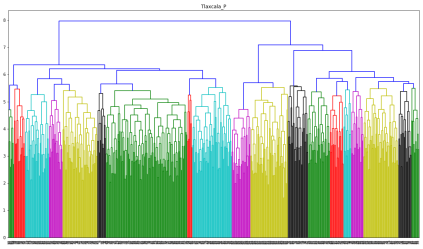
Sonora



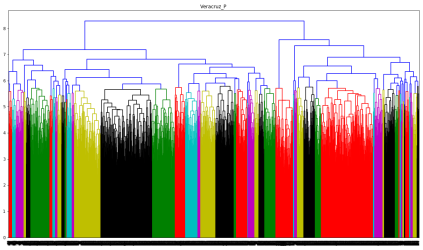
Tabasco



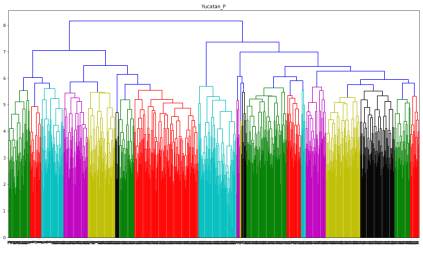
Tamaulipas



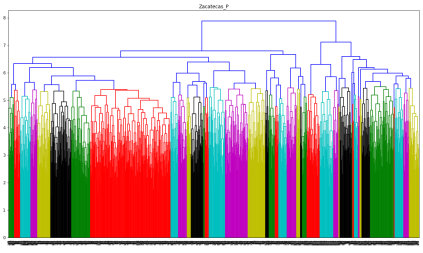
Tlaxcala



Veracruz



Yucatán



Zacatecas

Figura A.5: Dendrogramas de la república mexicana aplicadas a primarias.

Bibliografía

[Abraham Sillberschatz(2002)] S. Sudarshan. Abraham Sillberschatz, Henry F. Korth. *Fundamentos de Bases de Datos*. Mc Graw Hill, 2002.

[Andreas C.Muller(2017)] Sarah Guido. Andreas C.Muller. *Introduction to Machine Learning with Python*. O'REILLY, 2017.

[de Jonge and van der Loo(2013)] Edwin de Jonge and Mark van der Loo. An introduction to data cleaning with r. 2013.

[de Smith(2018)] Michael J de Smith. *Statistical Handbook Handbook*. The Winchelsea Press, 2018.

[DGEP(2013)] DGEP. Evaluación nacional del logro académico en centros escolares. 2013.

[Dos.(2013)] Izzet Dos. Some model suggestions for measuring effective schools. *Procedia*, 2013.

[Frederick.(1987)] Judith M. Frederick. Measuring school effectiveness: Guidelines for educational paractitioners. *Wareham Public Schools*, 1987.

[Gamboa and Bonals(2016)] Luis Alan Acuna Gamboa and Leticia Pons Bonals. Calidad educativa en méxico. de las disposiciones internacionales a los remiendos del proyecto nacional. 2016.

[Gareth James(2013)] Trevor Hastie Robert Tibshirani. Gareth James, Daniela Witten. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[Ghahramani.(2004)] Zoubin Ghahramani. Unsupervised learning. *University College London*, 2004.

[Hastie and Tibshirani(2016)] Trevor Hastie and Rob Tibshirani. Statistical learning, 2016. URL `https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016`. [Online; accessed Jun 28, 2016].

[Hegland(2008)] Markus Hegland. The apriori algorithm-a tutorial. 2008.

[INEGI(2014)] INEGI. Censo de escuelas, maestros y alumnos de educación básica y especial. 2014.

[José Hernandez Orallo(2004)] Cesar Ferri. José Hernandez Orallo, José Ramirez. *Introducción a la Míneria de Datos*. Pearson, 2004.

[Martha E(2016)] Gómez Collado Martha E. Panorama del sistema educativo mexicano desde la perspectiva de las políticas públicas. 2016.

[Masters.(2012)] Geoff N Masters. Measuring and rewarding school improvement. *Australian Council*, 2012.

[Migrant and Start(2006)] Migrant and Seasonal Head Start. *Introduction to Data Analysis Handbook*. Academy for Educational Development, 2006.

[Shai Shalev Shwartz(2014)] Shai Ben David. Shai Shalev Shwartz. *Understanding Machine Learning from Theory to Algorithms*. Cambridge University Press, 2014.

[Trevor Hastie(2001)] Jerome Friedman. Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.

[Vargas(2015)] Luz Adriana Hernández Vargas. *Selección de la metodología para determinar atípicos en las bases de cálculo de un índice de costos*. FUNDACION UNIVERSITARIA LOS LIBERTADORES, 2015.