INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DE DESEMPEÑO ACADÉMICO EN ESCUELAS PRIMARIAS DE MÉXICO

TESIS

QUE PARA OBTENER EL TÍTULO DE

INGENIERA EN COMPUTACIÓN

P R E S E N T A

PAOLA MEJÍA DOMENZAIN

ASESOR: M.C. JUAN SALVADOR MARMOL

CIUDAD DE MÉXICO 2019

«Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Diseño e implementación de un sistema de predicción de desempeño escolar**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.»

Fecha

Paola Mejia Domenzain

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ACRÓNIMOS

CAM Centro de Atención Múltiple. 35, 40

CCT Clave de Centro de Trabajo. 13, 22, 35, 42, 54

CNRMMCE Centro Nacional para la Revalorización del Magisterio y la Mejora Continua de la Educación. 5

CONAFE Consejo Nacional de Fomento Educativo. 25

CRISP-DM Cross-industry standard process for data mining. 1, 2

csv comma-separated values. 23

ENLACE Evaluación Nacional de Logro Académico en Centros Escolares. 9, 11, 14, 21

EXCALE Exámenes de la Calidad y el Logro Educativo. 9, 21

F911 Formato Estadístico 911. 9, 10, 20, 23, 40, 48

INEE Instituto Nacional para la Evaluación de la Educación. 5, 21

INEGI Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial. 10, 20, 22, 24, 25, 29, 35, 40, 48

INEGI Instituto Nacional de Estadística y Geografía. 9

KDD Knowledge Discovery in Databases Framework. 1, 2

OCDE Organización para la Cooperación y el Desarrollo Económico. 5

OSL Mínimos cuadrados ordinarios. 50

PISA Informe del Programa Internacional para la Evaluación de Estudiantes. 5, 8, 9, 21

Planea Plan Nacional para la Evaluación de los Aprendizajes. 9, 21

SEMMA Sample, Explore, Modify, Model, and Assess. 1, 2

SEP Secretaría de Educación Pública. 9, 14, 15

CAPÍTULO 1

INTRODUCCIÓN

La minería de datos es un dominio de la ciencia de la computación que permite el análisis de grandes cantidades de datos para encontrar y extraer patrones significativos útiles para el proceso de la toma de decisiones [1].

Este es un proyecto de minería de datos utilizando información de la educación en México.

1.0.1 Posibles metodologías

Existen varías metodología alternativas para resolver un problema de minería de datos. Entre ellas destacan las metodologías Sample, Explore, Modify, Model, and Assess (SEMMA), Knowledge Discovery in Databases Framework (KDD) y Cross-industry standard process for data mining (CRISP-DM) por su popularidad y aplicación en varias industrias. Más adelante, se describen brevemente.

Sample, Explore, Modify, Model, and Assess

La metodología utilizada por la compañia SAS para análisis de datos se llama "SEMMA"por sus siglas en inglés (Sample, Explore, Modify, Model, and Assess). La característica principal de la metodología es que los diferentes pasos se manejar con nodos. El primer paso es seleccionar diferentes muestras para después explorarlas estadísticamente. Más adelante, se crean y transforman variables y se reemplazan valores faltantes para crear diferentes modelos y compararlos [2]. Esta metodología se basa en la parte técnica del proyecto como la aplicación de técnicas estadísticas y visuali-

zación de datos. Sin embargo, no considera los objetivos del negocio o el contexto del problema.

Knowledge Discovery in Databases Framework

KDD es una metodología propuesta por Fayyad en 1996, propone las siguientes cinco fases: selección, pre-procesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo [3].

Cross-industry standard process for data mining

Por último, CRISP-DM surge como una iniciativa financiada por la Comunidad Europea para desarrollar una plataforma de Minería de Datos. El objetivo de la iniciativa es fomentar la interoperabilidad de las herramientas a través de todo el proceso y eliminar la experiencia misteriosa y costosa de las tareas simples de minería de datos [4].

1.0.2 Metodología seleccionada

De las posibles metodologías se eligió la metodología CRISP-DM para el proyecto dado que no hay propietario, es independiente de la aplicación o la industria y es neutral con respecto a herramientas. Asimismo, a diferencia de KDD y SEMMA, la primera fase de CRISP-DM involucra el entendimiento del negocio que es fundamental para el correcto desarrollo de un proyecto.

Otra ventaja es que la documentación oficial describe en detalle cada fase y tareas con ejemplos concretos de aplicación [5].

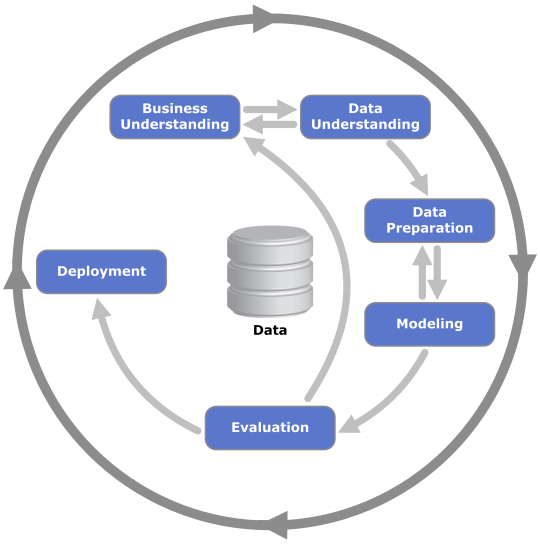


Figura 1.1: Fases del modelo CRISP-DM

1.1 Organización del documento

Como se visto en la, figura 1.1, la metodología propuesta en el año 2000 [6], propone las siguientes seis fases:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Implementación.

El documento esta organizado en siete capítulos correspondientes a las seis fases de la metodología mas este capítulo introductorio.

CAPÍTULO 2

COMPRENSIÓN DEL SECTOR EDUCATIVO

Este primer capítulo presenta el panorama general y explora las necesidades del "negocio". En este caso el negocio es el sector educativo.

A continuación, se introduce la problemática de la educación en México y los objetivos del proyecto con el fin de contribuir al desarrollo de un país más justo, más prospero y más libre.

2.1 Determinación de los objetivos del sector

La educación es de vital importancia para el desarrollo de un país y en México la calidad educativa es insuficiente. Como resultado, existen instituciones públicas y privadas cuya meta es mejorar el desempeño académico.

2.1.1 Contexto

La educación es relevante porque los beneficios de una sociedad más escolarizada se ven reflejados en una menor tasa de mortalidad [7], mayor democracia, participación ciudadana [8] y crecimiento económico [9] de la mano de una mayor equidad en la distribución de ingresos [10] [11].

La escolaridad se refiere al periodo de asistencia a un centro escolar [12]. Sin embargo, los beneficios no están relacionados con el número de años en la escuela, sino con el aprendizaje dentro de ella [9]. Una forma de medir el aprendizaje es evaluando las habilidades cognitivas.

La Organización para la Cooperación y el Desarrollo Económico (OCDE) desarrolló el Informe del Programa Internacional para la Evaluación de Estudiantes (PISA) con el fin de medir estas habilidades cognitivas. El objetivo es aplicar un examen estandarizado cada tres años en 72 países de la OCDE a alumnos de 15 años, evaluando una base sólida de conocimientos en lectura, matemáticas y ciencias [13].

En México la calidad educativa es insuficiente según los exámenes estandarizados internacionales. El país ha tenido resultados no satisfactorios desde el 2000 hasta el 2015, posicionándose entre los últimos 15 países. A lo largo de esos 15 años, los resultados han sido consistentemente bajos y sin cambios significativos [14]. No obstante, durante el periodo del 2000 al 2015 se han implementado varios programas educativos con el objetivo de mejorar el desempeño en las aulas.

Aunque existen logros del "Programa Nacional de la Educación 2001-2006", del "Plan Nacional de Desarrollo 2007-2012" y de la "Reforma Educativa del 2012" como mayores tasas de asistencia y de eficiencia terminal [14] , todavía existen retos para mejorar el desempeño escolar.

Identificación del problema

Actualmente, el "Proyecto de Nación 2018-2014" del presidente de México, López Obrador, propone la creación del Centro Nacional para la Revalorización del Magisterio y la Mejora Continua de la Educación (CNRMMCE) como sucesor del Instituto Nacional para la Evaluación de la Educación (INEE). El CNRMMCE deberá realizar estudios, investigaciones especializadas, emitir lineamientos relacionados con el desempeño escolar así como la mejora de las escuelas. El Estado deberá garantizar que los materiales didácticos, la infraestructura educativa, su mantenimiento y las condiciones del entorno contribuyan a los fines de la educación a través de programas sociales [15]. Es decir, el gobierno está interesado en implementar programas sociales

con el fin de mejorar el desempeño escolar.

Antecedentes

Históricamente han existido varios programas orientados a mejorar el desempeño de las escuelas como el Programa Escuelas de Tiempo Completo, Programa Desayunos Escolares, Programa de Acciones Compensatorias para Abatir el Rezago Educativo en la Educación Inicial y Básica, entre otros [1]

Una de las mayores dificultades de estos programas es seleccionar a las escuelas beneficiarias. Algunos programas, como los Programas Compensatorios Escolares, no han tenido resultados satisfactorios porque las escuelas atendidas no correspondían plenamente a los criterios de focalización y su distribución podía mejorar significativamente [16].

La definición de prioridad de los programas especiales, en qué y dónde se invierte primario, puede ser dictada por el gobierno federal o por los propios agentes del sistema escolar. Se realizan entrevistas con directivos y docentes e históricamente se ha dado prioridad a aspectos que tienen que ver con la infraestructura física de los establecimientos escolares y no con la calidad de docentes o servicios de la escuela [16].

Por un lado, los modelos econométricos ayudan a encontrar causalidad de factores

[1] Proyecto de Atención Educativa a la Población Indígena, Proyecto de Atención Educativa a la Población Infantil Agrícola Migrante, Proyecto de Enciclomedia, Programa de Escuelas de Bajo Rendimiento, Programa de Fortalecimiento del Servicio de la Educación Telesecundaria, Programa de Habilidades Digitales para Todos, Programa Asesor Técnico Pedagógico y para la Atención Educativa a la Diversidad Social Lingüística y Cultural, Programa Desayunos Escolares, Programa de Acciones Compensatorias para Abatir el Rezago Educativo en la Educación Inicial y Básica, Programa de Educación Inicial y Básica para la Población Rural e Indígena, Programa de Educación Primaria para Niñas y Niños Migrantes, Programa de Escuela Segura, Programa de Infraestructura, Programa Escuelas de Calidad, Programa Escuela Siempre Abierta, Programa Emergente para la Mejora del Logro Educativo, Programa Fortalecimiento de la Educación Especial y de la Integración educativa, Programa Nacional de Inglés en Educación Básica, Programa Nacional de Lectura, Programa Ver Bien para Aprender Mejor y Proyecto Mejoramiento del Logro Educativo en Escuelas Primarias Multigrado.

escolares y determinar que características de las escuelas mejoran el desempeño escolar. Por otro lado, estos modelos son costosos para determinar que escuelas deben participar en qué programa.

Dado que el mayor reto de las políticas públicas es la asignación eficiente de recursos de manera rápida, resulta conveniente usar modelos de aprendizaje de máquina para predicciones precisas en poco tiempo. Los modelos de aprendizaje de máquina han tenido buen desempeño detectando fraudes [17], prediciendo enfermedades [18] y deserción escolar [19]. Es por eso que ofrecen una alternativa a los métodos tradicionales.

2.1.2 Objetivos del sector

El sector educación tiene como objetivo garantizar una educación de calidad que promueva las oportunidades de aprendizaje a lo largo de la vida [20].

Una de las estrategias mencionadas previamente son los programas escolares. El objetivo del sector educativo es que los programas escolares sean exitosos y el éxito de los programas radica en la asignación de recursos. Por lo tanto, uno de los objetivos del sector educativo es determinar a que escuelas asignarle recursos.

Asimismo, uno de los criterios para asignar recursos puede ser el desempeño académico. Es decir, para disminuir la iniquidad de oportunidades, asignar recursos a aquellas escuelas con riesgo a empeorar su desempeño.

En consecuencia, se hace evidente la necesidad de predecir el desempeño académico de las escuelas en México.

Actualmente, la asignación de programas es tardada y costosa ya que se levantan entrevistas y mientras mayor se desee que sea el alcance más costosa es. Por estas razones, el objetivo del sector educativo es asignar programas de forma rápida y

precisa.

2.1.3 Criterios de éxito del sector

Un posible criterio de éxito es realizar asignaciones de programas sociales de forma más rápida, más precisa, más transparente y menos costosa.

En algunos programas sociales como Programa Escuelas de Tiempo Completo, toma más de tres meses seleccionar escuelas beneficiarias [21]. En específico, una mejora sería seleccionar escuelas en una semana.

La precisión se puede medir en las evaluaciones de programas sociales. Se espera que impacto del programa sea más positivo y efectivo con una asignación más precisa. Sin embargo, es difícil cuantificar cuánto más significativo ya que cada programa es diferente.

2.2 Valoración de la situación

En esta sección se explora a detalle los recursos, limitaciones y supuestos para determinar los objetivos de minería de datos.

2.2.1 Inventario de recursos

El objetivo, mencionado anteriormente, es predecir el desempeño académico en México. Sin embargo, el desempeño escolar es multidimensional y se puede medir y evaluar de distintas maneras cuantitativas y cualitativas.

Una manera de medir el desempeño escolar cuantitativamente es a través de pruebas estandarizadas.

México participa en la prueba estandarizada internacional PISA e internamente ha

implementado otras pruebas estandarizadas como Exámenes de la Calidad y el Logro Educativo (EXCALE), Evaluación Nacional de Logro Académico en Centros Escolares (ENLACE) y Plan Nacional para la Evaluación de los Aprendizajes (Planea).

Tabla 2.1: Pruebas estandarizadas en México

Prueba	Nivel de educación	Frecuencia de aplicación	Número de escuelas (último año)	Años evaluados	Datos disponibles por escuela
EXCALE	Básica y Media Superior	Cada tres años un mismo grado	3,552	2005- 2016	Sí
ENLACE	Básica y Media Superior	Cada año	122,608	2006-2014	Sí
Planea	Básica y Media Superior	Cada año	36,567	2014-2018	Sí
PISA	Media Superior	Cada 3 años	231	2003-2018	No

La tabla 2.1 muestra una comparación entre las cuatro pruebas mencionadas anteriormente. PISA es una prueba internacional con el defecto de que los datos no están disponibles por la escuela y evalúa a un menor número de escuelas que las pruebas nacionales. Por un lado, la prueba mas reciente y vigente es Planea. Por otro lado, la prueba con mayor alcance fue ENLACE en cuanto a número de años que se aplicó la prueba a un mismo grado y el número de escuelas evaluadas en un mismo año.

Asimismo, para poder predecir el desempeño resulta útil conocer características de la escuela, de los alumnos, directivos o personal docente. El Formato Estadístico 911 (F911) es un cuestionario llenado por todos los centros educativos al inicio y al final de cada ciclo escolar que incluye el número de alumnos por grado, desglosado por edad, el nivel de escolaridad del personal, estadísticas sobre los salones en uso y los alumnos discapacitados o con aptitudes sobresalientes [22].

Del mismo modo, el Instituto Nacional de Estadística y Geografía (INEGI) y la SEP recopilaron más información sobre el inmueble físico de los centros de trabajo con el

Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (INEGI)

En resumen, el inventario de recursos de datos costa de la siguiente información:

- Resultados de pruebas estandarizadas por escuela (EXCALE, ENLACE y Planea).
- Información de los alumnos y del personal del centro de trabajo (F911).
- Características de las escuelas (INEGI).

Estos datos se complementan con los siguientes recursos:

- Asesores y expertos en el tema de educación [2] y de minería de datos [3].
- Acceso a un cluster universitario para entrenar modelos.
- Convenio de 200 dólares de estudiante para máquinas virtuales con Microsoft Azure.
- Conocimiento de Python y acceso a la herramiento sklearn [23].

2.2.2 Requerimientos funcionales, supuestos y restricciones

Requerimientos funcionales

Los requerimientos funcionales del sistema son los siguientes:

- **Preprocesamiento de datos**: el sistema deberá ser capaz de aplicar técnicas de preprocesamiento y limpieza a los datos. En especifico, el sistema deberá manejar los valores faltantes, las diferentes codificaciones, los errores tipográficos.

[2] Dr. Enrique Seira
[3] MS Juan Salvador Mármol

- **Ingeniería de características**: el sistema deberá identificar las variables más importantes y crear variables significativas modificando las existentes.
- **Datos abiertos**: el sistema deberá ser construido utilizando en su mayoría datos abiertos.
- **Entrenamiento de modelos**: el sistema deberá construir y encontrar los mejores parámetros de diversos modelos de predicción.
- **Evaluación de modelos**: el sistema deberá evaluar el desempeño de los modelos con métricas estadísticas. Asimismo, el sistema deberá comparar los modelos.
- **Predicción**: el sistema deberá predecir, con los datos y los modelos entrenados, el desempeño de las escuelas en pruebas estandarizadas.
- **Comunicación de resultados**: los resultados y el análisis del sistema deberán ser publicados en una página web para que se tenga acceso a las bases de datos construidas, a la documentación y a los modelos.

Supuestos

Existen tres grandes supuestos. El primero y el mayor supuesto es que las pruebas estandarizadas como el ENLACE miden el desempeño académico de una escuela. El segundo es suponer que las características de la escuela y de los alumnos tienen relación con el desempeño académico. El tercero es que los programas sociales tienen algún efecto significativo sobre el desempeño escolar. Este supuesto esta basado los impacto positivo significativo del programa Escuelas de Tiempo Completo [24]

Restricciones

El proyecto tiene las siguientes tres restricciones: disponibilidad, tamaño y calidad de los datos.

En primer lugar, el sistema estará restringido por los datos abiertos disponibles. En caso de que no sean libres los datos para una escuela o para un año, el sistema no los incluirá.

En segundo lugar, el tamaño de algunas bases supera la memoria de una computadora portátil promedio. Por lo tanto, se usarán servidores alternativos para manejar bases de datos muy grandes.

En tercer lugar, es posible que la calidad de los datos sea baja ya que fueron capturados a través del tiempo por diferentes personas y organismos.

2.2.3 Riesgos y contingencias

La tabla 2.2 muestra algunos riesgos y posibles contingencias.

Tabla 2.2: Posibles riesgos y contingencias

Riesgo	Probabilidad (1-4)	Impacto (1-4)	Contingencia
No se obtienen los datos del formato 911	3	3	Utilizar únicamente la información del CEMABE
No identificar errores de captura en las bases de datos	2	4	Documentar y publicar la limpieza de las bases para recibir retroalimentación

2.2.4 Terminología

A continuación, se incluyen dos glosarios. Uno del sector educativo y otro con terminología de la minería de datos.

El siguiente glosario incluye términos relevantes en el sector educativo:

- **Centro de trabajo**: Unidad productiva. Un centro de trabajo educativo es coloquialmente una escuela. Todos los centros de trabajo tienen una *Clave de Centro de Trabajo (CCT)* que identifica únicamente a cada escuela. En este caso, múltiples centros de trabajo pueden estar en un mismo inmueble. Es decir, un mismo edificio físico puede tener varios CCT dependiendo el turno (matutino, vespertino o completo) o nivel educativo (pre-escolar, primaria o secundaria).
- **Personal docente**: Se refiere al personal del centro de trabajo con funciones de docencia. Coloquialmente son los "profesores".

El siguiente glosario incluye términos relevantes sobre la minería de datos:

- **Limpiar datos**: Es parte del procesamiento encargado de que los datos sigan un mismo formato y estén en delimitado rango.
- **Valores faltantes**: Aquellos valores cuyo valor es desconocido.

2.2.5 Análisis de costos y beneficios

Los beneficiarios del sistema son, en primera instancia, las instituciones que buscan identificar escuelas en riesgo de tener bajo desempeño y con potencial de crecimiento como el CNRMMCE. Como consecuencia, los beneficiarios finales son las escuelas que recibirán apoyo y la sociedad que a largo plazo tendrá mayores niveles educativos y calidad de vida.

Por un lado, el principal costo del proyecto es el tiempo invertido recuperando, limpiado y manipulando datos.

Por otro lado, el proyecto trae el beneficio de hacer accesible el análisis y conjunto de datos. Al igual que contribuir con una propuesta en México para optimizar la asignación de recursos. Esta propuesta presenta grandes ahorros a los métodos tradicionales de visitar los centros de trabajos y realizar entrevistas y trae el beneficio de tener mayor alcance ya que un mayor número de escuelas pueden ser consideradas.

2.3 Determinación de los objetivos de minería de datos

2.3.1 Objetivos del proyecto de minería de datos

Tomando en cuenta los objetivos del proyecto y los requerimientos funcionales mencionados anteriormente, el problema identificado es predecir el desempeño académico.

Suponiendo que las pruebas estandarizadas miden el desempeño académico de una escuela, el problema se traduce en predecir los resultados de la prueba estandarizada.

Uno de los beneficios propuestos es tener un gran alcance, por lo tanto resulta conveniente utilizar los resultados de ENLACE ya que, como visto en la tabla2.1, es la prueba que se aplicó en el mayor número de escuelas.

Por lo tanto, el objetivo final de minería de datos es predecir los resultados de ENLACE.

El problema de predicción se puede abordar como un análisis de clasificación o de regresión. La SEP, encargada de aplicar ENLACE, clasificó los resultados en los siguientes cuatro niveles: insuficiente, elemental, bueno y excelente. Sin embargo, las calificaciones de ENLACE son numéricas continuas y no categóricas.

En figura 2.1 se muestra la distribución de calificaciones por escuela. Dado que la

distribución es continua y tiene una sola moda, no parece adecuado clasificar las calificaciones en categorías o niveles como lo hace la SEP ya que habría muchas observaciones en los bordes de las clases.

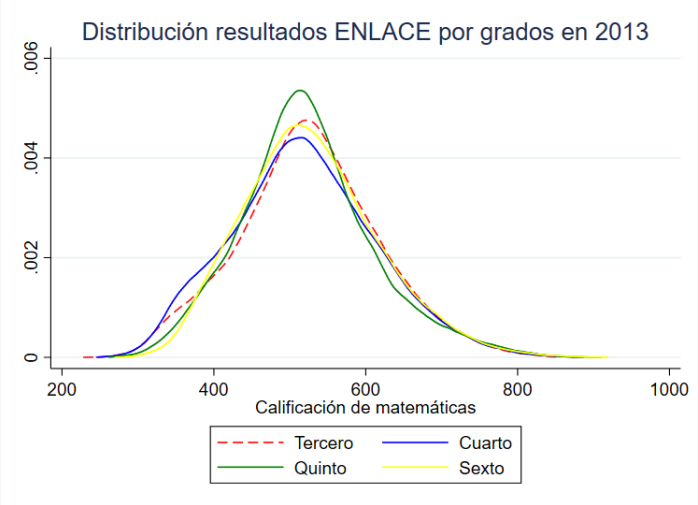


Figura 2.1: Distribución de resultados de Matemáticas

Como resultado, se identifica el análisis como un problema de regresión sobre la calificación numérica y continua de la prueba ENLACE.

2.3.2 Criterios de rendimiento del proyecto de minería de datos

La métrica que se usará para evaluar los modelos será el criterio de información de Akaike (AIC). En la ecuación 2.1 k es el número de parámetros en el modelo estadístico , y L es el máximo valor de la función de verosimilitud para el modelo estimado [25].

$$AIC = 2k - 2\ln L \tag{2.1}$$

AIC maneja un balance entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos [25].

El sistema será exitoso si logra construir modelos que superen los puntos de referencia iniciales construidos con ningún modelo y con una regresión simple.

2.4 Soluciones relacionadas

2.4.1 Modelos econométricos

El Banco Mundial realizó un estudio en el 2012, utilizando los datos del ENLACE y de una encuesta de contexto a participantes de la prueba, en el cual se construyó un modelo econométrico para encontrar las determinantes del logro escolar en México. Los resultados indican que el 40 % de las diferencias en las calificaciones de matemáticas se pueden explicar por la infraestructura de la escuela, la calidad de los docentes y la relación entre los estudiantes y autoridades escolares, medidas como opiniones de los alumnos en la encuesta de contexto. Las principales desventajas de este modelo son que utiliza una pequeña muestra de la población (120,000 alumnos de 14,098,879 alumnos que presentaron la prueba) y que no toma en cuenta interacciones entre variables.

2.4.2 Modelos de aprendizaje de máquina

Métodos con árboles

El artículo "Student and school performance across countries: A machine learning approach" presenta un análisis de determinantes de resultados de la prueba PISA. La prueba PISA, como mencionado anteriormente, es una prueba estandarizada a nivel mundial (similar a la prueba ENLACE en México). El artículo encuentra características de los estudiantes asociadas con resultados en la prueba y características de la escuela que contribuyen al valor agregado de la escuela. Asimismo, se explorar relaciones no-lineales e interacciones entre variables. Esto se logra utilizando métodos

basados en árboles que son más flexibles que los modelos tradicionales estadísticos ya que no se basan en suposiciones paramétricas. En primera instancia, se utiliza una regresión multinivel de árboles para estimar el valor agregado de la escuela. Más adelante, con árboles de regresión y boosting se relaciona el valor agregado de la escuela con las características de la escuela [26].

Métodos con redes neuronales

El artículo "GritNet: Student Performance Prediction with Deep Learning" plantea el problema de predicción de desempeño de un alumno como un análisis de eventos secuenciales y propone una red (GridNet) construida sobre una memoria bidireccional de largo y corto plazo (Bidirectional Long Short-Term Memory) [27]. Este método se basa en el principio que las redes recurrentes pueden usar sus conexiones de "feedback" para guardar representaciones de eventos recientes en forma de activaciones.

2.5 Producción de un plan de proyecto

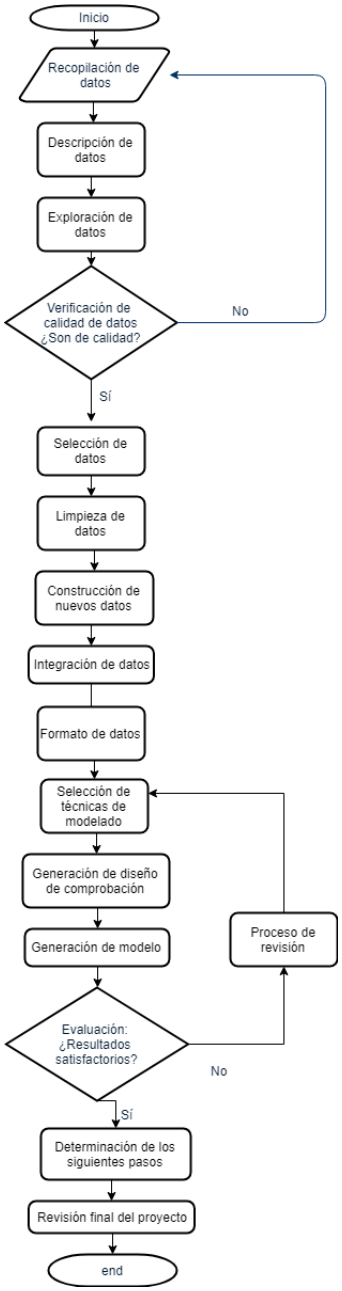


Figura 2.2: Flujo de tareas

La figura 2.2 muesta el flujo de las principales tareas del proyecto.

2.5.1 Valoración de herramientas y técnicas

Python es un lenguaje de programación interpretado con las ventajas de suportar múltiples bibliotecas de minería de datos [28]. Entre las bibliotecas disponibles cabe destacar Pandas para manejo de tablas, NumPy para el manejo de arreglos y Scikit-learn para herramientas de de minería de datos y aprendizaje de máquina.

Otros lenguajes de programación usados comúnmente en problemas de minería de datos son R y Stata. Ambos ofrecen buenas herramientas para visualizar y analizar datos. Por ejemplo, Stata permite abrir y manipular archivos muy grandes y en una gran variedad de formatos, incluyendo dbs. Asimismo, se tiene acceso a un servidor remoto con Stata que permite manipular conjuntos de datos que una computadora personal de 24 GB no puede cargar en memoria.

Por estos motivos, se utilizará Stata para la exploración de datos y el resto del análisis se implementará en Python.

CAPÍTULO 3

COMPRENSIÓN DE LOS DATOS

El objetivo de este capítulo es reportar los datos iniciales recopilados, describir los datos para más adelante explorar y verificar la calidad de los datos.

3.1 Recopilación de datos iniciales

En el capítulo 1 en la sección llamada "Valoración de la situación" se describe el inventario de recursos. En resumen, este inventario costa de los siguientes datos:

- Resultados de pruebas estandarizadas por escuela (EXCALE, ENLACE y Planea).
- Información de los alumnos y del personal del centro de trabajo (F911).
- Características de las escuelas (INEGI).

La tabla **??** muestra los conjuntos de datos iniciales, el formato y el método que se utilizó para obtenerlos.

Nombre	Formato	Método
Censo escuelas CEMABE 2013	CSV	Descarga electrónica cemabe.inegi.org.mx/
F911 2006-2013	DBF	Solicitud email plataformadetransparencia.org.mx
Resultados ENLACE 2006 -2013	Varios	Descarga electrónica enlace.sep.gob.mx/

Tabla 3.1: Conjunto de datos obtenidos

3.1.1 Recopilación resultados de pruebas estandarizadas

Los resultados de EXCALE, PISA y Planea están disponibles en el portal del INEE en la sección de evaluaciones y bases de datos [1].

En el capítulo anterior se propuso utilizar los resultados de ENLACE porque tuvo mayor alcance.

Los resultados ENLACE se pueden obtener a nivel escuela y a nivel persona.

A nivel escuela, los resultados históricos de de ENLACE están disponibles en el portal de ENLACE [2] [3]. Cabe destacar que los resultados del 2006 y del 2007 están integrados en una misma tabla y que los resultados a nivel escuela nacional no están disponibles en el 2008.

Asimismo, los datos a nivel persona se obtuvieron del Centro de Investigación Económica [4].

3.1.2 Recopilación resultados del formato estadístico 911

Los conjuntos de datos del formato 911 se solicitaron por internet mediante la Plataforma Nacional de Transparencia (PNT). Dicho organismo envió las bases por correo a un domicilio en un CD con un costo de diez pesos más gastos de envío. [29].

Se obtuvieron las respuestas del formato de inicio de cursos y fin de cursos para primaria desde el 2006 hasta el 2013.

[1] Información disponible para descargar en la siguiente liga: https://www.inee.edu.mx/evaluaciones/bases-de-datos/

[2] Resultados desde 2006 hasta 2012 disponibles en la siguiente liga: http://www.enlace.sep.gob.mx/ba/resultados$_a nteriores/$

[3] Resultados del 2013 disponibles en la siguiente liga: http://www.enlace.sep.gob.mx/content/ba/pages/base$_d e_d ato$

[4] Agradecimiento al Dr. Enrique Seira

3.1.3 Recopilación datos del CEMABE

La información de las escuelas se obtuvo del INEGI descargados desde el portal de Datos Abiertos del Gobierno de México [30].

3.2 Descripción de los datos

Todos los datos recopilados están a nivel escuela. Es decir, una escuela es una observación. Estas observaciones están identificadas con la Clave de Centro de Trabajo (CCT).

3.2.1 Descripción ENLACE

Tabla 3.2: Descripción general datos ENLACE por escuela

Año	Nombre	Extensión	Tamaño (MB)	Escuelas	Variables
2006, 2007	e2006_2007	dbf	139	397,424	35
2009	e2009 (hoja 1)	xls	71	49,988	84
2009	e2009 (hoja 2)	xls	54	38,304	84
2010	e2010 (hoja 1)	xls	75	55,651	81
2010	e2010 (hoja 2)	xls	45	33,884	81
2011	e2011 (hoja 1)	xls	66	48,521	81
2011	e2011 (hoja 2)	xls	56	42,027	81
2012	e2012 (hoja 1)	xls	41	45,742	81
2012	e2012 (hoja 2)	xls	34	38,114	81
2013	e2013 (hoja 1)	xls	66	48,521	81
2013	e2013 (hoja 2)	xls	56	42,027	81

La tabla 3.2 muestra los nombres, extensiones, tamaños y dimensiones de los datos de resultados de ENLACE a nivel escuela escuelas.

La tabla 3.3 muestra los conjuntos de datos a nivel alumno, cada alumno está iden-

tificado con un folio único en cada año. Las bases de datos a nivel alumno sí cuentan con los resultados del 2008.

Tabla 3.3: Descripción general datos ENLACE por alumno

Nombre	Tamaño (MB)	Alumnos	Escuelas	Variables
ENLACE2006	969	9,529,490	111,316	15
enl07_A	548	3,966,280	45,876	20
enl07_B	858	6,182,386	74,020	20
RESULT_ALUMNOS_08_A	408	4,306,540	51,539	21
enl08_B	843	5,646,800	68,433	23
RESULT_ALUMNOS_09_A	847	8,029,920	88,285	30
RESULT_ALUMNOS_09_B	947	5,157,768	29,496	32
RESULT_ALUMNOS_10 _A	266	6,054,266	52,526	8
RESULT_ALUMNOS_10 _B	279	6,054,266	67,379	8
RES_ENLACE_10_2	2,495	13,772,359	119,905	30
resul_enlace_11	1,152	8,759,180	90,538	33
resul_alum_eb12	1,411	13,507,167	114,346	32
enl2013_alum	3,304	14,098,879	120,648	21

Nota: Todas las bases están en formato de texto comma-separated values (csv)

3.2.2 Descripción F911

La SEP, a través de la Dirección General de Planeación y Programación (DGPP), realiza el levantamiento de la información estadística de todos los centros educativos, al inicio y fin de cada ciclo escolar, en todas las entidades federativas del país, utilizando el formato estadístico 911 [22]. Cabe resaltar que el formato es diferente para primarias generales, indígenas y comunitarias. Es decir, el número y el orden de las preguntas es diferente en cada caso.

Los datos del F911 se recopilaron en formato dbf. La tabla 3.4 muestra el número

de observaciones del formato 911 para cada tipo de escuela: general, comunitaria e indígena. Cada observación representa una escuela.

Tabla 3.4: Número de observaciones del formato 911 del inicio de cursos

	Número de observaciones		
Año	**General**	**Comunitaria**	**Indigena**
2006	76,991	12,296	9,830
2007	77,366	11,966	9,865
2008	77,702	11,637	9,953
2009	78,096	11,511	9,975
2010	78,430	11,756	10,036
2011	78,545	11,860	10,080
2012	78,836	11,866	10,173
2013	78,809	11,807	10,200

3.2.3 Descripción CEMABE

El Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (INEGI) se llevó a cabo durante septiembre, octubre y noviembre del 2013 con el objetivo de captar las características específicas de las escuelas, maestros y alumnos de instituciones públicas y privadas de educación básica del sistema educativo escolarizado y especial. El censo incluye 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 [31].

Las datos del INEGI se recopilaron en formato csv. El conjunto de datos costa de tres tablas. La figura 3.1 muestra el diagrama de entidad relación de las tablas del CEMABE. Las tablas TR_INMUEBLES y TR_CENTROS se pueden unir por la clave de identificación del inmuebles ID_INM, la relación es de uno a muchos. Esto quiere decir que un inmueble puede tener varios centros de trabajo; por ejemplo, en un

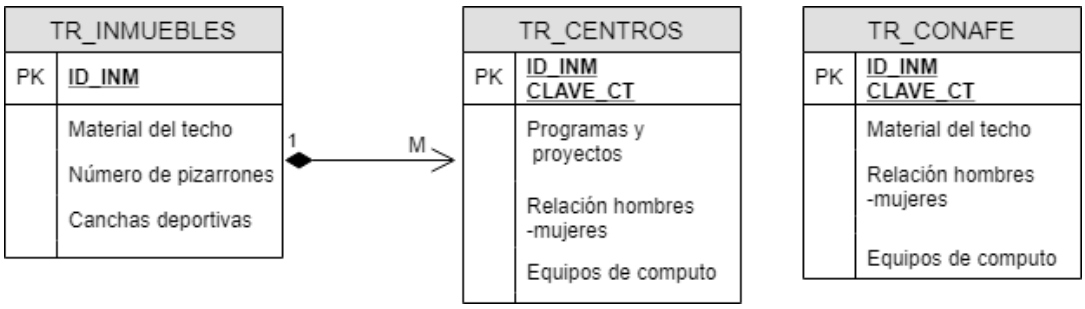


Figura 3.1: Diagrama entidad relación de tablas del CEMABE

mismo edificio puede trabajar una escuela con turno matutino y otra escuela con turno vespertino o una escuela primaria y secundaria. Sin embargo, la tabla TR_CONAFE no sé relaciona con ninguna de las otras tablas. La tabla TR_CONAFE contiene información similar a TR_INMUEBLES y TR_CENTROS para escuelas comunitarias de la Consejo Nacional de Fomento Educativo (CONAFE).

La tabla 3.5 muestra las dimensiones de los datos recopilados del INEGI.

Tabla 3.5: Descripción general datos CEMABE

Nombre	**Extensión**	**Tamaño**	**Número de observaciones**	**Número de columnas**
TR_CENTROS	csv	300M	177,829	266
TR_INMUEBLES	csv	193M	149,707	161
TR_CONAFE	csv	29M	33,849	155

Cabe resaltar que las variables de las tablas tienen un formato número en la mayoría de los casos en el cuál los valores faltantes están representados por el número “99”.

3.3 Exploración de datos

A continuación se muestran resultados significativos de la exploración de datos.

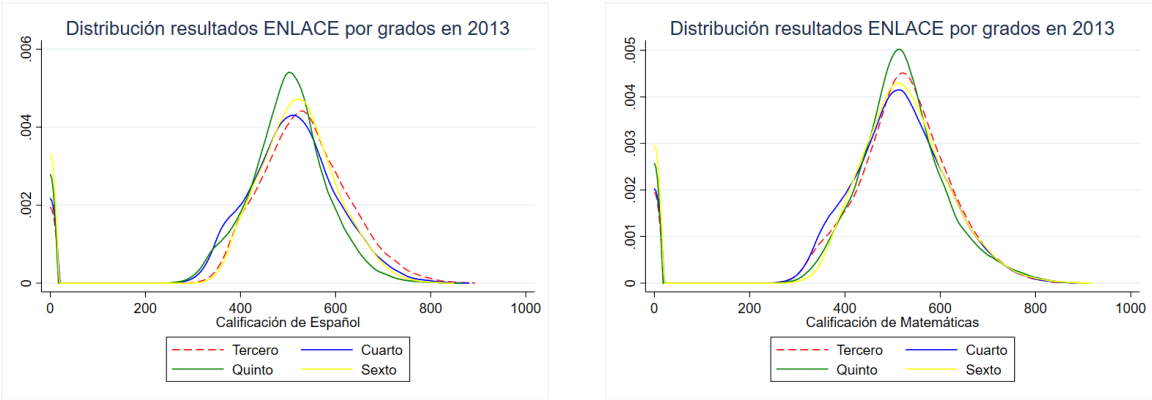
3.3.1 Exploración univariada

La variable a predecir es la calificación ENLACE. Los datos se obtuvieron a nivel escuela y a nivel alumno.

ENLACE nivel escuela

[PROBABLEMENTE QUITEMOS ESTO]

La figura 3.2a y 3.2b muestran la distribución de calificaciones de ENLACE por grado del 2013. Las figuras muestran un "pico" en el cero.



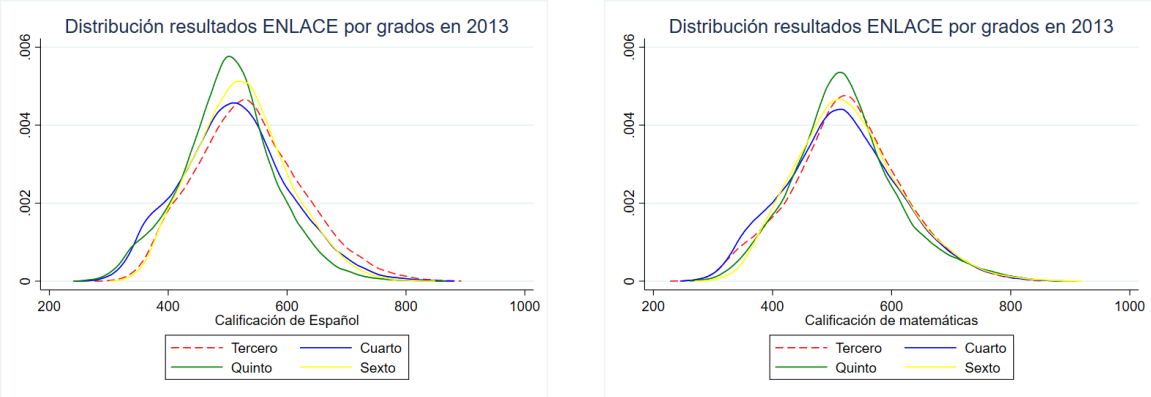
(a) Español (b) Matemáticas

Figura 3.2: Calificaciones de primaria por grado escolar

En algunos casos, las calificación es cero porque en ese año un grado no presentó la prueba.

El resto de la exploración se realizó utilizando las calificaciones corregidas. La corrección en las figuras 3.3a y 3.3b fue reemplazar las calificaciones cero con valores faltantes.

La figura 3.4 muestra las diferentes distribuciones por materia en sexto de primaria en el 2013. La diferencias en las distribuciones son resultado de las diferentes escalas



(a) Español (b) Matemáticas

Figura 3.3: Calificaciones de primaria por grado escolar corregidas.

en cada materia. Es decir, cada materia tuvo un número de reactivos diferentes en cada año y en cada grado.

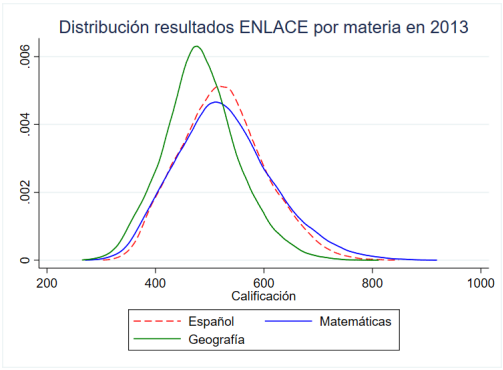
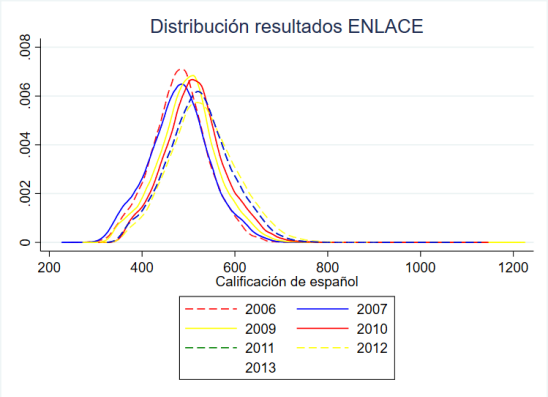
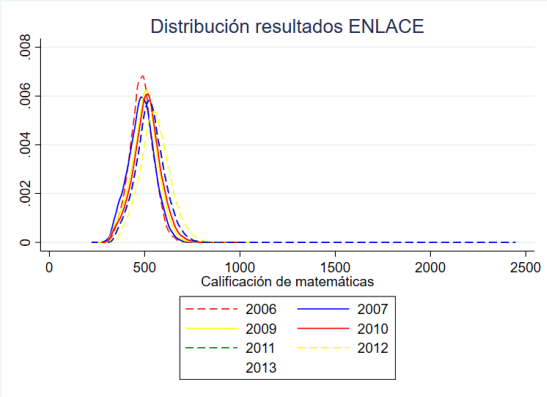


Figura 3.4: Distribución resultados por materia en 2013

Las figuras 3.6a y 3.6b muestran las distribuciones promedio de todos los grados en siete años que se aplicó la prueba, faltan los datos del 2008. Una vez más, las diferencias en las distribuciones se explican como resultado de escalas diferentes. Asimismo, es posible que existan calificaciones mal capturadas que alargan las colas de las distribuciones.



(a) Español

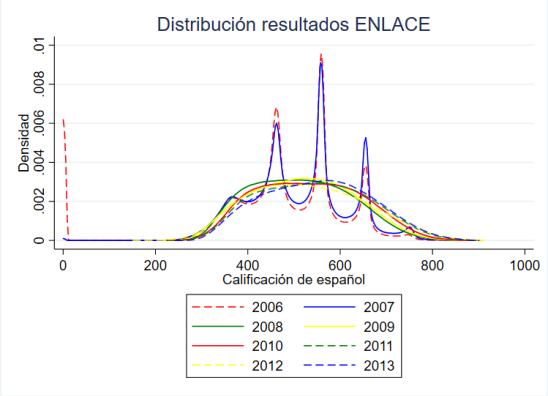


(b) Matemáticas

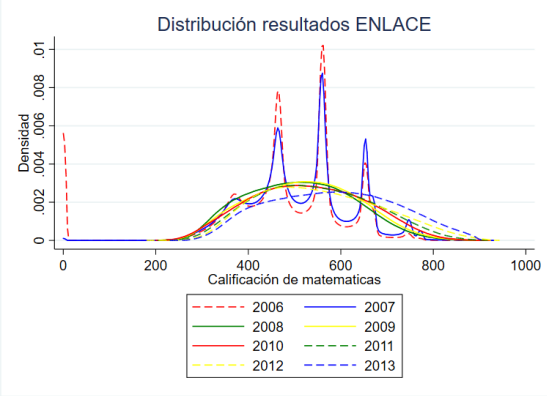
Figura 3.5: Comparación de distribución de resultados desde 2006 hasta 2013 (sin 2008)

ENLACE nivel alumno

Las figuras 3.6a y 3.6b muestran la distribución de calificaciones de español y matemáticas. Cabe resaltar que se tiene información de todos los años pero que las distribuciones son distintas.



(a) Español

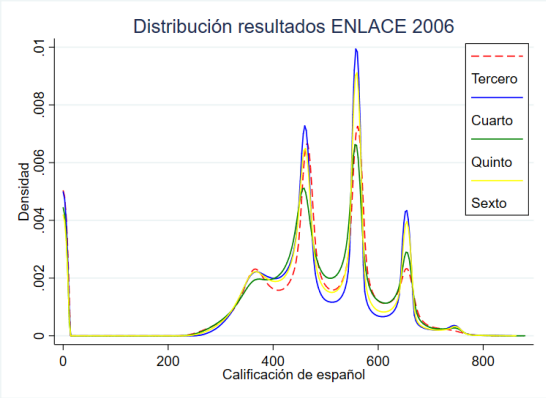


(b) Matemáticas

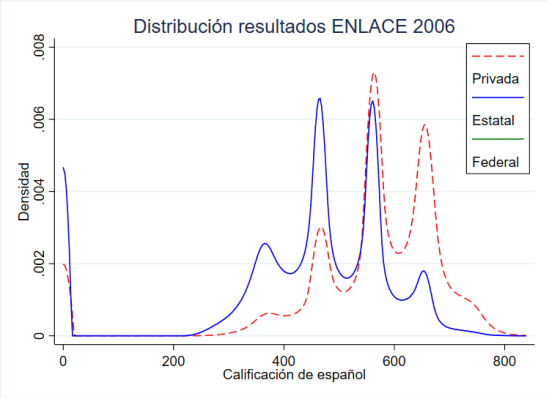
Figura 3.6: Comparación de distribución de resultados desde 2006 hasta 2013

El 2006 y 2007 tienen una distribución multi-modal diferente a la distribución normal de los otros años. La figura 3.7a muestra en detalle la distribución por grado del 2006 y la figura 3.7b muestra las distribuciones por sostenimiento. En ambos casos la

distribución es multimodal.



(a) Resultados por grado

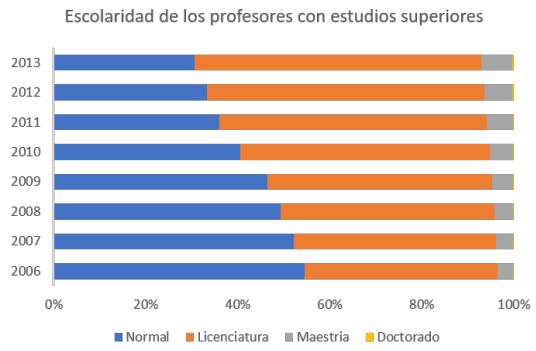


(b) Resultados por sostenimiento

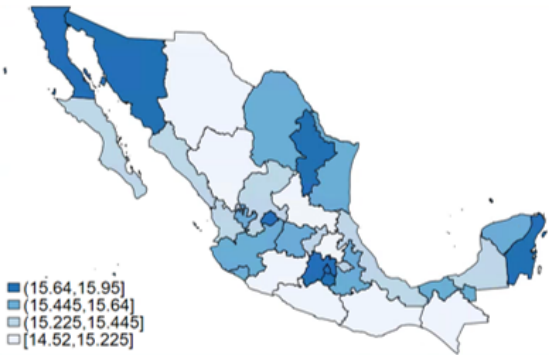
Figura 3.7: Comparación de resultados por grado y sostenimiento

Variables independientes

Se realizaron histogramas de todas las variables de todas las bases para explorar el comportamiento de los datos. Las figuras 3.8a y 3.10 muestran resultados notables de la exploración del formato estadístico 911 y las figuras 3.11a y 3.11b los resultados del INEGI.



(a) Personal docente titulado



(b) Años de escolaridad promedio por estado

Figura 3.8: Visualizaciones interesantes del F911

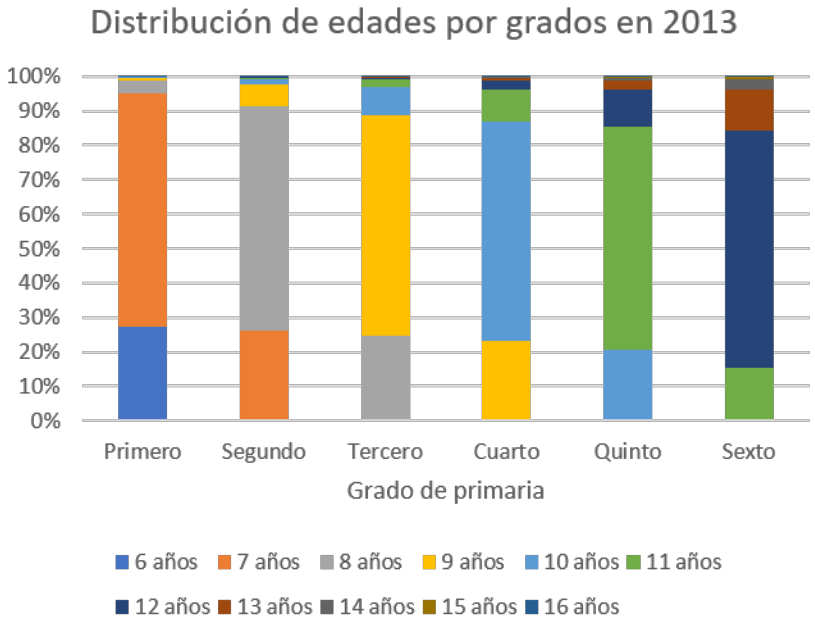


Figura 3.9: Edades de los alumnos por grado

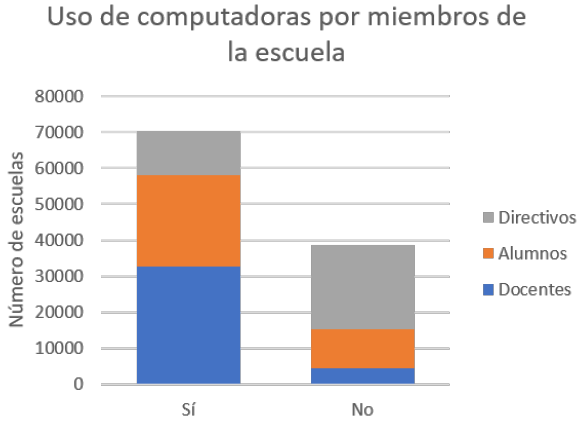
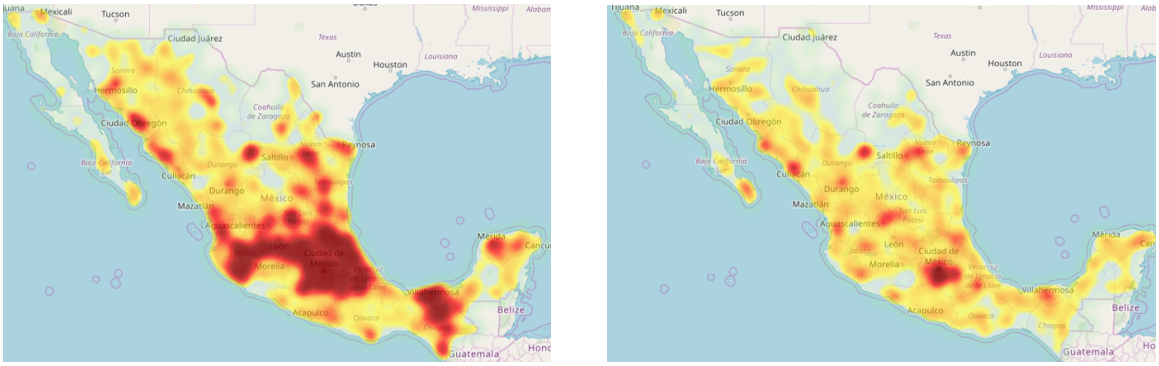


Figura 3.10: Uso de computadoras por miembros de la escuela

3.3.2 Exploración bivariada

Como muestra la figura 3.12, existe una gran correlación positiva entre los resultados de español y de matemáticas. La tabla 3.6 muestra las correlaciones por escuela entre todas las materias que fueron evaluadas en algún momento en la prueba ENLACE.

La variable objetivo de este proyecto es la calificación ENLACE. Existen calificaciones



(a) Programa de desayunos escolares

(b) Programa de tiempo completo

Figura 3.11: Mapa de calor de escuelas participantes en programas nacionales. Rojo es una mayor concentración y verde menor.

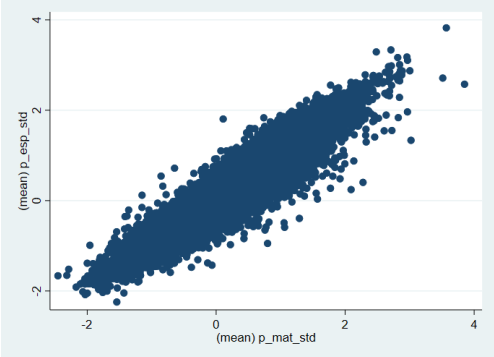


Figura 3.12: Correlación entre resultados español y matemáticas

ENLACE de Español, Matemáticas, Ciencias, Geografía e Historia. Sin embargo, únicamente las materias de Español y Matemáticas se presentaron todos los años, por lo cual se tiene más información de dichas materias. Dado que están altamente correlacionadas, se pueden utilizar indistintamente para el análisis.

Un problema que presentan las calificaciones es que cada grado, cada año y cada materia utilizó un número de incisos diferentes. Por lo tanto, la escala es distinta. Una posible solución es estandarizar las calificaciones por grado, materia y año.

A pesar de que las materias de Español y Matemáticas están altamente correlacionadas, como se ve en la figura 3.12, es posible que los resultados de español y de matemáticas surgan de procesos cognitivos distintos. Es decir, el buen dominio de la

Tabla 3.6: Tabla de correlaciones de una escuela entre materias

	Matemáticas	**Ciencias**	**Civismo**	**Geografía**	**Historia**
Español	0.76	0.74	0.76	0.75	0.64
Matemáticas	-	0.70	0.68	0.73	0.62
Ciencias		-	0.63	0.64	0.60
Civismo			-	0.55	0.59
Geografía				-	0.62

lengua española se aprende en casa y el buen dominio de las matemáticas se aprende en la escuela [32]. Dado que se desea conocer el desempeño de la escuela, tiene sentido utilizar las calificaciones de matemáticas.

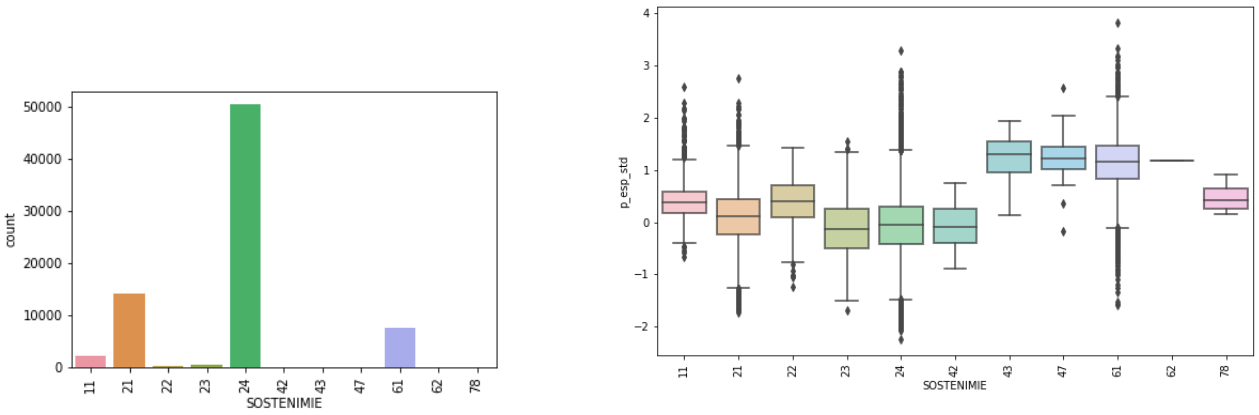
Como resultado, la variable objetivo es el promedio por escuela de matemáticas de los resultados estandarizados por año y grado.

Tabla 3.7: Variables del F911 general con alta correlación con ENLACE

Nombre	**Correlación**	**Descripción**
V917	0.51	Número de mensualidades que se pagan en escuelas particulares.
SOSTENIMIE	0.50	Fuente que proporciona los recursos financieros para el funcionamiento del centro de trabajo
V836	0.48	Total de personal
V825	0.42	Total de personal docente mujeres
V833	0.41	Total de profesores de idiomas mujeres

La tabla 3.7 muestra las 5 variables del formato 911 de escuelas generales con mayor correlación con la variable objetivo.

Sostenimiento se refiere a la fuente que proporciona los recursos financieros para el funcionamiento del centro de trabajo. En el sistema de centros de trabajo se utilizan los sostenimientos federal, estatal, municipal, autónomo y particular y subsidiado [33]. La figura 3.13a muestra el sostenimiento según el formato 911. Utilizando el clasifica-



(a) (b)

Figura 3.13: Sostenimiento. Federal (24), Estatal (21), Privado (61), Federal transferido (11)

dor del CCT (tercer carácter) identificó el nombre del sostenimiento correspondiente a la codificación [34]. La figura 3.13b muestra una diagrama de cajas y bigotes de las calificaciones de matemáticas por nivel de sostenimiento. Es interesante observar que los niveles de sostenimiento más comunes tienen datos atípicos muy separados del resto. La figura 3.14 muestra la distribución de calificaciones de español según el financiamiento. Cabe resaltar que todos los sostenimientos comparten un segmento del soporte.

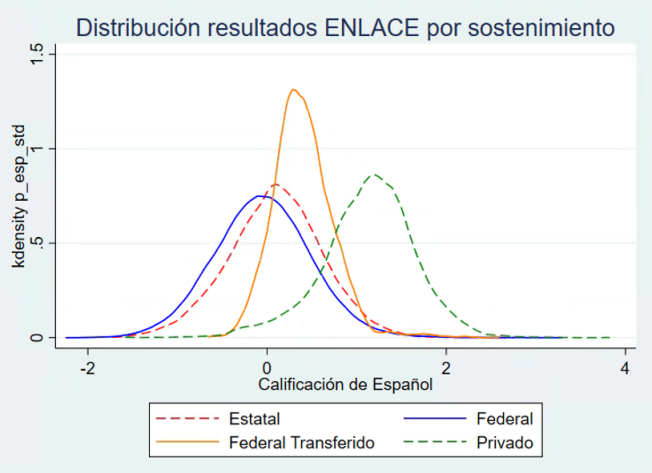


Figura 3.14: Distribución de resultados de matemáticas por sostenimiento

Tabla 3.8: Variables del F911 indígena con alta correlación con ENLACE

Nombre	**Correlación**	**Descripción**
V698	0.30	Cantidad de profesores que se encuentran en el programa de carrera magisterial
V435, V436, V437	0.24	Cantidad de personal docente que habla, lee y escribe la lengua materna de la comunidad
V692	0.23	Total de personal docente mujeres
V717	0.23	Total de aulas
V281	0.22	Número de alumnos de 11 años en sexto de primaria

Las tablas 3.8 y 3.9 muestran las variables del formato 911 de escuelas indígenas y comunitarias respectivamente con mayor correlación con la variable objetivo.

Tabla 3.9: Variables del F911 comunitario con alta correlación con ENLACE

Nombre	**Correlación**	**Descripción**
RENGLON	-0.29	Sostenimiento
V363	-0.17	Número total de alumnos
V339	-0.16	Número total de alumnos hombres
V351	-0.16	Número total de mujeres
V75	-0.14	Subtotal de alumnos en el segundo ciclo del primer nivel

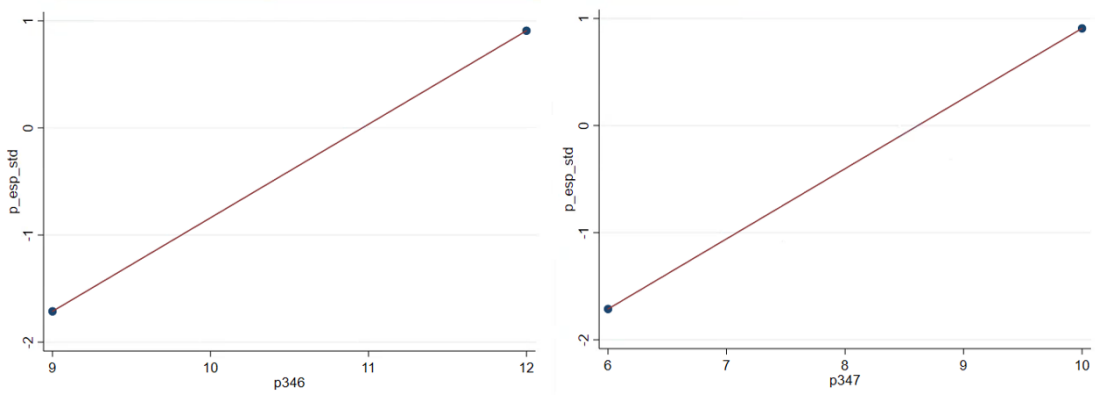
La tabla 3.10 muestra las variables de la tabla de Çentros"del CEMABE con mayor correlación con la calificación de matemáticas de ENLACE.

Cabe destacar que las variables p346 y p347 tienen una correlación perfecta, es decir, igual a uno. Las figuras 3.15a y 3.15b muestran las gráfica de dispersión de la calificación de matemáticas y el número de alumnos censados en séptimo grado (p346 y p347). La correlación perfecta, como se ve en las figuras 3.15a y 3.15b, se debe a que existen sólo dos observaciones no nulas de p346 y p347. Dichas variables indican el

Tabla 3.10: Variables de la tabla centros del CEMABE con alta correlación con ENLACE

Nombre	Correlación	Descripción
p347	1.00	Alumnos censados en séptimo grado (hombres)
p346	1.00	Alumnos censados en séptimo grado
p194	0.67	Uso de arenero por personal o alumnos
p210	0.61	Uso del aula de medios por alumnos
p211	0.60	Uso del aula de medios por docentes

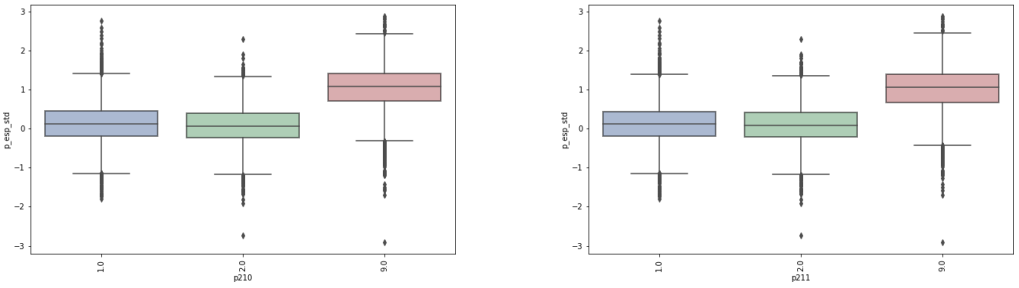
número de alumnos censados en séptimo grado, estas variables tienen sentido porque el INEGI se realizó para escuelas primarias y secundarias. Séptimo grado es el primer grado de secundaria, por lo tanto, las escuelas primarias no tienen séptimo grado. Dado que las escuelas de los resultados de ENLACE todas son escuelas primarias, y por lo tanto, solo tienen seis grados, las dos observaciones no nulas son datos atípicos. En una misma instalación es posible que haya escuelas primaria y secundaria, sin embargo, tienen CCT distintos. Las dos observaciones no nulas corresponden a Centro de Atención Múltiple (CAM) cuyo CCT es el mismo para el nivel primaria y secundaria.



(a) Alumnos censados en séptimo grado (b) Hombres censados en séptimo grado

Figura 3.15: Gráficas de dispersión de la calificación de español y el número de alumnos censados en séptimo grado (p346 y p347)

Las variables p210 y p211 de la tabla 3.10 tampoco son muy informativas. Las figuras 3.16a y 3.16b muestran diagramas de cajas y bigotes de las variables y la calificación ENLACE de matemáticas. Para las variables p210 y 211, la respuesta "1" quiere decir que el aula de medios es utilizada por los alumnos y docentes respectivamente, el número "2" quiere decir que el aula no es utilizada por los alumnos y "9" son los valores no especificados. En ambos casos, la diferencia entre sí utilizarla y no utilizarla es menor a no haber especificado si se usaba o no.



(a) Uso del aula de medios por alumnos (b) Uso del aula de medios por docentes

Figura 3.16: Gráficas de cajas y bigotes de calificación de español y uso de aula de medios

Tabla 3.11: Variables de la tabla CONAFE del CEMABE con alta correlación con ENLACE

Nombre	**Correlación**	**Descripción**
p327	-0.12	Total de alumnos censados en centros de trabajo censados (mujeres)
p337	-0.13	Alumnos censados en cuarto grado
p340	-0.13	Alumnos censados en quinto grado
p326	-0.13	Total de alumnos censados en centros de trabajo censados (hombres)
p325	-0.13	Total de alumnos censados en centros de trabajo censados

La tabla 3.11 muestra que las variables de la tabla de CONAFE con mayor correlación con el resultado ENLACE de matemáticas están inversamente correlacionadas. Esto

es coherente con la tabla 3.9 ya que ambas tablas son de información de escuelas comunitarias y muestran que existe una correlación inversa entre el número de alumnos y los resultados en la prueba ENLACE.

Tabla 3.12: Variables de la tabla inmueble del CEMABE con alta correlación con ENLACE

Nombre	**Correlación**	**Descripción**
p34	0.41	Total de tazas sanitarias
p5	0.35	Número de Centros de Trabajo en el inmueble
p36	0.31	Total de mingitorios
p20	0.26	Letrina u hoyo negro
p102	-0.25	Existencia de aula de usos múltiples

Finalmente, la tabla 3.12 las variables de la tabla del inmueble con mayor correlación con el resultado ENLACE de matemáticas. La figura 3.17 muestra una gráfica de dispersión por bloques de la variable con mayor correlación, número de tazas sanitarias (variable p34), y el resultado ENLACE de matemáticas.

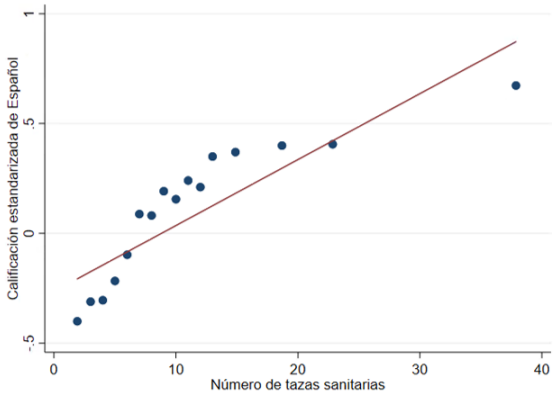


Figura 3.17: Gráfica de dispersión por bloques del número de tazas sanitarias y calificación ENLACE

3.4 Verificación de calidad de datos

3.4.1 Calidad de ENLACE

Por un lado, la prueba ENLACE ha sido criticada en varías ocasiones por inflación de resultados y falta de control en la aplicación [35].

En los últimos años, se realizó un chequeo de calidad de las respuestas de los alumnos y se agregó a los resultados una columna indicadora por alumno de si "copio" o no. En la base de las escuelas, se suma el número de observaciones no confiables en una columna de "resultados poco confiables". El indicador de "copia" fue asignado por los procesos de lectura automatizada que se usaron para calificar la prueba [36] [5].

Tabla 3.13: Porcentaje por año y grado de primaria de escuelas con resultados 100 % confiables

	Grado				
Año	Tercero	Cuarto	Quinto	Sexto	Primaria
2010	67 %	68 %	74 %	75 %	54 %
2011	71 %	70 %	72 %	77 %	55 %
2012	66 %	72 %	75 %	76 %	55 %
2013	71 %	70 %	72 %	77 %	55 %

La tabla 3.13 muestra el porcentaje de escuelas por año sin ningún resultado poco confiable. Es decir, las escuelas en las cuales no se detectó ni un caso de copia. La columna de primaria presenta niveles menores a las otras columnas porque es posible que una escuela no haya presentado casos de "copia" en un grado pero en otro sí.

[5] Para garantizar la transparencia en la aplicación, los resultados son filtrados por un software de "detección de probabilidad de copia" que utiliza los métodos K-index y Scruting, que tienen como base patrones de respuestas incorrectas similares [37]

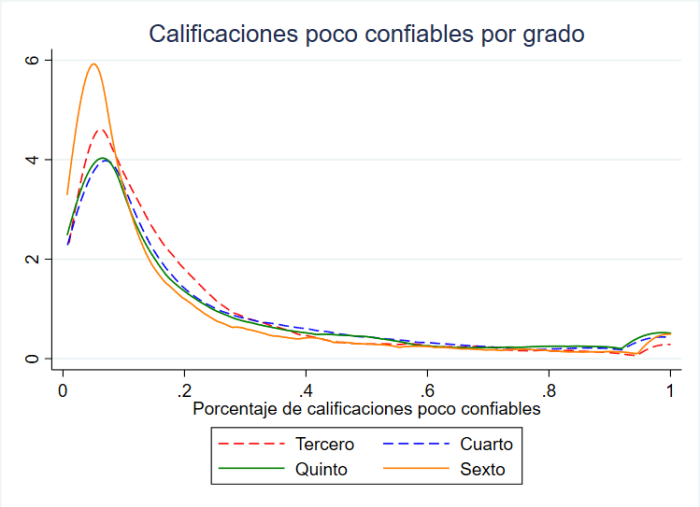


Figura 3.18: Porcentaje de copia por escuela

La figura 3.18 muestra como se distribuye el porcentaje de resultados poco confiables por grado.

Finalmente, a pesar de que resulta alarmante que en promedio solo el 45 % de las escuelas tengan al menos un resultado "poco confiable", en promedio solo el 5 % de los alumnos que presentan la prueba tienen resultados poco confiables (ver tabla 3.14).

Tabla 3.14: Porcentaje por año de alumnos con resultados poco confiables

Año	Porcentaje copio
2010	5.4 %
2011	4.8 %
2012	5.5 %
2013	4.8 %

Asimismo, resulta interesante observar la distribución de resultados poco confiables por estado. Para el 2013, quince estados no tuvieron resultados poco confiables mientras que el 38 % de los alumnos en Campeche, el 33 % de los alumnos en Tlaxcala y el 28 % de los alumnos en Sonora presentaron resultados poco confiables.

Es posible eliminar las escuelas con alto porcentaje de resultados poco confiables del análisis. Una alternativa es conseguir los resultados a nivel alumno y eliminar a los alumnos con la variable indicadora de "copia". Asimismo, se podrán eliminar las escuelas con un determinado porcentaje de copia.

Por otro lado, la cobertura "censal" no es total. Por ejemplo, el estado de Oaxaca participo en la prueba tres de ocho años y en el 2013 participaron sólo los centros comunitarios administrados a nivel federal por CONAFE. Otro ejemplo es el estado de Michoacan que no participó en la prueba del 2008. El apéndice A muestra el número de escuelas participantes por año y por estado.

3.4.2 Calidad del F911 y CEMABE

El levantamiento de información del INEGI se realizó del 26 de septiembre al 29 de noviembre del 2013. Mientras que la recopilación del formato 911 de inicio del 2013 fue del 19 de agosto hasta el 31 de diciembre. Cabe resaltar que las fechas se empalman y que el 95 % de los datos recuperados del F911 se llenaron entre el 26 de septiembre al 29 de noviembre del 2013 (mismas fechas del levantamiento del CEMABE).

Por un lado, el INEGI se realizó en 177,829 escuelas generales de nivel preescolar, primaria, secundaria, CAM o de educación especial. Del total número de escuelas censadas, el 44 % de son primarias (77,212 escuelas). Asimismo, se censaron 33,849 escuelas del CONAFE de las cuales el 32 % son primarias (10,936 escuelas). En total, en nivel primaria el INEGI contiene la información de 88,148 escuelas. Por otro lado, en el 2013 el F911 recopiló la información de 88,706 primarias generales, 10,193 primarias indígenas y 11,661 primarias comunitarias En total, el F911 contiene información de 110,560 escuelas primarias.

La tabla 3.15 muestra el porcentaje de escuelas de las tablas del CEMABE encontradas en las tablas del F911. Es decir, del 100 % de la tabla de Centros del CEMABE,

87 % de las escuelas también están en la tabla del F911 general y 8 % en la tabla del F911 Indígena. Por lo tanto, el 95 % de las escuelas de la tabla de Centros también están en la tabla del F911.

Tabla 3.15: Porcentaje de escuelas de las tablas del CEMABE encontradas en las tablas del F911

	F911		
CEMABE	General	Indigena	Comunitarias
Centros	87 %	8 %	-
CONAFE	-	-	88 %

La tabla 3.16 muestra los porcentajes inversos a la tabla 3.16. En este caso, la tabla dice qué porcentaje de las escuelas en las tablas del F911 también están en las tablas del CEMABE. En otras palabras, solo el 60 % de las escuelas indígenas que llenaron el F911 también fueron censadas por el INEGI.

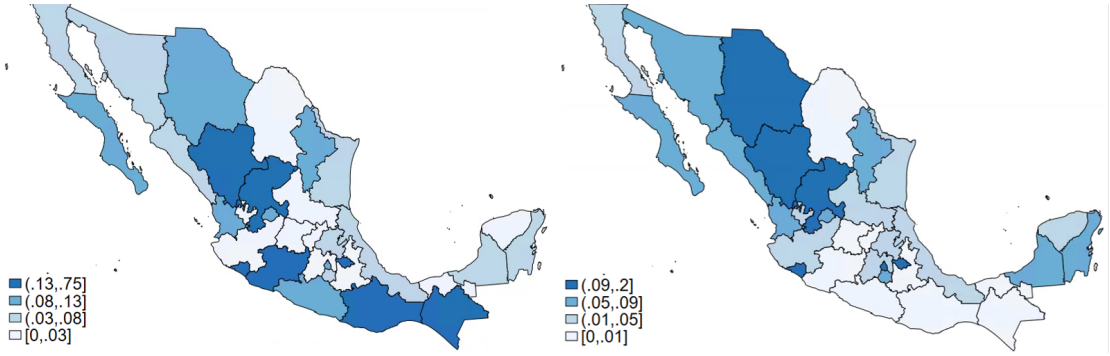
Tabla 3.16: Porcentaje de escuelas de las tablas del F911 encontradas en las tablas del CEMABE

	CEMABE	
F911	Centros	CONAFE
General	85 %	-
Indigena	60 %	-
Comunitarias	-	83 %

Las figuras 3.19a y 3.19b muestran el porcentaje [6] de escuelas por estado que no están en la intersección de las tablas. Por un lado, cabe destacar que de Querétaro solo una escuela que fue censada por el CEMABE no llenó el formato 911 y solo 5 escuelas que llenaron el formato 911 no fueron censadas por la INEGI. Por el otro lado, el 75 % de

[6]El porcentaje se calculó como el total de escuelas de una base que no están en la otra por estado entre el total de escuelas por estado de ambas bases.

las escuelas de Chiapas que están en la tabla del F911 no participaron ese mismo año en el CEMABE. Lo mismo ocurre con el 67 % y 49 % de las escuelas de Michoacán, Oaxaca respectivamente.



(a) Porcentaje de escuelas de las tablas del F911 que no están en el CEMABE

(b) Porcentaje de escuelas de las tablas del CEMABE que no están en el F911

Figura 3.19: Mapa coroplético del porcentaje de escuelas por estado que no están en la intersección de las bases

Ambas bases tienen una variable indicadora del sostenimiento de los centros de trabajo. En el F911, la variable se llama SOSTENIMIE y en el CEMABE control. Curiosamente, la variable de "ser privada o pública" para 144 escuelas es diferente en cada bases. El 0.21 % de las escuelas en el F911 están identificadas como públicas y son privadas. Las 144 escuelas (0.21 % del total) pertenecen al estado de Hidalgo. Es probable que la codificación del estado para ese año haya sido errónea ya que el tercer carácter del CCT indica el sostenimiento y en los 144 CCT, el tercer carácter es la letra "P" (privada).

Otra similitud es que ambos cuestionarios, el del F911 y el del CEMABE, incluyen la matricula de la escuela. En el formato F911 y en el CEMABE, las variable V347 y p166, respectivamente, indican el total de alumnos en primaria. Las variables tienen una correlación del 99 %, la figura 3.20 muestra una gráfica de dispersión por bloques de la variable de matrícula del F911 y del CEMABE.

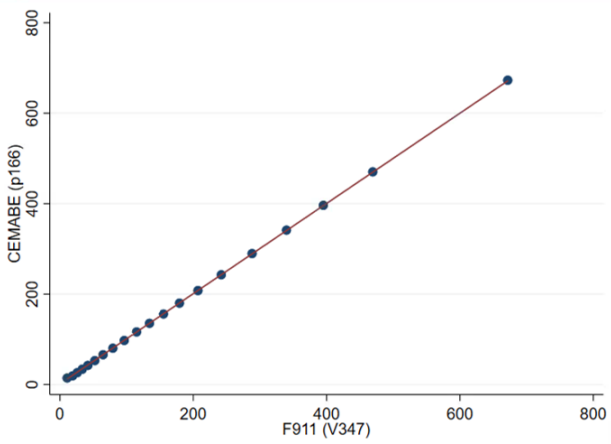


Figura 3.20: Gráfica de dispersión por bloques de la matricula por escuela en ambas bases

Es interesante notar que en el CEMABE la matrícula tiene, en promedio, 1.1 alumnos más por escuela que el F911. El estado con el menor error absoluto medio (MAE) es Hidalgo (MAE = 1.3) y el estado con el mayor error absoluto medio es Oaxaca (MAE = 15.9).

CAPÍTULO 4

PREPARACIÓN DE LOS DATOS

Una vez que se han entendido los datos, es posible seleccionar, limpiar, construir e integrar la información. En este capítulo, se utilizará la exploración y la verificación de calidad descrita en el capítulo anterior para construir los conjuntos de datos que utilizarán para construir modelos.

Dichos conjuntos costan de dos partes: la variable objetivo y las variables independientes. A continuación se detallará la preparación de ambos elementos.

4.1 Variable objetivo

La variable objetivo es la calificación de matemáticas en la prueba ENLACE a nivel escuela.

4.1.1 Limpieza de datos

La limpieza de la variable objetivo se realizó eliminando los resultados poco confiables y los valores atípicos.

En primer lugar, se eliminaron los resultados identificados como "copia" al calificar las pruebas.

En el capítulo anterior (en la sección 3.4.1) se detectó que, en promedio, solamente el 55 % de las escuelas tienen resultados completamente confiables. Esto es consecuencia del 5.1 % de los alumnos con resultados identificados como "copia".

Para limpiar la variable objetivo, se escogió perder el 5.1 % de los datos. Es decir, se

eliminaron los resultados de los estudiantes que "copiaron" con el fin de no incluirlos en el promedio de la escuela.

Asimismo, para no perder información sobre las "copias" en la escuela, se creo una nueva variable indicando el porcentaje de alumnos que copiaron por escuela. Como se ve en la figura 4.1, la mayoría de las escuelas tuvieron un porcentaje bajo de "copia".

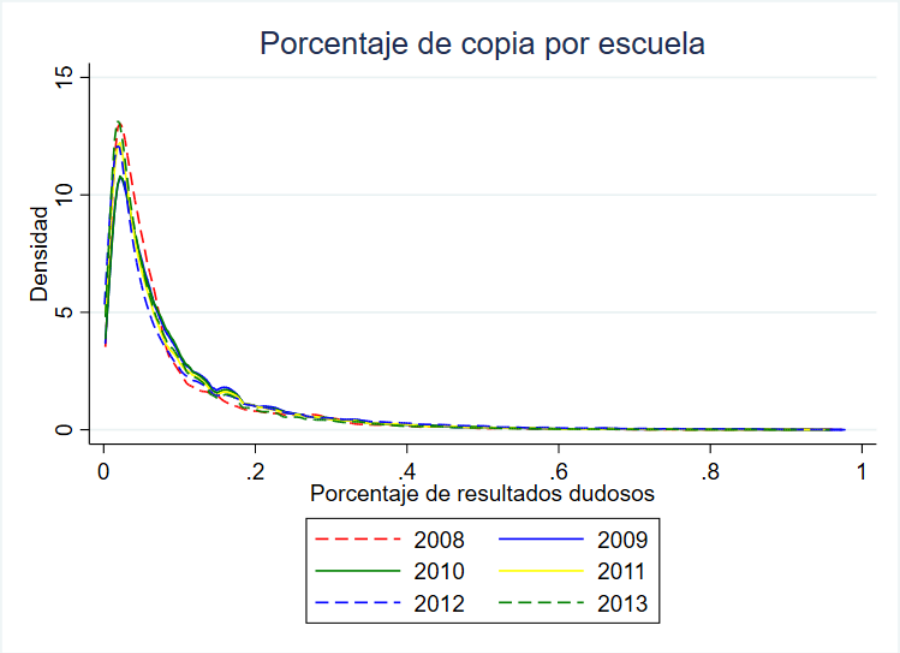


Figura 4.1: Distribución de los porcentajes de alumnos que "copiaron" por escuela

Nota: Esta gráfica solo incluye escuelas con uno o más resultados identificados como "copia"

De igual modo, para tener resultados más confiables, se eliminaron los resultados de los años en los que una escuela tuvo un porcentaje de copia mayor a 50 %. La tabla 4.1 muestra el porcentaje de escuelas que fueron eliminadas por año. En total, solo se pierde información de 26 escuelas que tuvieron en todos los años más de 50 % de resultados poco confiables. El resto de las escuelas, tuvieron al menos un año con 50 % o más resultados confiables.

En segundo lugar, se eliminaron los valores atípicos. Es decir, las observaciones de

Tabla 4.1: Escuelas con más de 50 % de resultados "copia"

Año	Escuelas	% total
2008	360	0.40
2009	550	0.62
2010	544	0.61
2011	476	0.53
2012	807	0.96
2013	364	0.41
Total	3101	0.44

alumnos con calificaciones fuera del rango.

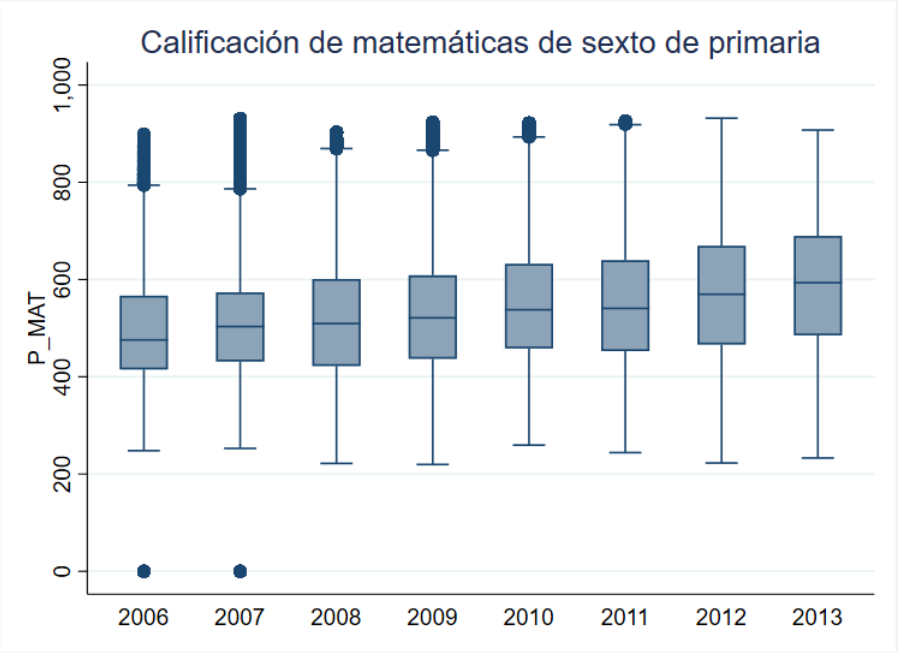


Figura 4.2: Diagrama de cajas y bigotes de las calificaciones de sexto de primaria por año

La gráfica 4.2 muestra las calificaciones de matemáticas de sexto de primaria por año. Se calcularon los extremos superiores e inferiores para cada grado en cada año siguiendo la definición de extremos de una gráfica de cajas y bigotes [38]. Las ecuaciones 4.1 y 4.2 fueron utilizadas para obtener los extremos inferiores y superiores. En las ecuaciones, Q1 es el primer quartil, es decir el percentil 25; Q3 es el tercer quartil

(percentil 75) y IQR es el rango interquartil (Q3 - Q1).

$$inferior = Q1 - 1.5IQR \tag{4.1}$$

$$superior = Q3 + 1.5IQR \tag{4.2}$$

Tabla 4.2: Porentaje de calificaciones atípicas por año y grado

Grado	2006	2007	2008	2009	2010	2011	2012	2013
3	-	-	-	-	-	-	-	-
4	0.30	0.45	-	-	-	-	-	-
5	1.60	1.55	0.04	0.01	0.01	-	-	-
6	2.24	1.81	0.08	0.05	0.03	0.12	-	-

Nota: El porcentaje se calculó como el número de alumnos arriba del limite superior más el número de alumnos abajo del límite inferior, entre el total de alumnos por grado

La tabla 4.2 muestra el porcentaje de calificaciones de alumnos eliminadas por ser consideradas valores atípicos del año y grado. El porcentaje es más alto para 2006 y 2007 por su comportamiento multi-modal.

4.1.2 Construcción de nuevos datos

Después de eliminar las observaciones atípicas y "tramposas", se calculó la calificación promedio por grado para cada año y escuela.

Como muestra la gráfica 4.2, cada año tiene rangos y valores extremos diferentes. Por lo tanto, estandarizar las calificaciones por grado y por año permite una comparación más justa.

Más adelante, se calculó el promedio por escuela de las calificaciones estandarizadas

por grado. La figura 4.3 muestra las distribuciones estandarizadas por año.

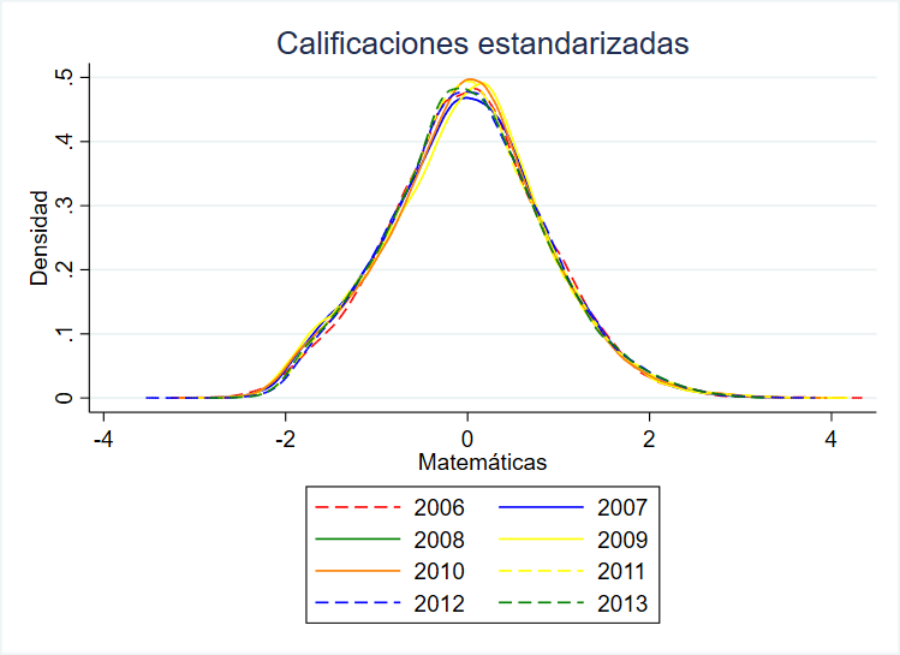


Figura 4.3: Distribución de calificaciones estandarizadas

Finalmente, la variable objetivo es el promedio de las calificaciones de todos los años por escuela. Si una escuela solo participó un año, solo se utilizará ese valor, mientras que para las escuelas que participaron ocho años, se utiliza la calificación promedio. En total, se tiene la calificación promedio de 113,894 escuelas primarias.

4.2 Variables independientes

Las variables independientes son las características de las escuelas, inmueble, profesores y alumnos. La fuente de estas variables son el INEGI y en el F911.

El proceso de preparación de datos fue iterativo. Primero se excluyeron variables poco relevantes con poca varianza o información y se identificaron las más significativas. Estas variables se limpiaron y se utilizaron para la construcción de datos. Una vez integrados los datos se volvieron a seleccionar y limpiar las variables más importantes para construir modelos.

4.2.1 Selección de datos

A primera vista parece que un mayor número de variables es deseable porque se tiene más información. Sin embargo, la maldición de la dimensionalidad (curse of dimensionality, en inglés) indica que el número de observaciones necesarias incrementa exponencialmente con el número de variables. Por lo tanto, dado el alto número de variables y el número de observaciones fijo, es necesario reducir el número de variables [39].

Las técnicas de reducción de dimensionalidad tienen como objetivo decrementar el número de variables aleatorias y obtener variables principales. Existen las siguientes dos técnicas para lograr esto: selección de características y extracción de características.

Extracción de características

Para empezar, la primera técnica no modifica las variables sino que selecciona las más relevantes. Esta técnica se divide en exclusión e inclusión. De entrada, se puede excluir variables no relacionadas con el desempeño académico como el número exterior en la dirección de la escuela o la fecha de levantamiento de la encuesta.

Se examinaron las variables no numéricas y se eliminaron aquellas que tuvieran información duplicada o poco relevante. Por ejemplo, la variable "n_estado" indica el nombre de la entidad federativa en la que se encuentra la escuela, esta variable contiene la misma información que ïd_estado"que asigna un número a cada estado.

Más adelante, se excluyeron las variables con varianza baja. Se normalizaron las variables númericas y utilizando la función VarianceThreshold() se escogió un threshold de 0.001 para eliminar aquellas variables con varianza menor a 0.001.

La selección de variables se llevó a cabo en dos etapas: la etapa inicial que ayudó a

la construcción de nuevas variables y la etapa final seleccionó las variables para los modelos. En ambos casos, se utilizó el siguiente método para seleccionar las variables:

1. Método de filtración: Se escogieron "k" variables con mayor correlación con la variable objetivo en términos absolutos.

2. Método de envoltura (del inglés, wrapper method): Las "k" variables fueron utilizadas en diferentes modelos de los cuales se seleccionaron las "n" mejores. Los modelos utilizados fueron los siguientes:

   - Mínimos cuadrados ordinarios (OSL)
   - Bosque aleatorio
   - Eliminación de características recursivas (RFE)
   - SelectKBest con regresiones lineales

3. Método incrustados (del inglés, embedded method): Todas las "n" mejores variables seleccionadas por los diferentes métodos de envoltura se utilizaron para crear un nuevo conjunto de datos. De este nuevo conjunto de datos se seleccionaron variables utilizando los regularizadores Lasso y Ridge. Del mismo modo que con los métodos de envoltura, ambos regularizadores escogieron las "n" mejores variables y el conjunto de datos final se construyó con la intersección de variables seleccionadas.

En primer lugar, se escogió filtrar las variables por correlación para seleccionar de forma rápida, aunque menos precisa, variables con información importante [40]. Más adelante, los métodos de envoltura, que son costosos computacionalmente, se combinaron para complementar los beneficios y deficiencias de cada uno. Por ejemplo, OSL identifica relaciones lineales mientras que el bosque aleatorio identifica interacciones. Finalmente, los regularizadores utilizados como métodos incrustados sirvieron para

reducir la complejidad del modelo. Ambos regularizadores reducen el tamaño de los coeficientes, Lasso reduce el valor a cero y Ridge los acerca a cero, por eso se escogieron las "n" variables con mayor coeficiente en Ridge y en Lasso todas aquellas con coeficientes diferentes a cero.

Los valores de "k' y "n" se escogieron con base en el número de observaciones y variables disponibles. En la primera etapa, "k' y "n" fueron más grandes para examinar un mayor número de variables y construir variables significativas.

4.2.2 Limpieza de datos

La limpieza de datos también se realizó en dos etapas: superficial y profunda.

En la primer etapa, se hizo una limpieza simple de codificación y rellenar valores nulos. Por ejemplo, en el CEMABE, el identificador de "No especificado" en algunos casos es el número (9) nueve y en otros casos el novecientos noventa y nueve (999). Estos valores fueron reemplazados por valores nulos.

En cuanto a los valores nulos, en el F911, en la sección de edades por grado, faltan muchos valores. Sin embargo, se pueden inferir con las variables alrededor. Por ejemplo, si el total de alumnos de sexto de primaria es treinta y treinta niños tienen doce años, entonces cero niños tienen once años.

En la segunda etapa, una vez escogidas las variables principales, se examinaron con cuidado. Es decir, en vez de asumir que los valores faltantes eran cero, se utilizaron técnicas de imputación de datos.

La imputación multi-variada por ecuaciones en cadena (MICE) es una alternativa para tratar los valores nulos ya que al imputar muchos valores, disminuye la incertidumbre estadística [41]. Una desventaja del método de imputación de datos MICE es que hace suposiciones sobre las distribuciones de los datos. Una mejor alternativa es utilizar

"Miss Forest", un método de imputación no paramétrico que soluciona el problema entrenando bosques aleatorios con los valores observados, haciendo predicciones y repitiendo el proceso iterativamente [42]. "Miss Forest" se implementó utilizando la función ExtraTreesRegressor de sklearn.ensamble.

4.2.3 Construcción de nuevos datos

De forma similar a las otras secciones de preparación de los datos, se construyeron en dos etapa. La etapa inicial se basó en la exploración de datos, conocimiento del campo y en los datos seleccionados por primera vez. La etapa final utilizó técnicas de reducción de dimensionalidad sobre las variables principales.

En la primera etapa, se construyeron datos con el formato 911 y el CEMABE.

Cabe resaltar que los datos del formato 911 fueron registrados a nivel escuela en términos absolutos. Sin embargo, puede resultar más informativo y más comparable conocer las proporciones o porcentajes. Por ejemplo, el número de maestros por alumno puede dar más información que el número de maestros en una escuela.

Asimismo, existen muchas variables que pueden ser resumidas. Por ejemplo, el desglose de edades por grado se puede resumir como la edad promedio por grado. Tomando esto en cuenta, se construyeron variables basándose en la literatura en conocimiento del sector.

A continuación, se en listan algunas de las variables construidas con el formato 911. La descripción completa de todas las variables generadas se encuentra en el apéndice B.

- Proporción mujeres-hombre de alumnos (alumnas mujeres / alumnos hombres) [43].
- Porcentaje de maestros con grado igual o mayor a licenciatura [43].

- Número de alumnos por maestro [44].
- Porcentaje de alumnos repitiendo grado [45].
- Número de años promedio en preescolar [46]

De forma similar, la descripción completa de algunas variables generadas a partir del CEMABE se encuentra en el apéndice C.

En la segunda etapa, una vez seleccionadas las variables más importantes, fue posible crear un nuevo conjunto de datos a partir del conjunto original.

Se exploraron tres técnicas: análisis de componentes principales (PCA), análisis de componentes independientes (ICA) y ensamble de vecinos estocásticos distribuidos (t-SNE del inglés t- Distributed Stochastic Neighbor Embedding). La primera técnica, el análisis de componentes principales (PCA) reduce la dimensionalidad utilizando transformaciones ortogonales y crea un nuevo conjunto con valores sin correlación lineal llamado componentes principales que es capaz de explicar un gran porcentaje de la varianza de los datos. La segunda técnica, el análisis de componentes independientes (ICA), busca factores independientes a diferencia de la técnica PCA que busca factores sin correlación. Finalmente, la técnica de ensamble de vecinos estocásticos distribuidos (t-SNE del inglés t- Distributed Stochastic Neighbor Embedding) a diferencia de PCA e ICA busca patrones no lineales desde un enfoque local y global [47].

Se crearon 4 variables con los vectores principales de PCA y lograron describir el 80 % de la variaza en los datos. Sin embargo, la mayor desventaja de estas técnicas es que se pierde la interpretabilidad de los modelos.

[JUAN, DE ESTO NO ESTOY NADA SEGURA] Se escogieron las variables de la segunda etapa ya que la métrica es el criterio de información de Akaike que favorece la simplicidad y el menor número de variables.

4.3 Integración de datos

En primera instancia, se intentó hacer un conjunto maestro con los datos del F911, CEMABE y ENLACE. Sin embargo, las preguntas del F911 son diferentes para cada tipo de primaria (general, comunitaria e indigena) y el CEMABE también tiene un formato para escuelas geneales e indigenas y otro para escuelas comunitarias. Existen algunas variables en común y sí es posible construir el conjunto maestro. Sin embargo, puede ser más útil para el análisis tratar cada tipo de escuela por separado.

Por lo tanto, se integraron los datos en tres conjuntos: conjunto de escuelas generales, conjunto de escuelas indígenas y conjunto de escuelas comunitarias.

En los tres conjuntos, se juntaron las calificaciones de ENLACE, el formato 911 y el CEMABE correspondiente utilizando el Clave de Centro de Trabajo (CCT) y el turno como identificadores únicos.

Estos tres conjuntos fueron los utilizados para seleccionar variables y limpiar los datos por separado. Cabe recordar que el universo de escuelas de ENLACE, CEMABE y F911 son diferentes. Es decir, las escuelas no son las mismas, por lo tanto se perdió información de algunas escuelas al integrar los datos.

La tabla 4.3 muestra las dimensiones de los conjuntos integrados, el número de variables originales y el número después de la reducción.

Tabla 4.3: Observaciones y variables de conjuntos integrados

Escuela	Observaciones	Variables originales	Variables seleccionadas
General	64,911	1,458	53
Indígena	5,206	1,257	18
Comunitaria	9,139	692	21

4.4 Dar formato a los datos

Dado que las variables independientes tienen diferentes escalas, estas se normalizaron para equilibrar la importancia de las variables y disminuir el costo computacional acelerando los cálculos. Asimismo, se crean variables indicadoras para las variables categóricas.

Appendices

APÉNDICE A

CALIDAD ENLACE

Tabla A.1: Número de escuelas por estado y año en ENLACE

	año						
Estado	2006	2007	2008	2009	2010	2011	2012
Aguascalientes	671	713	683	690	685	693	685
Baja California	714	1,481	1,530	1,563	1,602	1,653	1,652
Baja California Sur	215	372	335	347	360	364	376
Campeche	662	675	670	674	677	681	681
Coahuila	1,696	1,717	1,717	1,744	1,758	1,779	1,783
Colima	416	440	426	427	434	438	439
Chiapas	469	3,460	5,890	6,217	5,770	6,253	3,999
Chihuahua	2,225	2,508	2,417	2,469	2,472	2,472	2,461
Distrito Federal	3,218	3,344	3,337	3,314	3,310	3,266	3,242
Durango	2,010	1,984	2,035	2,069	2,067	2,064	2,078
Guanajuato	4,585	4,649	4,623	4,631	4,657	4,689	4,313
Guerrero	4,087	3,874	3,881	2,436	3,219	3,484	606
Hidalgo	3,073	3,118	2,706	2,723	2,722	2,710	2,716
Jalisco	5,364	5,704	5,714	5,782	5,801	5,817	5,447
México	3,337	7,248	7,585	7,603	7,690	7,706	7,452
Michoacán	1,528	2,566	-	2,037	2,127	2,331	977
Morelos	949	976	977	933	1,017	1,027	1,017
Nayarit	982	1,130	1,139	1,144	1,166	1,177	983
Nuevo León	2,405	2,481	2,505	2,547	2,602	2,655	2,689
Oaxaca	-	3,557	4,175	646	-	-	-
Puebla	3,577	4,390	3,983	3,477	4,114	4,197	4,190
Querétaro	1,188	1,319	1,197	1,222	1,236	1,241	1,242
Quintana Roo	710	718	734	745	759	778	779
San Luis Potosí	3,159	2,930	2,758	2,747	2,747	2,756	2,772
Sinaloa	1,477	2,407	2,303	2,321	2,336	2,372	2,331
Sonora	1,667	1,699	1,731	1,730	1,747	1,617	1,761
Tabasco	1,868	2,120	1,923	1,927	1,924	1,922	1,920
Tamaulipas	2,250	2,335	2,246	2,241	2,271	2,319	2,318
Tlaxcala	643	385	622	672	760	718	691
Veracruz	8,594	9,460	8,593	8,627	8,630	8,565	8,670
Yucatán	1,324	1,221	1,217	1,228	1,254	1,270	1,280
Zacatecas	1,551	1,787	1,791	1,992	1,778	1,774	1,705

APÉNDICE B

INGENIERÍA DE CARACTERÍSTICAS: VARIABLES F911

Utilizando el formato 911 se crearon las siguientes variables

Nombre de variable	Descripción	¿Cómo se construyó?	Supuestos o comentarios
meanAge_i	Edad promedio por grado escolar. El índice "I" representa un grado escolar. Por ejemplo, meanAge_2 es la edad promedio de los alumnos de segundo de primaria.	En cada grado se multiplicó la edad de los alumnos por el número de alumnos de esa edad y se dividió entre el número de alumnos total de ese grado.	Los alumnos en la categoría "menos de 6 años" se contaron como si tuvieran 5 años y los alumnos en la categoría "más de 15" se contaron como si tuvieran 16 años.
meanAge_prim	Edad promedio de todos los alumnos de primaria de la escuela	Se calculó el promedio de meanAge_i. Es decir, se sumó meanAge_1, meanAge_2, meanAge_3, meanAge_4, meanAge_5, meanAge_6 y se dividió entre 6.	Los alumnos en la categoría "menos de 6 años" se contaron como si tuvieran 5 años y los alumnos en la categoría "más de 15" se contaron como si tuvieran 16 años.
classSize_i	Número de alumnos en cada grupo por grado escolar. El índice "I" representa un grado escolar. Por ejemplo, classSize_2 es el número de alumnos de segundo de primaria en cada grupo.	Se dividió el total de alumnos por grado entre el número de grupos en cada grado. Por ejemplo, el total de alumnos de segundo de primaria entre el número de grupos de segundo de primaria.	
classSize_prim	Promedio de alumnos por grupo en toda primaria	Se calculó el promedio de classSize _i. Es decir, se sumó classSize _1, classSize _2, classSize _3, classSize _4, classSize _5, classSize _6 y se dividió entre 6.	
totalGirls	Número total de alumnos de género femenino en primaria	Se sumó el total de mujeres de nuevo ingreso más el total de mujeres repetidoras.	
totalBoys	Número total de alumnos de género masculino en primaria	Se sumó el total de hombres de nuevo ingreso más el total de hombres repetidores.	
sexRatio_students	Proporción de alumnos hombres y alumnas mujeres	Se dividió el total de alumnos hombres (totalBoys) entre el total de alumnas mujeres (totalGirls)	

repeatersRatio	Proporción de alumnos que son repetidores	Se dividió el total de alumnos repetidores entre el total de alumnos de primaria.	La variable está entre cero y uno.
preschoolAttendance	Proporción de alumnos de primero de primaria que fueron a una escuela preescolar antes de entrar a primaria.	Se dividió el total de alumnos de primero de primaria que habían cursado preescolar entre el total de alumnos en primero de primaria	La variable está entre cero y uno.
meanPreschoolYears	Número de años promedio de preescolar de los alumnos de primero de primaria	Se multiplicó el total número de alumnos de nuevo ingreso más repetidores por el número de años de preescolar cursados (un año, dos años o tres años) y se dividió entre el total de alumnos que cursaron preescolar.	El valor mínimo posible es cero y el máximo es tres.
meanPreschoolYears_girls	Número de años promedio de preescolar de las alumnas de género femenino de primero de primaria	Se multiplicó el número de alumnos mujeres de nuevo ingreso más las mujeres repetidoras por el número de años de preescolar cursados (un año, dos años o tres años) y se dividió entre el total de mujeres que cursaron preescolar.	El valor mínimo posible es cero y el máximo es tres.
meanPreschoolYears_boys	Número de años promedio de preescolar de las alumnas de género masculino de primero de primaria	Se multiplicó el número de alumnos hombres de nuevo ingreso más las mujeres repetidoras por el número de años de preescolar cursados (un año, dos años o tres años) y se dividió entre el total de hombres que cursaron preescolar.	El valor mínimo posible es cero y el máximo es tres.
indigenousStudents	Proporción de alumnos indígenas	Se dividió el número de alumnos indígenas entre el total de alumnos	La variable está entre cero y uno.
indigenousStudents_girls	Proporción de alumnas mujeres indígenas	Se dividió el número de alumnas mujeres indígenas entre el total mujeres.	La variable está entre cero y uno.
indigenousStudents_boys	Proporción de alumnos hombres indígenas.	Se dividió el número de alumnos hombres indígenas entre el total hombres.	La variable está entre cero y uno.
foreignStudents	Proporción de alumnos extranjeros.	Se dividió el número de alumnos extranjeros entre el total de alumnos.	La variable está entre cero y uno.
foreignStudents_girls	Proporción de alumnas mujeres extranjeras.	Se dividió el número de alumnas mujeres extranjeras entre el total mujeres.	La variable está entre cero y uno.

foreignStudents_boys	Proporción de alumnos hombres extranjeros.	Se dividió el número de alumnos hombres extranjeros entre el total hombres.	La variable está entre cero y uno.
foreignStudents_usa	Proporción de alumnos extranjeros originarios de Estados Unidos.	Se dividió el número de alumnos extranjeros provenientes de Estados Unidos entre el total de alumnos.	
foreignStudents_canada	Proporción de alumnos extranjeros originarios de Canadá.	Se dividió el número de alumnos extranjeros provenientes de Canadá entre el total de alumnos.	
foreignStudents_centralA	Proporción de alumnos extranjeros originarios de Centroamérica y el Caribe.	Se dividió el número de alumnos extranjeros provenientes de Centroamérica y el Caribe entre el total de alumnos.	
foreignStudents_southA	Proporción de alumnos extranjeros originarios de Sudamérica.	Se dividió el número de alumnos extranjeros provenientes de Sudamérica entre el total de alumnos.	
foreignStudents_africa	Proporción de alumnos extranjeros originarios de África.	Se dividió el número de alumnos extranjeros provenientes de África entre el total de alumnos.	
foreignStudents_asia	Proporción de alumnos extranjeros originarios de Asia.	Se dividió el número de alumnos extranjeros provenientes de Asia entre el total de alumnos.	
foreignStudents_europe	Proporción de alumnos extranjeros originarios de Europa.	Se dividió el número de alumnos extranjeros provenientes de Europa entre el total de alumnos.	
foreignStudents_oceania	Proporción de alumnos extranjeros originarios de Oceanía.	Se dividió el número de alumnos extranjeros provenientes de Oceanía entre el total de alumnos.	
usaerStudents	Proporción de alumnos atendidos por la Unidad de Servicios de Apoyo a la Educación Regular (USAER).	Se dividió el número de alumnos atendidos por la Unidad de Servicios de Apoyo a la Educación Regular (USAER) entre el total de alumnos.	
usaerStudents_girls	Proporción de alumnas mujeres atendidas por la Unidad de Servicios de Apoyo a la Educación Regular (USAER).	Se dividió el número de alumnas mujeres atendidas por la Unidad de Servicios de Apoyo a la Educación Regular (USAER) entre el total de mujeres.	
usaerStudents_boys	Proporción de alumnos hombres atendidos por la Unidad de Servicios de	Se dividió el número de alumnos hombres atendidos por la Unidad de Servicios de Apoyo a	

	Apoyo a la Educación Regular (USAER).	la Educación Regular (USAER) entre el total de hombres.	
disabilitiesStudents	Proporción de alumnos con alguna discapacidad o con capacidades y aptitudes sobresalientes.	Se dividió el número de alumnos con alguna discapacidad o con capacidades y aptitudes sobresalientes entre el total de alumnos.	
disabilitiesStudents_girls	Proporción de alumnas mujeres con alguna discapacidad o con capacidades y aptitudes sobresalientes.	Se dividió el número de alumnas mujeres con alguna discapacidad o con capacidades y aptitudes sobresalientes entre el total de mujeres.	
disabilitiesStudents_boys	Proporción de alumnos hombres con alguna discapacidad o con capacidades y aptitudes sobresalientes.	Se dividió el número de alumnos hombres con alguna discapacidad o con capacidades y aptitudes sobresalientes entre el total de hombres.	
disabilitiesStudents_blindness	Proporción de alumnos con ceguera.	Se dividió el número de alumnos con ceguera entre el total de alumnos.	
disabilitiesStudents_vision	Proporción de alumnos con discapacidad visual.	Se dividió el número de alumnos con discapacidad visual entre el total de alumnos.	
disabilitiesStudents_deaf	Proporción de alumnos con sordera.	Se dividió el número de alumnos con sordera entre el total de alumnos.	
disabilitiesStudents_hearing	Proporción de alumnos con discapacidad auditiva.	Se dividió el número de alumnos con discapacidad auditiva entre el total de alumnos.	
disabilitiesStudents_mobility	Proporción de alumnos con discapacidad motriz.	Se dividió el número de alumnos con discapacidad motriz entre el total de alumnos.	
disabilitiesStudents_intelectual	Proporción de alumnos con discapacidad intelectual.	Se dividió el número de alumnos con discapacidad intelectual entre el total de alumnos.	
disabilitiesStudents_genius	Proporción de alumnos con capacidades y aptitudes sobresalientes.	Se dividió el número de alumnos con capacidades y aptitudes sobresalientes entre el total de alumnos.	
specialStudents	Proporción de alumnos con Necesidades Educativas Especiales (NEE).	Se dividió el número de alumnos con Necesidades Educativas Especiales (NEE) entre el total de alumnos.	
specialStudents_girls	Proporción de alumnas mujeres con Necesidades Educativas Especiales (NEE).	Se dividió el número de alumnas mujeres con Necesidades Educativas Especiales (NEE) entre el total de mujeres.	

specialStudents_boys	Proporción de alumnos hombres con Necesidades Educativas Especiales (NEE).	Se dividió el número de alumnos hombres con Necesidades Educativas Especiales (NEE) entre el total de hombres.	
totalTeachers	Número total de maestros docentes.	Se sumó el total de hombres de personal docente más el total de mujeres de personal docente	
studentTeacherRatio	Número de alumnos por maestro.	Se dividió el total de alumnos entre el total de profesores.	
sexRatio_teachers	Proporción de maestros hombres y maestros mujeres.	Si dividió el número de personal docente de hombres entre el número de personal docente de mujeres.	
normalTeachers	Proporción de maestros del nivel educativo "Normal"	Se sumó el número de maestros en el nivel educativo "normal" incluyendo normal preescolar terminada, normal primaria incompleta, normal primaria terminada, normal superior incompleta, normal superior pasante y normal superior titulado	
meanSchoolYears_teachers	Promedio de número de años de educación de los profesores.	Se multiplicó el número de años por el número de profesores en cada nivel y se dividió entre el total de profesores	Número de años por grado escolar: Primaria incompleta (3 años), primaria terminada (6 años), secundaria incompleta (7 años), secundaria terminada (9 años), profesional técnico (12 años), bachillerato incompleto (10 años), bachillerato terminado (12 años), normal preescolar incompleta (14 años) , normal preescolar terminada (16 años), normal primaria incompleta (14 años), normal primaria terminada (16 años), normal superior incompleta (14 años), normal superior pasante (16 años), normal superior titulado (16 años), licenciatura incompleta (14 años), licenciatura pasante

			(16 años), licenciatura titulado (16 años), maestría incompleta (17 años), maestría graduado (18 años), doctorado incompleto (19 años), doctorado graduado (22 años)
graduateTeachers	Proporción de personal docente con grado mayor o equivalente a licenciatura.	Se dividió el número de personal docente con titulo entre el número de personal docente total.	
directivesTeachers	Proporción de personal directivo con grupo	Se dividió el número de personal directivo con grupo entre el número de personal directivo total (con y sin grupo).	
sexRatio_directives	Proporción de personal directivo hombres y personal directivo mujeres.	Se dividió el número de personal directivo de hombres entre el número de personal directivo de mujeres.	
sexRatio_PE	Proporción de profesores de educación física hombres y profesores de educación física mujeres.	Se dividió el número de profesores de educación física hombres entre el número de profesores de educación física mujeres.	
sexRatio_art	Proporción de profesores de actividades artísticas hombres y profesores de actividades artísticas mujeres.	Se dividió el número de profesores de actividades artísticas hombres entre el número de profesores de actividades artísticas mujeres.	
sexRatio_techno	Proporción de profesores de actividades tecnológicas hombres y profesores de actividades tecnológicas mujeres.	Se dividió el número de profesores de actividades tecnológicas hombres entre el número de profesores de actividades tecnológicas mujeres.	
sexRatio_lang	Proporción de profesores de idiomas hombres y profesores de idiomas mujeres.	Se dividió el número de profesores de idiomas hombres entre el número de profesores de idiomas mujeres.	
specialTeachers	Proporción de personal docente especial del total de personal.	Se dividió el número de personal docente especial entre el total de personal.	
adminStaff	Proporción de personal administrativo, auxiliar y	Se dividió el número de personal administrativo, auxiliar y de	

	de servicios del total de personal.	servicios entre el total de personal.	
studentStaffRatio	Número de personal por alumno.	Se dividió el número de alumnos entre el personal total.	
studentTeacher_PE	Número de profesores de educación física por alumno.	Se dividió el número de alumnos entre el número de profesores de educación física.	
studentTeacher_art	Número de profesores de actividades artísticas por alumno.	Se dividió el número de alumnos entre el número de profesores de actividades artísticas.	
studentTeacher_techno	Número de profesores de actividades tecnológicas por alumno.	Se dividió el número de alumnos entre el número de profesores de actividades tecnológicas.	
studentTeacher_lang	Número de profesores de idiomas por alumno.	Se dividió el número de alumnos entre el número de profesores de idiomas.	
hoursPE	Cantidad de horas impartidas a la semana por el personal docente especial de educación física.	Renombrar variable V872	
hoursArt	Cantidad de horas impartidas a la semana por el personal docente especial de actividades artísticas.	Renombrar variable V873	
hoursTechno	Cantidad de horas impartidas a la semana por el personal docente especial de actividades tecnológicas.	Renombrar variable V874	
hoursLang	Cantidad de horas impartidas a la semana por el personal docente especial de idiomas.	Renombrar variable V875	
carrMagisterial	Proporción de personal con carrera magisterial	Se dividió la cantidad de profesores que se encuentran en el programa de carrera magisterial entre el personal total.	Directivos, personal administrativo y docente especial pueden estar en el programa de carrera magisterial
nivelCarrMagis_1V	Nivel promedio de profesores frente a grupo con carrera magisterial. Estos se encuentran en la primera vertiente.	Se multiplicaron los niveles por el número de profesores de cada nivel, dividido entre el número de profesores frente a grupo.	Ponderación por nivel: Nivel A = 1 Nivel B = 2 Nivel BC = 3 Nivel C = 4 Nivel D = 5 Nivel E = 6

nivelCarrMagis	Nivel promedio de profesores frente a grupo, Docentes en funciones directivas y de supervisión y docentes en actividades técnico-pedagógicas con carrera magisterial. Estos se encuentran en la primera, segunda y tercera vertiente.	Se multiplicaron los niveles por el número de profesores de cada nivel, dividido entre el total de profesores.	Ponderación por nivel: Nivel A = 1 Nivel B = 2 Nivel BC = 3 Nivel C = 4 Nivel D = 5 Nivel E = 6
classroomInUse	Proporción de aulas en uso del total de aulas existentes.	Se dividió el número de aulas en uso entre el número de aulas existentes.	
studentClassroom	Proporción de estudiantes en aulas en uso	Se dividió el total de alumnos entre el número de aulas en uso.	
adaptedClassrooms	Proporción de aulas en uso que están adaptadas.	Se dividió el número de aulas adaptadas entre el número de aulas en uso.	
schoolCost	Monto promedio de dinero que gasta cada alumno (o los padres del alumno) en un determinado concepto, durante el ciclo escolar. Se aplica a los siguientes conceptos: inscripción, paquete de útiles y libros (cuando éstos se soliciten) y uniformes. Asimismo, se aplica a cuotas que requieran un desembolso para las familias; por ejemplo, las aportaciones a la asociación de padres de familia o alguna ayuda para el arreglo de la escuela o para equipar laboratorios y talleres, etcétera.	Se sumó el gasto promedio anual en el paquete de útiles y libros que se sugiere adquiera el alumno más el gasto promedio anual en uniformes que se sugiere adquiera el alumno más el gasto promedio anual en cuotas	
privateTuition	Gasto promedio anual de escuelas particulares.	Se sumó gasto promedio anual en inscripción más el gasto promedio mensual en colegiatura por el número de mensualidades que se pagan	

		más el gasto promedio mensual del servicio de transporte por el número de mensualidades que se pagan.	
transportCost	Gasto anual en transporte escolar.	El gasto promedio mensual del servicio de transporte por el número de mensualidades que se pagan.	

Únicamente escuelas indígenas.

Nombre de variable	Descripción	¿Cómo se construyó?
indigena	Variable indicadora de educación primaria indígena	Se asignó el valor 1 en toda la tabla.
escuelaAlbergue	Variable indicadora en caso de la escuela sea un albergue	Renombrar variable V1.
primIndigena	Variable indicadora en caso de la escuela sea una primaria indigena	Renombrar variable V2.
mt_languageTeachers	Proporción de maestros que hablan alguna lengua materna entre el total de maestros	Se sumó el número de maestros parlantes de cada lengua materna y se dividió entre el número total de maestros.
mt_speakTeachers	Proporción de maestros que hablan alguna lengua materna de la comunidad entre el total de maestros	Se sumó el número de docentes que hablan la lengua materna de la comunidad y dividió entre el número total de maestros.
mt_readTeachers	Proporción de maestros que hablan la lengua materna de la comunidad entre el total de maestros	Se sumó el número de docentes que hablan la lengua materna de la comunidad y se dividió entre el número total de maestros.
mt_writeTeachers	Proporción de maestros que escriben la lengua materna de la comunidad entre el total de maestros	Se sumó el número de docentes que pueden escribir la lengua materna de la comunidad y se dividió entre el número total de maestros.
mt_speakStaff	Proporción de personal que hablan alguna lengua materna de la comunidad entre el total del personal.	Se sumó el número de personal que hablan la lengua materna de la comunidad y dividió entre el número total de personal.
mt_readStaff	Proporción de personal que hablan la lengua materna de la comunidad entre el total de personal.	Se sumó el número de personal que hablan la lengua materna de la comunidad y se dividió entre el número total de personal.

mt_writeStaff	Proporción de personal que escriben la lengua materna de la comunidad entre el total de personal.	Se sumó el número de personal que pueden escribir la lengua materna de la comunidad y se dividió entre el número total de personal.
promoters	Proporción de personal cuya función es ser promotor.	Se sumó el número de promotores mujeres más el número de promotores hombres y se dividió entre el total de personal.
studentPromotersRatio	Número de alumnos por promotor.	Se dividió el número de alumnos entre el número total de promotores.

Únicamente escuelas comunitarias.

Nombre de variable	**Descripción**	**¿Cómo se construyó?**
comunitario	Variable indicadora de educación comunitaria rural primaria.	Se asignó el valor 1 en toda la tabla.
cursosComunitarios	Variable indicadora del servicio de curso comentarios.	Se renombró la variable V1.
paepi	Variable indicadora del servicio de Proyecto de Atención Educativa a la Población Indígena (PAEPI)	Se renombró la variable V2.
paepiam	Variable indicadora del servicio de Proyecto de Atención Educativa a la Población Infantil Agrícola Migrante (PAEPIAM)	Se renombró la variable V3.
aulasCompartidas	Variable indicadora del proyecto de aulas compartidas.	Se renombró la variable V410.

APÉNDICE C

INGENIERÍA DE CARACTERÍSTICAS: VARIABLES CEMABE

Utilizando el CEMABE se crearon las siguientes variables

Nombre de variable	Descripción o pregunta	¿Cómo se construyó?	Supuestos y comentarios
propertyType	Tipo de inmueble Opciones: Construcción hecha para fines educativos (incluye CAM) Construcción adaptada para fines educativos Construcción provisional (materiales ligeros y precarios) Escuela móvil (vagón, camión, circo, etc.) Sin construcción (al aire libre) Instalaciones de apoyo a la educación especial (USAER, CRIE, CAPEP, UOP, etc.) Biblioteca Instalaciones administrativas de la SEP o de apoyo a la labor educativa (supervisión de zona, jefaturas de sector, centro de maestros, etc.)	Renombrando la variable p3	Se puede utilizar para crear variables indicadoras de cada categoría
numberCT	Número de centros de trabajo en el inmueble. En un inmueble puede haber preescolar, primaria, secundaria y/o preparatoria.	Renombrando la variable p5	
outsideFacilities	11. ¿Cuántos anexos escolares tiene este inmueble en otro domicilio? (canchas deportivas, gimnasio, auditorio, oficinas, etc.)	Renombrando la variable p11	
fenceMaterial	¿De qué material es la mayor parte de la barda o cerco perimetral?	Se asigno un valor dependiendo del material.	Valores asignados por material: Tabique, ladrillo, block, piedra, cantera, cemento o concreto 0.9 Reja metálica 0.5 Malla ciclónica 0.4 Madera 0.3
wallMaterial	¿De qué material es la mayor parte de las paredes o muros del inmueble?	Se asigno un valor dependiendo del material.	Valores asignados por material: Tabique, ladrillo, block, piedra, cantera, cemento o concreto 0.9 Adobe 0.5

			Madera 0.4 Embarro o bajareque, carrizo, bambú o palma 0.3 Lámina metálica, de asbesto o cartón 0.2 Material de desecho 0.1
roofMaterial	¿De qué material es la mayor parte del techo del inmueble?	Se asigno un valor dependiendo del material.	Valores asignados por material: Losa de concreto o viguetas con bovedilla 0.9 Teja 0.5 Terrado con viguería 0.4 Madera, tejamanil, palma o paja 0.3 Lámina metálica, de asbesto o cartón 0.2 Material de desecho 0.1
floorMaterial	¿De qué material es la mayor parte del piso del inmueble?	Se asigno un valor dependiendo del material.	Valores asignados por material: Madera, mosaico u otro recubrimiento 0.9 Cemento o firme 0.5 Tierra o materiales removibles 0.2
propertyMaterials	Material del inmueble	Suma de los valores asignados del material del piso, de las paredes, del techo y de la barda	
noWaterSupply	Variable indicadora de si el inmueble tiene acceso a agua o no	Tiene el valor 1 si la escuela no tiene fuente de	

		abastecimiento de agua. Se utilizó la variable p17a	
electricity	Variable indicadora de si el inmueble tiene energía eléctrica	Tiene el valor 1 si la escuela no tiene fuente de abastecimiento de agua. Se utilizó la variable p18a	

REFERENCIAS

[1] A. A. Aquino, G. Molero-Castillo y R. Rojano, "Hacia un nuevo proceso de minería de datos centrado en el usuario", *Pistas Educativas*, vol. 36, n.º 114, 2018. dirección: http://itcelaya.edu.mx/ojs/index.php/pistas/article/view/303.

[2] SAS. (2018). Data Mining and SEMMA, dirección: http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm (visitado 10-04-2019).

[3] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth y col., "Knowledge Discovery and Data Mining: Towards a Unifying Framework.", en *KDD*, vol. 96, 1996, págs. 82-88.

[4] E. León. (2018). Metodologías aplicadas al proceso de Minería de Datos, dirección: http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf (visitado 05-02-2019).

[5] H. Palacios, R. Toledo, G. Hernandez y A. Navarro, "A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change", *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, págs. 598-604, jun. de 2017. DOI: 10.25046/aj020376.

[6] R. Wirth y J. Hipp, "CRISP-DM: Towards a standard process model for data mining", en *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, 2000, págs. 29-39.

[7] A. Lleras-Muney, "The relationship between education and adult mortality in the United States", *The Review of Economic Studies*, vol. 72, n.º 1, págs. 189-221, 2005.

[8] R. J. Barro, "Democracy and growth", *Journal of economic growth*, vol. 1, n.º 1, págs. 1-27, 1996.

[9] E. A. Hanushek, D. T. Jamison, E. A. Jamison y L. Woessmann, "Education and economic growth: It's not just going to school, but learning something while there that matters", *Education next*, vol. 8, n.º 2, págs. 62-71, 2008.

[10] J. D. Gregorio y J.-W. Lee, "Education and Income Inequality: New Evidence from Cross-country Data", *Review of income and wealth*, vol. 48, n.º 3, págs. 395-416, 2002.

[11] "Determinantes del logro escolar en México . Primeros resultados utilizando la prueba ENLACE media superior, author=Hoyos, Rafael E de and Espino, Juan Manuel and García, Vicente, journal=El trimestre económico, volume=79, number=316, pages=783–811, year=2012",

[12] R. A. Española. (2005). escolaridad, dirección: `http://lema.rae.es/dpd/srv/search?key=escolaridad` (visitado 05-02-2019).

[13] P. Informe, "Aprender para el Mundo de Mañana", *Madrid. Santillana*, 2003.

[14] A. Márquez Jiménez, "A 15 años de PISA: resultados y polémicas", *Perfiles educativos*, vol. 39, n.º 156, págs. 3-15, 2017.

[15] A. Ortega, "Maestros, plazas, el adiós del INEE y otras claves de la nueva reforma educativa", *Expansión Pólitica*, mayo de 2019. dirección: `https://politica.expansion.mx/mexico/2019/04/25/maestros-plazas-el-adios-del-inee-y-otras-claves-de-la-nueva-reforma-educativa`.

[16] R. M. Torres y E. Tenti, "Políticas educativas y equidad en México: La experiencia de la Educación Comunitaria, la Telesecundaria y los Programas Compensatorios", Secretaría de Educación Pública, Dirección General de Relaciones Internacionales, inf. téc., 2000.

[17] J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms", *Auditing: A Journal of Practice & Theory*, vol. 30, n.º 2, págs. 19-50, 2011.

[18] B. K. Lee, J. Lessler y E. A. Stuart, "Improving propensity score weighting using machine learning", *Statistics in medicine*, vol. 29, n.º 3, págs. 337-346, 2010.

[19] N.-B. Sara, R. Halland, C. Igel y S. Alstrup, "High-school dropout prediction using machine learning: A danish large-scale study", en *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, 2015, págs. 319-24.

[20] S. de Educación. (2018). Misión, visión y objetivo, dirección: `https://seduc.edomex.gob.mx/mision_vision_objetivo` (visitado 11-07-2019).

[21] SEGOB. (2015). ACUERDO número 18/12/15 por el que se emiten las Reglas de Operación del Programa Escuelas de Tiempo Completo para el ejercicio fiscal

2016., dirección: http://dof.gob.mx/nota_detalle.php?codigo=5421435&fecha=27/12/2015.

[22] S. de Educación y Cultura Subsecretaría de Planeación Educativa Dirección de Evaluación y Estadística. (2010). FORMATO 911 (Preescolar, Primaria y Secundaria), dirección: http://web.seducoahuila.gob.mx/sidecc/formatos/Formato911-2.pdf (visitado 18-05-2019).

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot y E. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[24] M. B. J. Silveyra De La Garza Marcela Lucia; Yanez Pagans. (2018). ¿Qué impacto tiene el Programa Escuelas de Tiempo Completo en los Estudiantes de Educación Básica? : Evaluación del Programa en México 2007-2016 (Spanish).

[25] Wikipedia. (2019). Criterio de información de Akaike, dirección: https://es.wikipedia.org/wiki/Criterio_de_informaci\%C3\%B3n_de_Akaike (visitado 09-06-2019).

[26] C. Masci, G. Johnes y T. Agasisti, "Student and school performance across countries: A machine learning approach", *European Journal of Operational Research*, vol. 269, n.º 3, págs. 1072-1085, 2018.

[27] B.-H. Kim, E. Vizitei y V. Ganapathi, "GritNet: Student performance prediction with deep learning", *arXiv preprint arXiv:1804.07405*, 2018.

[28] M. Solutions. (2017). Advantages and Disadvantages of Python Programming Language, dirección: https://medium.com/@mindfiresolutions.usa/advantages-and-disadvantages-of-python-programming-language-fd0b394f2121 (visitado 15-07-2019).

[29] P. N. de Transparencia. (2019). Solicitudes, dirección: https://www.plataformadetransparen
org.mx/web/guest/inicio (visitado 20-04-2019).

[30] S. de Educación Pública. (2014). Censo de escuelas, maestros y alumnos de educación básica y especial, dirección: https://datos.gob.mx/busca/dataset/censo-de-escuelas-maestros-y-alumnos-de-educacion-basica-y-especial (visitado 05-02-2019).

[31] M. y.A.d.E.B.y.E. C. Censo de Escuelas, *Tutorial para el manejo de las tablas de datos*. INEGI, 2014.

[32] M. tu escuela. (2013). Nota metodológica para educación básica., dirección: http://www.mejoratuescuela.org/metodologia (visitado 23-07-2019).

[33] SEP. (2016). GLOSARIO DE TÉRMINOS: Educación Básica, dirección: http://planeacion.sec.gob.mx/upeo/GlosariosInicio20162017/BASICA2016.pdf (visitado 18-07-2019).

[34] ——, (2012). Glosario de Términos Utilizados, dirección: http://cumplimientopef.sep.gob.mx/glosario_de_terminos/ (visitado 18-07-2019).

[35] E. Backhoff y S. Contreras Roldán, "Corrupción de la medida" e inflación de los resultados de ENLACE", *Revista mexicana de investigación educativa*, vol. 19, n.º 63, págs. 1267-1283, 2014.

[36] N. Martínez. (2019). Privadas, mejores que públicas: ENLACE, dirección: https://archivo.eluniversal.com.mx/nacion/171721.html (visitado 23-07-2019).

[37] ENLACE. (2014). Procedimiento general, dirección: http://www.enlace.sep.gob.mx/ba/aplicacion/procedimiento_general/ (visitado 16-08-2019).

[38] Statistica. (2019). Outliers and Extremes, dirección: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Graphs/Graph/CreatingGraphs/Dialogs/2DGraphs/Notes/OutliersandExtremes (visitado 13-08-2019).

[39] M. Ved. (2018). Feature Selection and Feature Extraction in Machine Learning: An Overview, dirección: https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96 (visitado 21-04-2019).

[40] A. Shetye. (2019). Feature Selection with sklearn and Pandas, dirección: https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b (visitado 15-08-2019).

[41] I. R. White, P. Royston y A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice", *Statistics in medicine*, vol. 30, n.º 4, págs. 377-399, 2011.

[42] D. J. Stekhoven y P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data", *Bioinformatics*, vol. 28, n.º 1, págs. 112-118, oct. de 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr597. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/28/1/112/583703/btr597.pdf. dirección: https://doi.org/10.1093/bioinformatics/btr597.

[43] P. M. Carneiro, J. Das y H. Reis, "The value of private schools: Evidence from Pakistan", 2016.

[44] N. Bau, "School competition and product differentiation", Working Paper. Toronto, ON, inf. téc., 2015.

[45] E. Backhoff, A. Bouzas, C. Contreras, E. Hernández y M. García, "Factores escolares y aprendizaje en México. El caso de la educación básica", *México: INEE. Recuperado de: http://www. inee. edu. mx/images/Samana Vergara-Lope Tristán y Felipe J. Hevia de la Jara*, vol. 63, 2007.

[46] L. F. DiLalla, J. L. Marcus y M. V. Wright-Phillips, "Longitudinal effects of preschool behavioral styles on early adolescent school performance", *Journal of School Psychology*, vol. 42, n.º 5, págs. 385-401, 2004.

[47] P. Sharma. (2018). The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes), dirección: `https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/` (visitado 21-04-2019).