INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

DISEÑO E IMPLEMENTACIÓN DE UN PRODUCTO DE DATOS PARA LA TOMA DE DECISIONES EN PROGRAMAS SOCIALES ORIENTADOS A ESCUELAS PRIMARIAS GENERALES EN 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 producto de datos para la toma de decisiones en programas sociales para escuelas primarias en México**", 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

He aquí mi secreto, que no puede ser más simple: solo con el corazón se puede ver bien; lo esencial es invisible para los ojos.

Antoine de Saint-Exupéry

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ACRÓNIMOS

CAM Centro de Atención Múltiple. 37

CAS Sistemas Complejos Adaptativos. 1

CCT Clave de Centro de Trabajo. 14, 22, 39

CEMABE Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial. 10, 11, 20, 22, 24, 25, 29, 37, 38, 46

CNRMMCE Centro Nacional para la Revalorización del Magisterio y la Mejora Continua de la Educación. 6

CONAFE Consejo Nacional de Fomento Educativo. 25

CRISP-DM Cross-Industry Standard Frocess for Data Mining. 1, 2

csv Comma-Separated Values. 23

ENLACE Evaluación Nacional de Logro Académico en Centros Escolares. 9, 10, 13, 16, 20, 21

EXCALE Exámenes de la Calidad y el Logro Educativo. 9, 20

F911 Formato Estadístico 911. 10, 11, 20, 24, 38, 46

INEE Instituto Nacional para la Evaluación de la Educación. 6, 20

INEGI Instituto Nacional de Estadística y Geografía. 11

KDD Knowledge Discovery in Databases Framework. 1, 2

OCDE Organización para la Cooperación y el Desarrollo Económico. 6

PISA Informe del Programa Internacional para la Evaluación de Estudiantes. 6, 9, 10, 20

Planea Plan Nacional para la Evaluación de los Aprendizajes. 10, 20

ROC Característica Operativa del Receptor. 60

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

SEP Secretaría de Educación Pública. 11, 16

CAPÍTULO 1

INTRODUCCIÓN

La ciencia de datos es un campo en la intersección de la computación [1] y la estadística. Este campo permite el estudio fenomenológico de Sistemas Complejos Adaptativos (CAS), con el propósito de construir productos de datos que ayuden a la toma de decisiones y acciones sobre el sistema [2].

Este trabajo se identifica como un proyecto de ciencia de datos. Presenta el estudio de la educación primaria en México y la construcción de un producto de datos para la toma de decisiones de programas sociales para el sector educación.

Para lograr dicho fin, se exploraron tres diferentes metodologías para guiar el proyecto.

1.1 Posibles metodologías

Existen varías metodología alternativas para realizar un producto 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.

1.1.1 Sample, Explore, Modify, Model, and Assess

En primer lugar, la metodología utilizada por la compañia SAS para análisis de datos se llama "SEMMA". La característica principal de la metodología es que los diferentes pasos se manejan 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 [3]. Esta metodología se basa en la parte técnica del proyecto como la aplicación de técnicas estadísticas y visualización de datos. Sin embargo, no considera los objetivos del negocio ni el contexto del problema.

1.1.2 Knowledge Discovery in Databases Framework

En segundo lugar, KDD fue 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 [4].

1.1.3 Cross-Industry Standard Frocess 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 [5].

1.2 Metodología seleccionada

De las posibles metodologías se eligió la metodología CRISP-DM para el proyecto ya que no hay propietario, es independiente de la aplicación o la industria y es neutral con respecto qué herramientas utilizar. 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 [6].

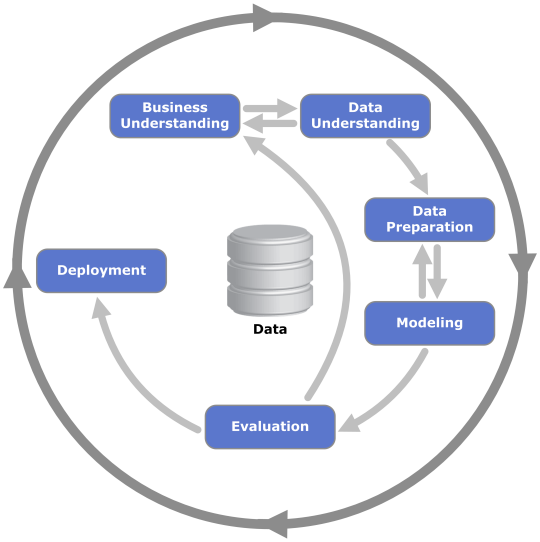


Figura 1.1: Fases del modelo CRISP-DM

1.3 Organización del documento

Como se ve en la figura 1.1, la metodología propuesta en el año 2000 [7], propone las siguientes seis fases:

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

[1]Las traducciones de la métodología están basadas en el Manual de CRISP-DM de IBM [8]

El documento está organizado en siete capítulos correspondientes a las seis fases de la metodología más este capítulo introductorio que pretende justificar la estructura y el uso de la metodología.

CAPÍTULO 2

COMPRENSIÓN DEL SECTOR EDUCATIVO

Este capítulo presenta el panorama general y explora las necesidades del "negocio". Para fines del proyecto, 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 a través de diversos proyectos y programas.

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 [9], mayor democracia, participación ciudadana [10] y crecimiento económico [11] de la mano de una mayor equidad en la distribución de ingresos [12] [13].

La escolaridad se refiere al periodo de asistencia a un centro escolar [14]. Sin embargo, los beneficios no están relacionados con el número de años en la escuela, sino con el aprendizaje dentro de ella [11]. 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 [15].

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 [16]. 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 [16] , 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 mejorar 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

[17]. Es decir, el gobierno está interesado en implementar programas sociales con el fin de mejorar el desempeño escolar. Asimismo, cabe destacar que la meta no solo es gubernamental. Existen varías organizaciones de la sociedad civil con el objetivo de elevar la calidad de la educación en México [1].

Antecedentes

Históricamente han existido y existen varios programas gubernamentales y de sociedades civiles orientados a mejorar el desempeño de las escuelas [2].

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 [19].

La definición de prioridad de los programas sociales, en qué y dónde se invierte primero, puede ser dictada por el gobierno federal, por los propios agentes del sistema escolar o por la organización civil. En algunos casos, se realizan entrevistas con directivos y docentes e históricamente se ha dado prioridad a aspectos que tienen que

[1]Entre estas organizaciones se encuentran: Bécalos de Fundación Televisa, Sembrando Arte y Tecnología para la educación, Junior Achievement Worldwide y Proeducación [18]

[2]Entre los programas que existen o han existido, se encuentran: El Programa Escuelas de Tiempo Completo, Programa Desayunos Escolares, Programa de Acciones Compensatorias para Abatir el Rezago Educativo en la Educación Inicial y Básica, 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.

ver con la infraestructura física de los establecimientos escolares y no con la calidad de docentes o servicios de la escuela [19]. En otros casos, las escuelas interesadas hacen una solicitud antes de el proceso de entrevista. Después de las entrevistas, cada proyecto tiene su propio proceso de selección con criterios que no siempre son claros, justos ni transparentes.

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 qué escuelas asignarle recursos y qué criterio de asignación usar.

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. Asimismo, cada programa va dirigido a un tipo de escuela o a una región geográfica especifica.

Tomando esto en cuenta, el objetivo del sector se entiende como tomar decisiones sobre la asignación de programas sociales de forma informada, transparente y precisa, haciendo distinción entre regiones geográficas y tipos de escuelas.

2.1.3 Criterios de éxito del sector

El criterio de éxito es correctamente determinar cuáles escuelas están en riesgo de tener rendimiento académico decreciente y identificar cambios o elementos de la escuela que estén relacionados con una caída en el desempeño general.

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.

Los programas educativos actuales tienen diferentes enfoques y están focalizados para diferentes poblaciones. Tienen en común el objetivo de mejorar la calidad educativa. La calidad educativa es un problema multidimensional que se puede medir y evaluar de distintas maneras cuantitativas y cualitativas.

En la sección 2.1.1 se mencionó que una forma de medir el aprendizaje y calidad educativa son las habilidades cognitivas. Finalmente se argumentó que las pruebas estandarizadas sirven para medir estas habilidades cognitivas. Con esto en mente, una forma de estimar cuantitativamente la calidad educativa puede ser examinando los resultados en pruebas estandarizadas.

Tomando en cuenta el objetivo del sector de tomar decisiones informadas, resulta interesante conocer no solo el desempeño académico si no también las características de las escuelas, alumnos y profesores.

2.2.1 Inventario de recursos

Valorando la situación, los recursos que se necesitan son resultados de pruebas estandarizadas y características de la infraestructura, los profesores y los alumnos de las escuelas primarias. A continuación, se mencionarán los recursos de información disponibles.

El primero de estos recursos son los resultados de pruebas estandarizadas. México participa en la prueba estandarizada internacional PISA e internamente aplica y 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 aplicadas 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. La última en la lista es PISA, una prueba internacional con el defecto de que los datos no están disponibles a nivel escuela y evalúa a un menor número de escuelas que las pruebas nacionales. Además de PISA, Planea es la única prueba que sigue vigente. Esto quiere decir que PLANEA tiene resultados más recientes, sin embargo su alcance es mejor al de ENLACE.

ENLACE es la prueba con mayor alcance 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.

Además de los resultados de las pruebas mencionadas anteriormente, el segundo recurso de información disponible son las características de las escuelas, los alumnos, directivos y personal docente. Para esto se tienen dos fuentes de información principales que son el Formato Estadístico 911 (F911) y el Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (CEMABE).

La primera fuente de información, el Formato Estadístico 911 (F911), es un cuestionario llenado, en teoría, por todos los centros educativos del país al inicio y al final de cada ciclo escolar. El formato incluye información sobre 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 [21].

La segunda fuente de información, el Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (CEMABE), fue un esfuerzo del Instituto Nacional de Estadística y Geografía (INEGI) y la SEP para recopilar información sobre el inmueble físico de los centros de trabajo.

En resumen, el inventario final de datos consta de los siguientes elementos:

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

Los datos se complementan con los siguientes recursos:

- Asesores y expertos en el tema de educación [3] y de minería de datos [4].
- Acceso a un servidor remoto con 512 GB de memoria.
- Conocimiento y experiencia previa con Python
- El paquete de modelos algorítmicos y estadísticos Scikit-learn [22].

2.2.2 Requerimientos funcionales, supuestos y restricciones

Requerimientos funcionales

Los requerimientos funcionales del producto de datos son los siguientes:

[3] Dr. Enrique Seira
[4] M.S. Juan Salvador Mármol

- **Cobertura**: el producto deberá tener información de la mayor cantidad de escuelas posible.
- **Datos abiertos**: el producto deberá ser construido utilizando en su mayoría datos abiertos.
- **Integración de datos**: el producto deberá integrar datos de diferentes fuentes, años y formatos.
- **Limpieza de datos**: el producto deberá presentar y utilizar datos limpios. En especifico, se deberán manejar los valores faltantes, diferentes codificaciones y los errores tipográficos.
- **Ingeniería de características**: el producto deberá identificar las variables más importantes y tener nuevas variables significativas que sean modificaciones de las originales.
- **Interpretabilidad**: El producto deberá ser entendible. Deberá ser posible interpretar los resultados.
- **Transparencia**: El proceso de construcción del producto deberá estar bien documentado, deberá ser replicable y transparente en todos sus pasos.
- **Flexibilidad**: El producto deberá ser flexible y adaptarse a requerimientos específicos por usuario. Por ejemplo: Visualizar resultados para una entidad federativa en particular.
- **Alcance**: el producto deberá tener un gran alcance. Es decir, deberá poder ser utilizado en toda la República Mexicana.
- **Comunicación de resultados**: el producto de datos deberá estar disponible en línea y se deberán hacer consultas al servicio web.

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 está basado en el impacto positivo significativo del programa Escuelas de Tiempo Completo [23].

Restricciones

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

En primer lugar, el proyecto está sujeto a qué datos están disponibles. Es decir, el número de escuelas de las cuales no se tiene información es una restricción.

En segundo lugar, el proyecto no tiene presupuesto. Por lo tanto, las herramientas computacionales están restringidas por la memoria de una computadora de 24 GB. En caso de que alguna base exceda la capacidad de la computadora, se utilizarán otras herramientas disponibles en un servidor remoto con aplicaciones limitadas o en la nube.

En tercer lugar, la calidad del proyecto es proporcional a la calidad de los datos. Los datos de baja calidad, capturados a través del tiempo por diferentes personas y organismos, restringen el desempeño de los modelos y del proyecto.

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 obtener 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
Tener un la variable objetivo sesgada, no confiable o informativa	2	4	Documentar la creencia de que no es confiable y explorar por qué

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".

- **Sostenimiento**: Fuente que proporciona los recursos financieros para el funcionamiento del centro de trabajo. Las principales son estatal, federal, CONAFE y privada.

- **Grado de marginación**: Es un indicador multidimensional que mide la intensidad de las privaciones padecidas por la población a través de 9 formas de exclusión agrupadas en 4 dimensiones: educación, vivienda, distribución de la población e ingresos monetarios.

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

- **Modelos de datos**: Asume un modelo estocástico y estima parámetros de los datos. Ejemplos: Regresiones lineales y regresiones logísticas [24].
- **Modelos algorítmicos**: No asume ningún modelo y se concentra en hacer predicciones. Ejemplos: Redes neuronales y árboles de decisión.
- **Limpieza de 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. Asimismo, un costo a considerar a futuro son los recursos de almacenamiento y procesamiento de datos y la renta mensual si se desea mantener una aplicación en la nube.

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 sector educativo y los requerimientos funcionales mencionados anteriormente, el objetivo del proyecto de minería de datos es responder la siguiente pregunta: "¿Cuáles escuelas primarias están en riesgo de bajar su rendimiento académico?"

Se identifican los siguientes dos sub-objetivos:

1. Construir un modelo para predecir rendimiento escolar decreciente.
2. Crear una aplicación web para hacer disponible el modelo y sus resultados.

En específico, suponiendo que las pruebas estandarizadas miden el desempeño académico de una escuela, el primer sub-objetivo se traduce en predecir los cambios negativos entre diferentes resultados de "alguna" prueba estandarizada.

Dado que uno de los requerimientos es que el proyecto tenga una gran cobertura, conviene utilizar los resultados de ENLACE ya que, como en la tabla 2.1, es la prueba que se aplicó en el mayor número de escuelas.

El problema de predicción se puede abordar como un análisis de clasificación o de regresión. La SEP, encargada de aplicar ENLACE, clasificó los resultados individuales en los siguientes cuatro niveles: insuficiente, elemental, bueno y excelente. Sin

embargo, nos interesan los cambios de resultados a nivel escuela. En situaciones del mundo real, un quinto de desviación estándar (0.2) se considera un efecto significativo y grande [25]. Por lo tanto, podemos clasificar aquellas escuelas cuyo desempeño bajó 0.2 desviaciones estándar en un determinado periodo como escuelas con "rendimiento decreciente".

Como resultado, el primer sub-objetivo se centrará en construir un modelo de clasificación de escuelas con cambios en los resultados promedio de la prueba ENLACE negativos.

2.3.2 Criterios de rendimiento del proyecto de minería de datos

Por un lado, el proyecto será exitoso si crea nuevo conocimiento y el producto final genera un impacto. Es decir, si se toman decisiones de programas sociales o de políticas públicas utilizando la información del proyecto.

Por otro lado, también se considerará exitoso si se determina que los datos no pueden responder la pregunta planteada.

2.4 Soluciones relacionadas

2.4.1 Modelos de datos

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 [13].

2.4.2 Modelos algorítmicos

Métodos con árboles

El artículo "Student and school performance across countries: A machine learning approach" [26] presenta un análisis de determinantes de resultados de la prueba PISA. La prueba PISA, como se menciono 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 exploran 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.

Métodos con redes neuronales

El artículo "GritNet: Student Performance Prediction with Deep Learning" [27] plantea el problema de predicción de desempeño de un alumno como un análisis de eventos secuenciales y propone una red neuronal (GridNet) construida sobre una memoria bidireccional de corto plazo prolongado (Bidirectional Long Short-Term Memory). 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.

2.4.3 Valoración de herramientas y técnicas

Se utilizará Stata y R para la exploración de datos, el pre-procesamiento de datos se realizará en Stata y el modelado y despliegue se implementará en Python. 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 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.

CAPÍTULO 3

COMPRENSIÓN DE LOS DATOS

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

3.1 Recopilación de datos iniciales

En la sección "Valoración de la situación" del capítulo 1 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 (CEMABE).

La tabla 3.1 muestra los conjuntos de datos iniciales, el formato y el método que se utilizó para obtenerlos.

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 a comparasión de las otras pruebas, tuvo mayor alcance.

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

Tabla 3.1: Conjunto de datos obtenidos

Nombre	Formato	Método
Censo escuelas CEMABE	CSV	Descarga electrónica
2013		cemabe.inegi.org.mx/
F911	DBF	Solicitud email
2006-2013		plataformadetransparencia.org.mx
Resultados ENLACE	Varios	Descarga electrónica
Por escuela		enlace.sep.gob.mx/
2006-2007 y 2009-2013		
Resultados ENLACE	Varios	Solicitud CIE
Por alumno		Centro de Investigación Económica
2006-2013		

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

A nivel escuela, los resultados históricos 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.

[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 CEMABE descargados desde el portal de Datos Abiertos del Gobierno de México [30].

3.2 Descripción de los datos

Todos los datos recopilados a nivel escuela se pueden identificar únicamente 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 identificado 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)

Por un lado, una variable relevante que se encuentra en las bases a nivel escuela y no a nivel alumno es el grado de marginación de la escuela. Por el otro lado, las bases a nivel escuela contienen el porcentaje de copia mientras que a nivel alumno se sabe si "copió" o no. Esto se explicará con más detalle en secciones futuras.

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 [21]. 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**	**Indígena**
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

Como muestra la tabla 3.4 el número de escuelas comunitarias e indígenas a comparación con las escuelas generales es muy pequeño. Por esa razón el análisis se centrará en escuelas generales.

3.2.3 Descripción CEMABE

El Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial (CEMABE) 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].

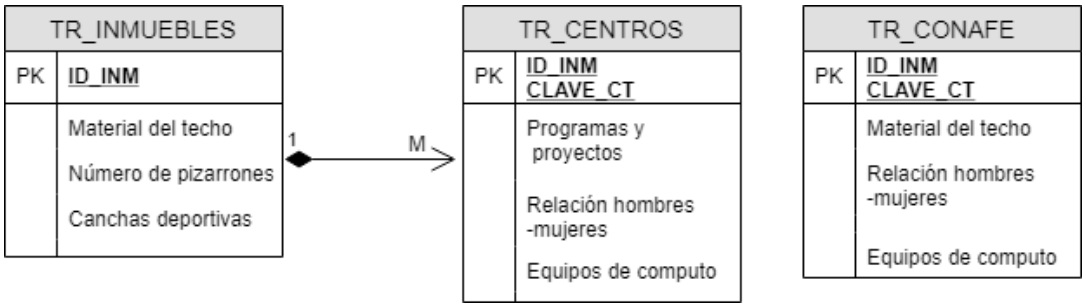


Figura 3.1: Diagrama entidad relación de tablas del CEMABE

Las datos del CEMABE se recopilaron en formato csv. El conjunto de datos consta 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 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 se 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 CEMABE.

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 numérico en la ma-

yoría de los casosy que los valores faltantes están representados por los números “9”,“99”,“999”,“9999” o “99999”.

3.3 Exploración de datos

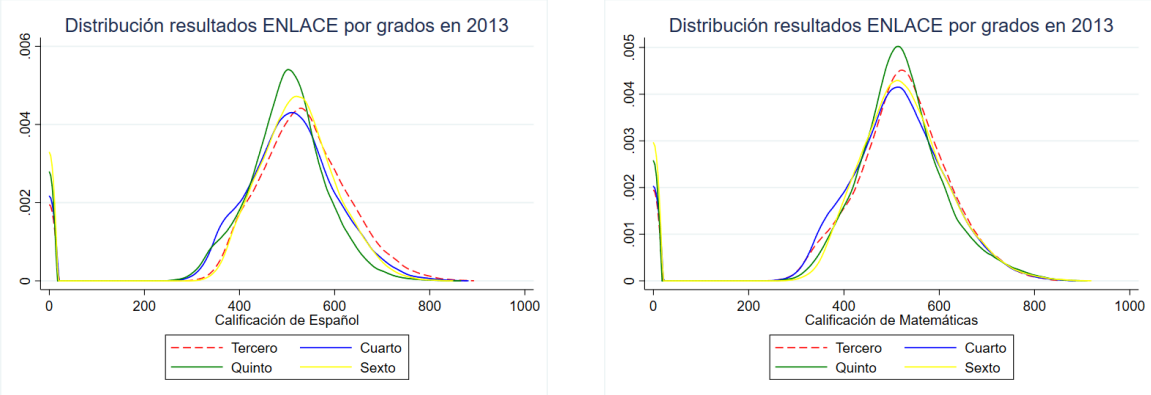
A continuación se muestran resultados significativos de la exploración de datos.

3.3.1 Exploración univariada

La variable sobre la cual nos interesa predecir los cambios es la calificación ENLACE. Los datos se obtuvieron a nivel escuela y a nivel alumno.

ENLACE nivel escuela

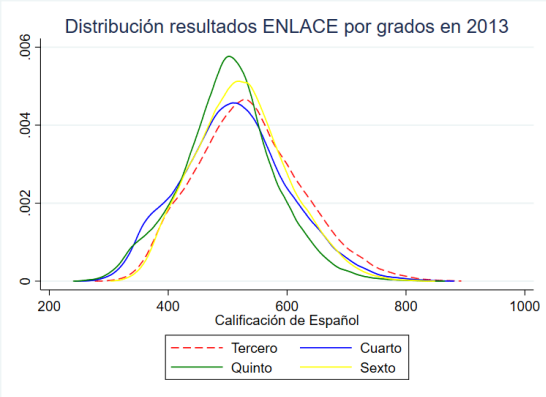
La figura 3.2a y 3.2b muestran la distribución de calificaciones de ENLACE por grado del 2013. Las figuras muestran un “pico” en el cero. En algunos casos, las calificación es cero porque en ese año un grado no presentó la prueba.



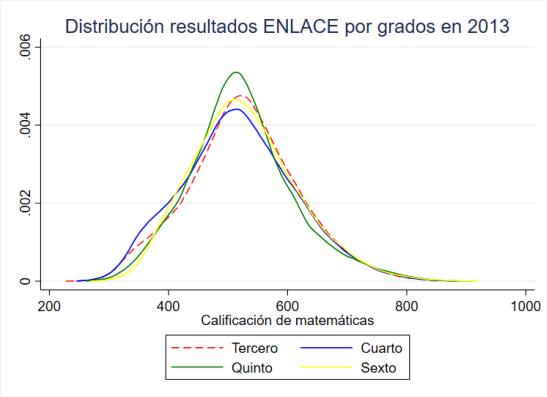
(a) Español (b) Matemáticas

Figura 3.2: Calificaciones de primaria por grado escolar

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



(a) Español



(b) Matemáticas

Figura 3.3: Calificaciones de primaria por grado escolar corregidas.

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

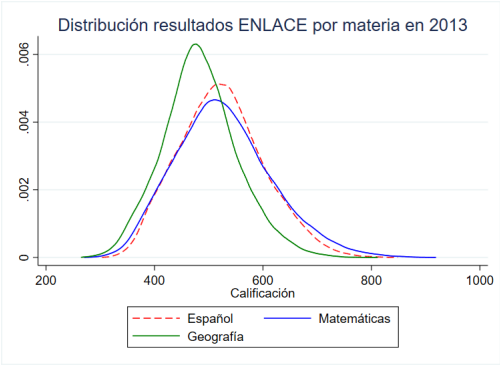
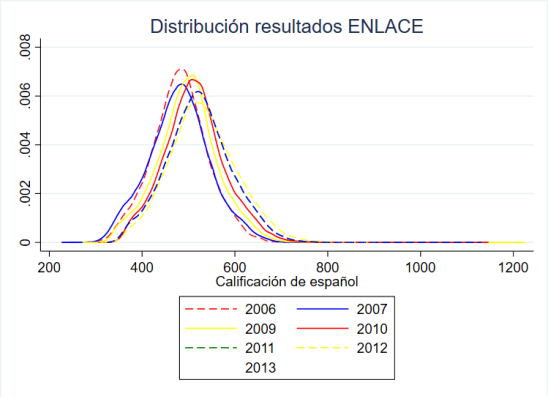


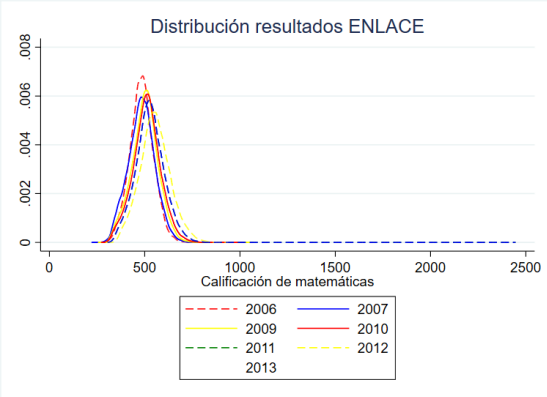
Figura 3.4: Distribución resultados por materia en 2013

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

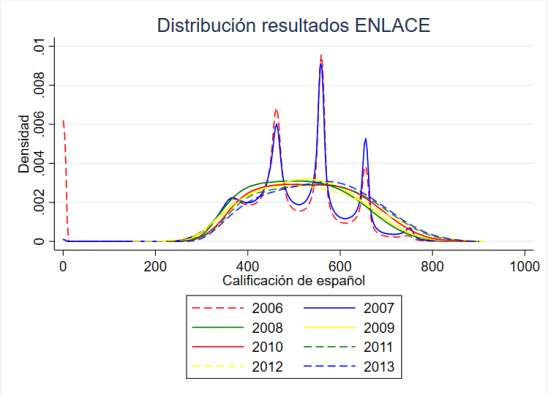


(b) Matemáticas

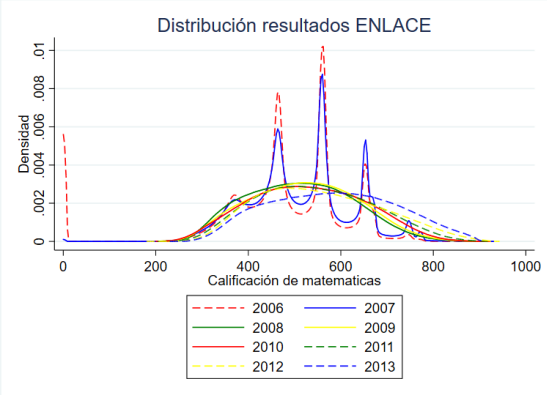
Figura 3.5: Comparación de distribución de resultados desde 2006 hasta 2013 (sin 2008)

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

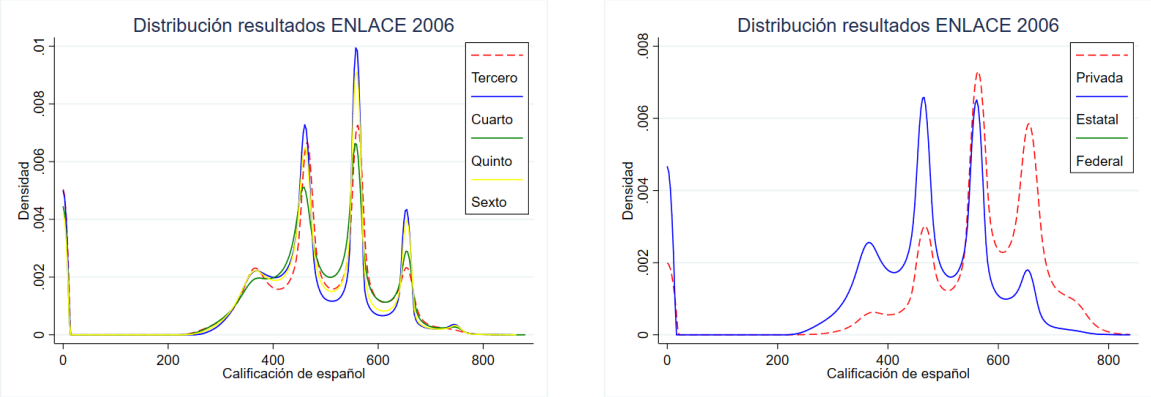


(b) Matemáticas

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. Es posible que los datos del 2006 y del 2007 estén alterados y por eso presenten tal comportamiento. Evidencia que sustenta esto es que a través del portal web es posible obtener los resultados por folio de los alumnos para todos los años excepto 2006 y 2007.



(a) Resultados por grado

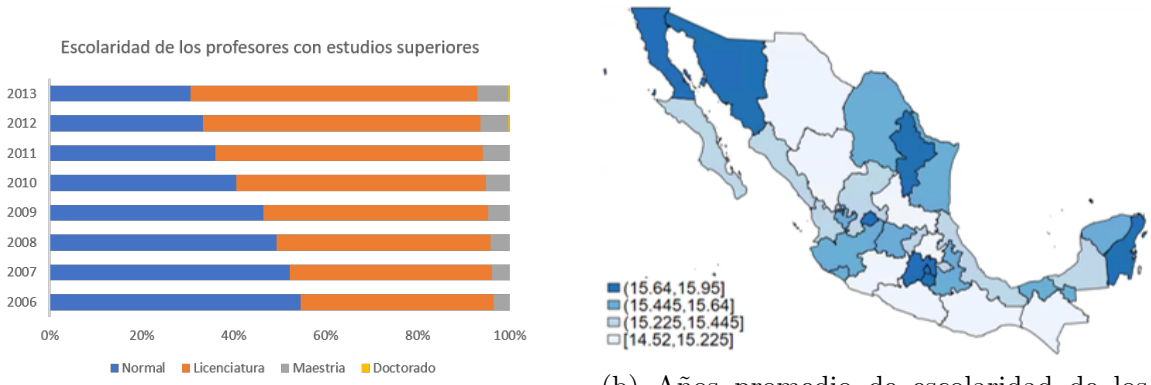
(b) Resultados por sostenimiento

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, 3.8b y 3.9 muestran resultados notables de la exploración del formato estadístico 911 y las figuras **??** y 3.11b los resultados del CEMABE.

En cuanto a la educación de los profesores, como se ve en la figura 3.8a, es interesante que a través de los años hay un mayor porcentaje de profesores con licenciatura y menor porcentaje de profesores con un título de la normal. Asimismo, como se ve en la figura 3.8b, cabe destacar que a pesar de que Oaxaca y Chiapas son de los estados con menor años escolaridad de los profesores, la diferencia no es tan grande con Nuevo



(a) Personal docente titulado

(b) Años promedio de escolaridad de los profesores por estado

Figura 3.8: Visualizaciones interesantes del F911

León o la Ciudad de México. Los profesores de Chiapas tienen en promedio 14.52 años de escolaridad mientras que en la Ciudad de México el promedio es 15.95. La diferencia es de menos de dos años. Sin embargo, las diferencias entre el desempeño de los alumnos son mucho más significativas. Esto puede reforzar la creencia que los años de educación no equivalen a la calidad de la educación.

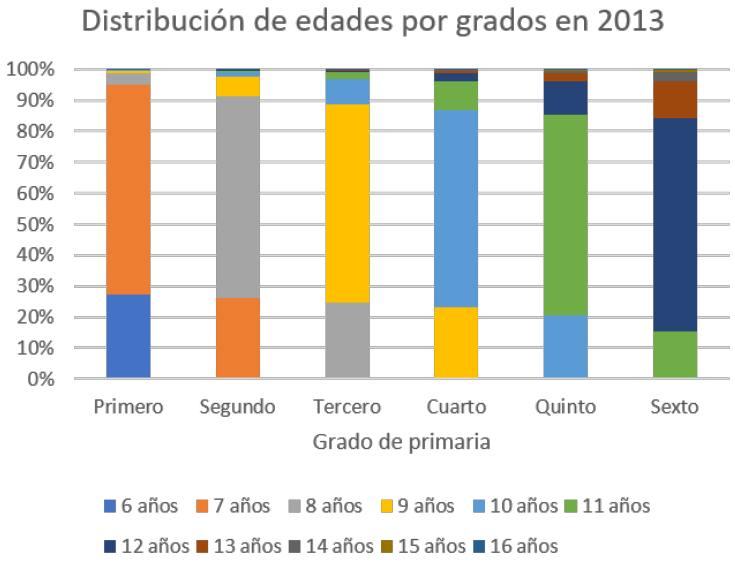


Figura 3.9: Edades de los alumnos por grado

La distribución de edades vista en la figura 3.9 depende del día en el que la escuela recopiló las edades de los alumnos. Por ejemplo, para primero de primaria, las edades

registradas en el formato de inicio de cursos cambian durante el ciclo escolar, es muy posible que los alumnos de seis años, cumplan siete durante el año escolar. Lo interesante es que en sexto de primaria los alumnos "dos" años menores tienen mayor porcentaje que en otros años. Estos alumnos sí podrían ser regazados o repetidores.

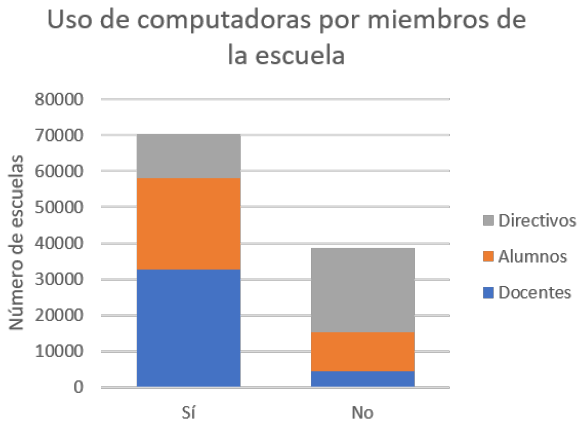
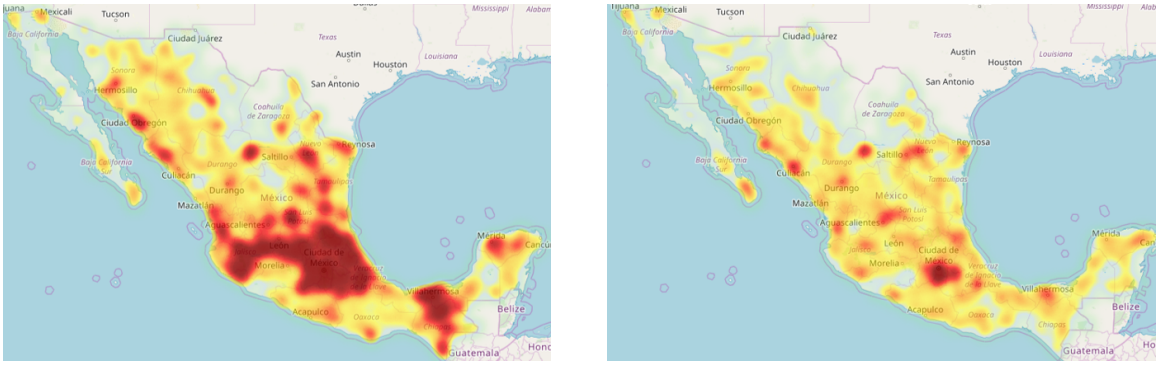


Figura 3.10: Uso de computadoras por miembros de la escuela

La figura 3.10 muestra el número de escuelas que usan computadoras desglosado por el usuario. Es interesante que los directivos no usen la computadora y que los docentes, en comparación, en su mayoría si la usen de apoyo.



(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.

Las figuras 3.11a y **??** muestran la concentración de escuelas con el programa de desayunos escolares y de tiempo completo. Cabe destacar que en centro muestra más

concentración porque existe un mayor número y densidad de escuelas en esa zona.

3.3.2 Exploración bivariada

La variable de "cambio negativo en el rendimiento escolar" se construirá utilizando la calificación de ENLACE.

Existen calificaciones 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.

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.

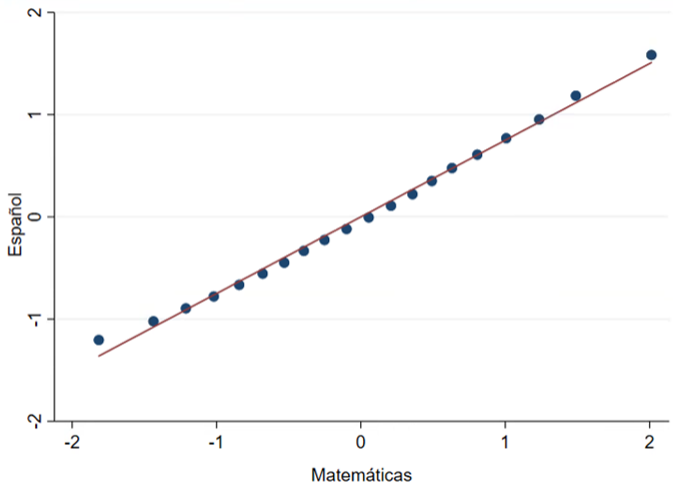


Figura 3.12: Correlación entre resultados español y matemáticas

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 surjan de procesos cognitivos distintos. Es decir, el buen dominio de la 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.

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

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.

Una vez que se estandarizan las calificaciones por materia, grado y año es posible calcular el promedio de la escuela y la diferencia por periodos anteriores. Más adelante se explicara con detalle la construcción de la variable "cambio" y "rendimiento decreciente".

Con el objetivo de entender las interacciones entre el cambio y las características de las escuelas, se calculó la correlación de las variables del CEMABE, y del formato 911 con el cambio de resultados ENLACE.

Tabla 3.7: Correlaciones cambio ENLACE y variables del F911 de inicio de cursos

Nombre de variable	Correlación	Descripción
V833	-0.083	Total de mujeres que son profesoras de idiomas
V835	-0.074	Total de mujeres que trabajan como personal administrativo, auxiliar y de servicios
V838	-0.078	Total de mujeres que trabajan como secretarias
V915	-0.084	Gasto promedio anual en inscripción
V916	-0.081	Gasto promedio mensual en colegiatura
V917	-0.120	Número de mensualidades que se pagan

La tabla 3.7 muestra las 6 variables del formato 911 de inicio de escuelas generales con mayor correlación con el cambio en la calificación promedio ENLACE por escuela. Es interesante como las tres últimas variables están relacionadas con el gasto en la escuela y las dos primeras con el género del personal escolar.

Tabla 3.8: Correlaciones cambio ENLACE y variables del F911 de fin de cursos

Nombre de variable	Correlación	Descripción
VAR678_F	-0.060	Número de profesores de actividades artísticas
VAR680_F	-0.086	Número de profesores de idiomas
VAR681_F	-0.066	Número total de personal administratio, auxiliar y de servicios
VAR682_F	-0.065	Número total de personal

Más adelante, la tabla 3.8 muestra las variables del formato 911 de fin de cursos con mayor correlación con el cambio. En este caso, las cuatro variables están relacionadas con el personal de la escuela y se repite al igual que al inicio el personal de idiomas. Esto es interesante porque el cambio se está calculando con respecto a la calificación de matemáticas no de otros idiomas. Es posible que aprender otros idiomas este relacionado con desarrollo de conexiones neuronales que son después utilizadas en matemáticas.

Finalmente, la tabla 3.9 muestra las variables del CEMABE con mayor correlación con el cambio en la calificación de matemáticas de ENLACE.

Es interesante que de nuevo resalta la importancia del personal femenino en el centro de trabajo y se incorporan características del inmueble.

Tabla 3.9: Correlaciones cambio ENLACE y variables del CEMABE

Nombre de variable	Correlación	Descripción
P13A	0.066	Material de la barda o cerco perimetral
P303	-0.088	Personal femenino en Centros de Trabajo
P34	-0.076	Total de tazas sanitarias
P22	0.055	Drenaje
P17A	0.055	Fuente principal de abastecimiento de agua
P16	0.050	Material del piso del inmueble

3.4 Verificación de calidad de datos

3.4.1 Calidad de ENLACE

Una desventaja es que la prueba ENLACE ha sido criticada en varías ocasiones por inflación de resultados y falta de control en la aplicación [33].

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 [34] [5].

La tabla 3.10 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 [35]

Tabla 3.10: 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 %

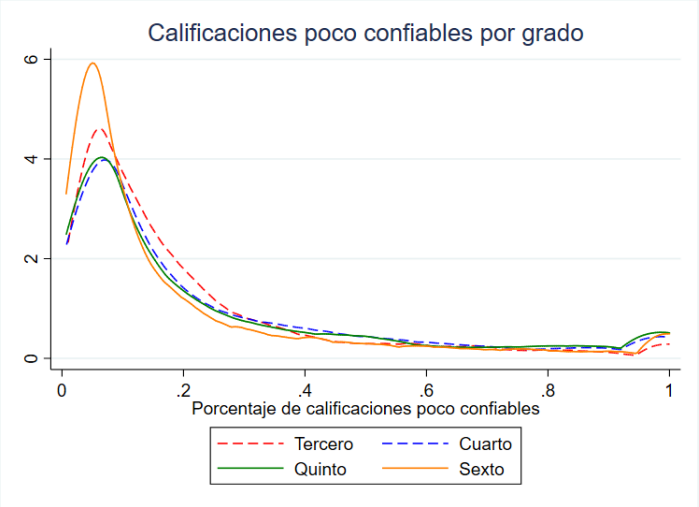


Figura 3.13: Porcentaje de copia por escuela

La figura 3.13 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 3.11).

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

Tabla 3.11: Porcentaje por año de alumnos con resultados poco confiables

Año	Porcentaje alumnos copia
2010	5.4 %
2011	4.8 %
2012	5.5 %
2013	4.8 %

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.

Otra desventaja de ENLACE es que 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 solo 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 **??** muestra el número de escuelas participantes por año y por estado.

3.4.2 Calidad del F911 y CEMABE

El levantamiento de información del CEMABE se realizó del 26 de septiembre al 29 de noviembre del 2013. Mientras que la recopilación del formato 911 de inicio de cursos del 2013 fue del 19 de agosto hasta el 31 de diciembre. Las fechas se empalman. Además, 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 CEMABE se realizó en 177,829 escuelas generales de nivel preescolar, primaria, secundaria, CAM o de educación especial. Del número de escuelas censadas,

el 44 % 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 CEMABE 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 3.12 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 3.12: 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 3.13 muestra los porcentajes inversos a la tabla 3.13. 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.

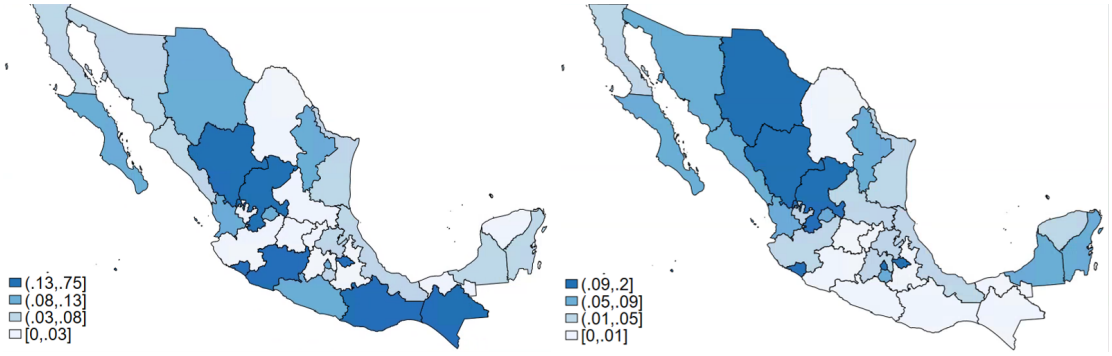
Las figuras 3.14a y 3.14b muestran el porcentaje [6] de escuelas por estado que no están en la intersección de las tablas. Por un lado, cabe destacar que de Querétaro solo una escuela que fue censada por el CEMABE no llenó el formato 911 y solo 5 escuelas que llenaron el formato 911 no fueron censadas 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.

Tabla 3.13: 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 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 3.14: 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 base. 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 3.15 muestra una gráfica de dispersión por bloques de la variable de matrícula del F911 y del CEMABE.

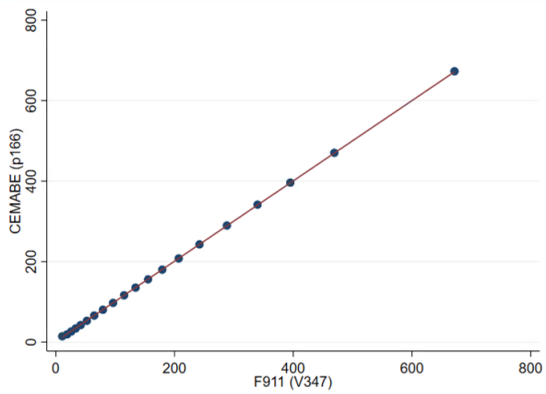


Figura 3.15: Gráfica de dispersión por bloques de la matricula por escuela en ambas bases

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 constan 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 se construyó utilizando 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 creó 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".

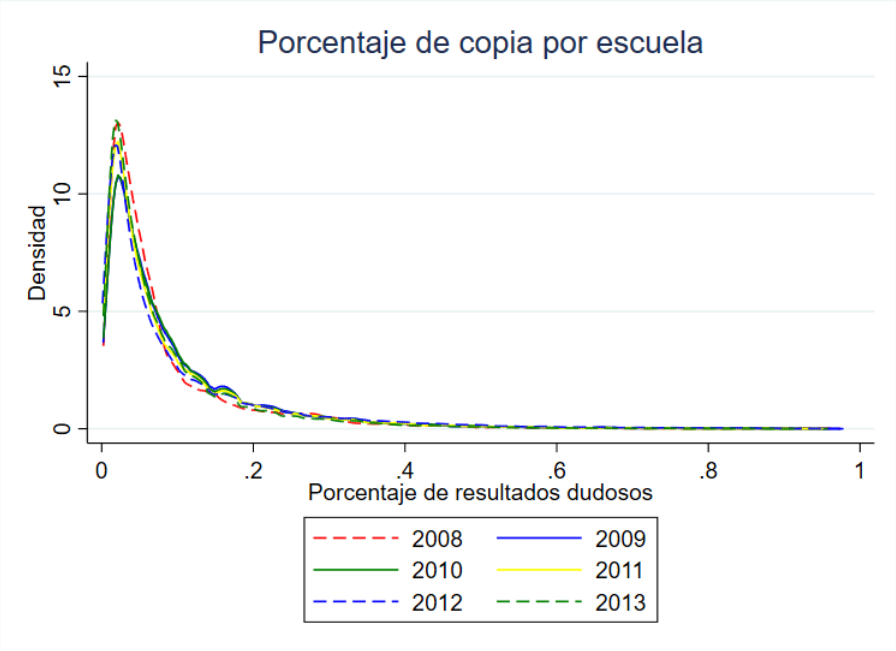


Figura 4.1: Distribución de los porcentajes de alumnos que "copiaron" por escuela

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 alumnos con calificaciones fuera del rango.

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

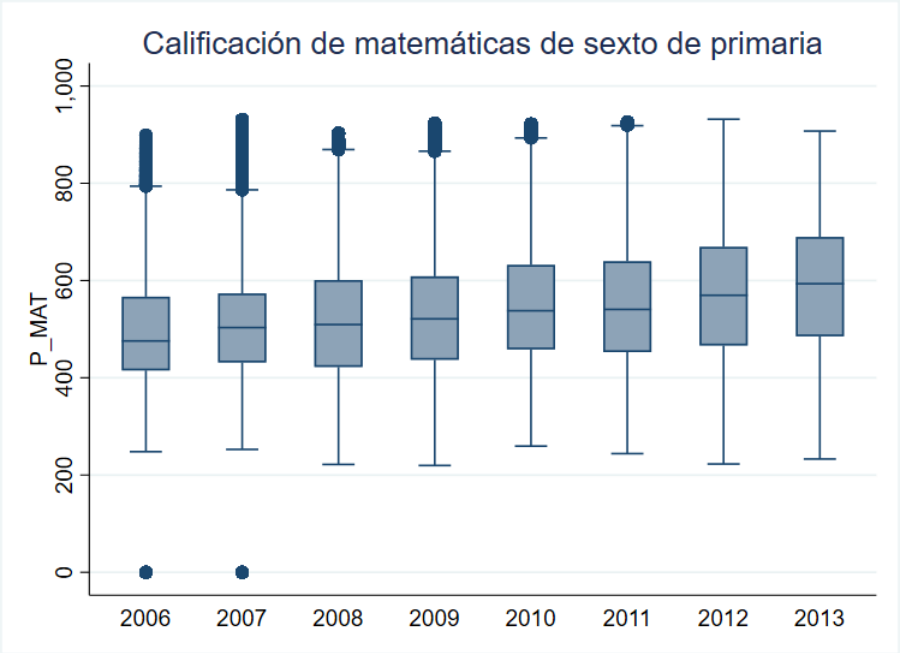


Figura 4.2: Diagrama de cajas y bigotes de las calificaciones de sexto de primaria por año

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 [36]. 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.

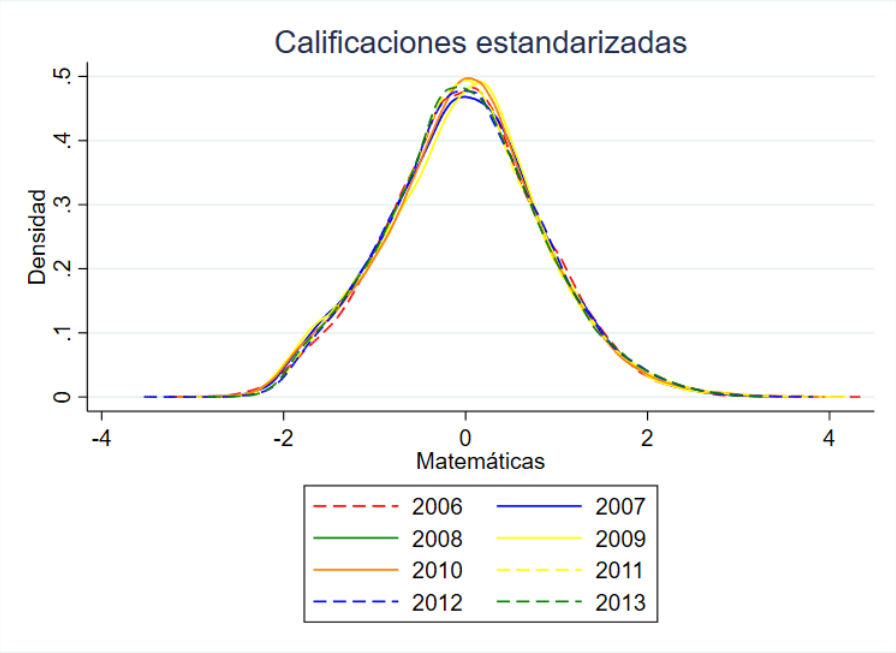


Figura 4.3: Distribución de calificaciones estandarizadas

Finalmente, el "cambio en desempeño" es la diferencia entre el promedio de las calificaciones de un año por escuela con el promedio de otro año. Se calcularon las diferencias tomando el 2013 como base. Es decir, se calculó la diferencia de resultados entre el 2013 y el 2012, 2011, 2010, 2009 y 2008. De forma que la diferencia de años más grande fue de cinco años y la más pequeña de un año. Con la variable de cambio se construyó la variable de "desempeño decreciente" que toma valor de 1 si el cambio negativo y de magnitud mayor a 0.2.

La figura 4.4 muestra la distribución de cambios entre años. Diferencia se refiere al número de años entre los cuales se está calculando la diferencia. Es interesante que mientras menor sea la diferencia de años, menor parece ser el cambio. Es decir, cuando las diferencias se hacen notables con el tiempo.

Utilizando estos cambios, se construyó la variable objetivo de "rendimiento decreciente". Las escuelas con un cambio negativo en valor absoluto mayor a 0.2 se clasificaron

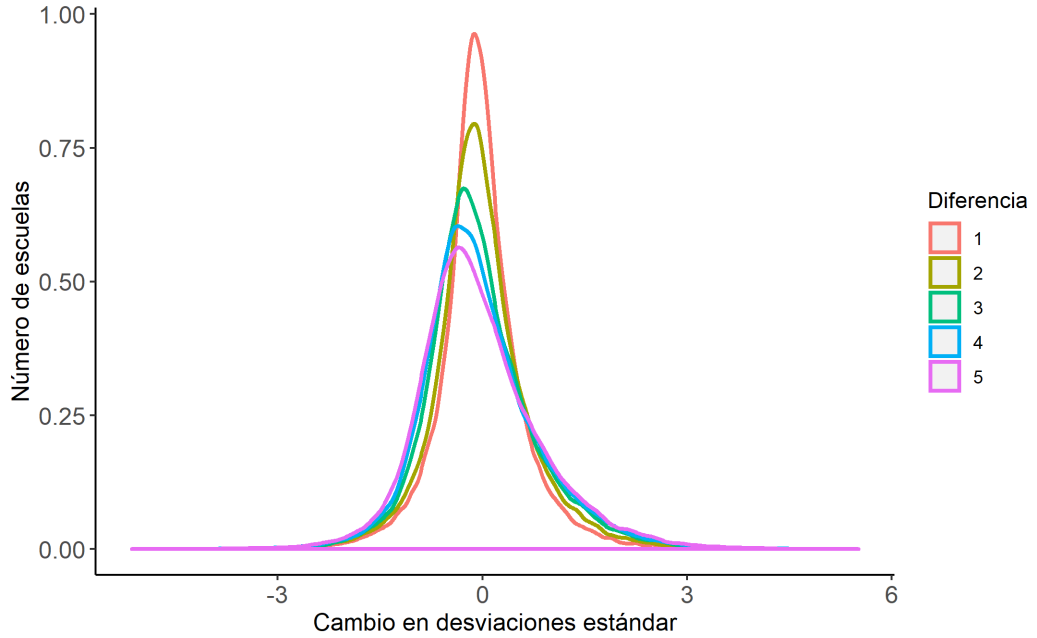


Figura 4.4: Distribución de cambios entre distintos tamaños de periodos

positivamente.

La figura 4.5 muestra el número de escuelas clasificadas con "rendimiento decreciente" (1) y como "rendimiento constante o creciente" (0) por diferencia entre años. Los resultados son consistentes con la figura 4.4 ya que el número de escuelas con rendimiento decreciente y no es muy similar cuando la diferencia son 5 años y cuando la diferencia es 1 año, existen más escuelas con rendimiento constante o creciente.

4.2 Variables independientes

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

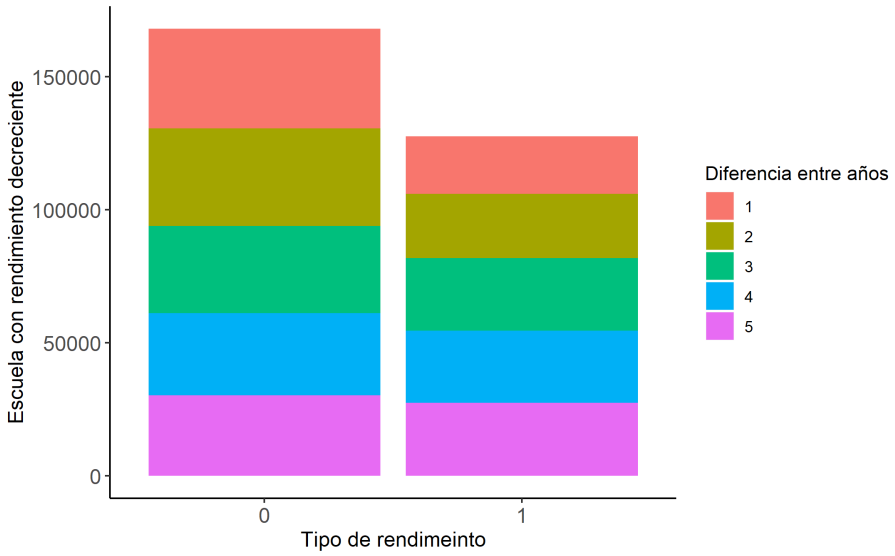


Figura 4.5: Número de escuelas por tipo de rendimiento por diferencia de años

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 [37].

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.

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 excluyeron

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 "id_estado" que asigna un número a cada estado.

Más adelante, se excluyeron las variables con varianza baja. Se normalizaron las variables numéricas 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 construyeron Bosques Aleatorios (Random Forests, en inglés) para identificar las variables significativas. Los Bosques Aleatorios son un método incrustado (embedded method, en inglés) que permite identificar la contribución de las variables en las decisiones. Esto lo hace construyendo cientos de árboles de decisión con una muestra de los datos y de las variables. En cada árbol se calcula que tan buena partición hizo cada variable y se promedian las calificaciones de todos los árboles para obtener la importancia de variables global [38].

Con esto en mente, se construyeron múltiples bosques y se guardaron en una lista las variables con importancia mayor al máximo entre 0.005 y al cuartil 65 de importancia de variables. Para asegurarse de que las variables contuvieran información relevante para todos los estados y para todos los años, se corrieron 33 árboles (uno para cada estado más uno con todos los estados) 5 veces (uno para cada año de diferencia). La unión de las variables seleccionadas para cada estado y cada diferencia fueron las variables identificadas como importantes.

En la primer etapa, una vez identificadas las variables se examinaron a detalle y se crearon nuevas variables a partir de ella. En la segunda etapa, se mejoró la selección de variables. Una de las desventajas de los árboles aleatorios para selección de variables es que variables con alta correlación son otorgadas importancia similar. Esto es una desventaja porque son variables no informativas. Por lo tanto, en la segunda etapa se eliminaron las variables con correlación mayor a 0.95. Asimismo, para obtener la primera lista de variables significativas, se rellenaron los valores faltantes con ceros. En la segunda etapa, se les dio un tratamiento correcto a los valores faltantes y después se volvió a correr el algoritmo para obtener una segunda lista filtrada de variables importantes.

Cabe resaltar una vez identificadas las variables importantes, se seleccionaron del conjunto original sucio para el resto del analisis.

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 relleno de 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 [39]. 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 [40]. "Miss Forest" se implementó utilizando la función ExtraTreesRegressor de sklearn.ensamble. Se imputaron valores solamente en las columnas y renglones para los cuales faltarán menos del 20 % de los datos. Asimismo, cabe mencionar que no se utilizó la variable objetivo para la imputación de datos.

4.2.3 Construcción de nuevos datos

De forma similar a las otras secciones de preparación de los datos, se construyeron nuevos datos en dos etapas. 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 A.

- Proporción mujeres-hombre de alumnos (alumnas mujeres / alumnos hombres) [41].
- Porcentaje de maestros con grado igual o mayor a licenciatura [41].
- Número de alumnos por maestro [42].
- Porcentaje de alumnos repitiendo grado [43].
- Número de años promedio en preescolar [44]

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

Cabe mencionar que el modelo busca predecir cambios, por eso se generaron variables de cambio (con la resta y la diferencia) entre las características al inicio del ciclo escolar, al final del ciclo escolar y entre años.

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 [45]. 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. Dado que uno de los requerimientos funcionales es la interpretabilidad de los modelos, se utilizaron las variables construidas en la primera etapa.

4.3 Integración y formato de datos

Los datos del CEMABE, F911 de inicio de cursos, F911 de fin de cursos y la variable objetivo fueron integrados a través del CCT y del turno de la escuela.

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 crearon variables indicadoras para las variables categóricas.

CAPÍTULO 5

MODELADO

Este capítulo utilizará el procesamiento de datos del capítulo anterior para cumplir con los objetivos de minería de datos y responder a la pregunta: ¿Cuáles escuelas están en riesgo de tener bajo desempeño y qué características están relacionadas?

5.1 Selección de técnicas de modelado

Una posible técnica de modelado es usar una red neuronal recurrente de memoria bidireccional corto plazo prolongado (del inglés, Long short-term memory LSTM) como en las soluciones relacionadas [27]. Esto se podría implementar utilizando los años como las observaciones en el tiempo. Sin embargo, los principales problemas de esta alternativa son que el número de observaciones en el tiempo es muy pequeño. Es decir, solo se puede seguir a las escuelas por ocho años y algunas escuelas participaron menos años. Asimismo, las redes neuronales son muy difícil de interpretar por ser modelos muy complejos. Como resultado, tomando el cuenta el número de observaciones, conviene examinar modelos de clasificación.

Se seleccionaron varios modelos de aprendizaje supervisado. Entre ellos modelos lineales como Regresión Logística, de agrupamiento como k-vecinos más cercanos y basados en árboles como Bosques Aleatorios y bosques con gradiente con "Boosting"(del inglés, Gradient Tree Boosting).

En primer lugar, el modelo de regresión logística es en realidad un modelo de regresión Bernoulli con un liga logística. Se escogió por su simplicidad y ser el modelo más fácil de interpretar.

En segundo lugar, los métodos de ensamble de árboles en general tienen muy buenas métricas predictivas, poco sobre entrenamiento y a diferencia de los modelos de regresión lineales, toma en cuenta interacción entre variables. Una variación de los bosques aleatorios son los bosques extremadamente aleatorios (Extremely Randomized Trees, en inglés). Este modelo, a diferencia de un bosque aleatorio tradicional, divide los nodos de forma aleatoria y no utiliza muestras bootstrap. Como ventajas sobre los bosques aleatorios, los bosques extremadamente aleatorios suelen ser más rápidos, computacionalmente menos costosos y tiene mejor desempeño frente a variables con ruido [46].

Otra variación de los bosques aleatorios es utilizar gradiente con "Boosting". Estos modelos han tenido muy sus buenos resultados en competencias [47] de clasificación y a diferencia de los bosques tradicionales, construye árboles de decisión de forma secuencial.

En cuanto a los datos, el modelo de regresión logística asume que los datos están normalizados. Los modelos basados en árboles no necesitan datos normalizados pero tampoco cambia su desempeño si están normalizados o no. Por lo tanto, se utilizará la misma base de datos normalizada para probar los diferentes modelos.

5.2 Generación de un diseño de comprobación

Con el fin de probar y evaluar los modelos, se separó el 10 % de la muestra para calcular las métricas de desempeño. Es decir, el 90 % de la muestra se utilizó para entrenar los modelos y el 10 % para calcular el error.

Nos interesa que la clasificación sea lo más precisa y exhaustiva posible. La precisión, de acuerdo a la ecuación 5.2, indica cuántas de las escuelas clasificadas con "rendimiento decreciente" verdaderamente tienen rendimiento decreciente. Es decir, cuántos

de los elementos seleccionados son relevantes.

$$Precisión = \frac{verdaderos\,positivos}{verdaderos\,positivos + falsos\,positivos} \tag{5.1}$$

Sin embargo, la precisión no es una buena métrica si las clases no están completamente balanceadas. Por ejemplo, si el 99 % de las escuelas tienen "rendimiento decreciente" entonces si se clasifica todas las escuelas con "rendimiento decreciente" se obtendría una precisión de 0.99. Es por eso que también nos interesa la exhaustividad. La exhaustividad (recall, en inglés), de acuerdo a la ecuación **??** indica la proporción de escuelas que verdaderamente tenían "rendimiento decreciente" entre el total de escuelas clasificadas con "rendimiento decreciente". En otras palabras, cuántos elementos relevantes fueron seleccionados.

$$Exhaustividad = \frac{verdaderos\,positivos}{verdaderos\,positivos + falsos\,negativos} \tag{5.2}$$

Dado que nos interesan ambas métricas, se utilizará el Valor-F que es una media armónica entre la precisión y exhaustividad. La ecuación 5.3 muestra como calcular en valor-F.

$$F_1 = 2 \cdot \frac{Exhaustividad \cdot Precisión}{Exhaustividad + Precisión} \tag{5.3}$$

5.3 Generación de los modelos

En primer lugar, se encontraron los "mejores" parámetros para cada modelo. Es posible modificar los parámetros de los modelos con el fin de encontrar la configuración que optimice el desempeño del modelo. La tabla **??** muestra el espacio explorado de parámetros para cada modelo.

Tabla 5.1: Parámetros explorados por modelo

Modelo	**Parámetros explorados**
Regresión Logística	El número máximo de iteraciones (convergencia): se probó con 70, 100 (por omisión), 200 y 500. Valores de C (regularización inversa): se probó con 0.01, 0.1, 0.3, 0.5, 0.8, 1 (por omisión), 2 La proporción de regularizador l1 y l2: se probó con 0.3, 0.5, 0.8 El tipo de penalización (tipo de regularizadores): se probó con elasticnet (mezcla de LASSO y Ridge), LASSO (L1) y Ridge (L2)
KN (K-Nearset Neighbors)	El número de vecinos más cercanos: se probó con 2, 5 (por omisión), 10, 20, 30 y 50
Bosque Aleatorio (Random Forest)	El número de estimadores: se probó con 50, 100 (por omisión), 200, 400 y 500 estimadores. La profundidad máxima del árbol: se probó sin limite (por omisión), 10 y 50. El número máximo de variables: se probó con la raíz del número de observaciones (auto, por omisión) y con el logaritmo en base 2 del número de observaciones. El número mínimo de observaciones necesarias para dividir un nodo: se probó con 2 (por omisión), 4 y 8
Bosque Extremadamente Aleatorio (Extra Tree)	El número de estimadores: se probó con 50, 100 (por omisión), 200, 400 y 500 estimadores. La profundidad máxima del árbol: se probó sin limite (por omisión), 10 y 50. El número máximo de variables: se probó con la raíz del número de observaciones (auto, por omisión) y con el logaritmo en base 2 del número de observaciones. El número mínimo de observaciones necesarias para dividir un nodo: se probó con 2 (por omisión), 4 y 8
XGBoost	La tasa de aprendizaje: se probó con 0.2, 0.3 (por omisión) y 0.5. La profundidad máxima del árbol: se probó con 6 (por omisión), 9 y 12. El valor de submuestra: se probó con 1 (por omisión) y 0.8. El valor de submuestra de columnas: se probó con 1 (por omisión) y 0.8.

Para encontrar la combinación óptima, se construyeron modelos con todas las posibles combinaciones. Esto se hizo con ayuda de la función *GridSearchCV*. La función utilizó el valor F_1 como métrica para escoger la mejor combinación de parámetros. Asimismo, para los modelos se utilizó la técnica de validación cruzada. Se escogió hacer la validación cruzada con 20 % de la muestra. A diferencia de la validación tradicional que entrena con el 80 % de los datos y prueba el modelo con el 20 % restante, validación cruzada hace este proceso iterativo. Es decir, si tuviéramos 100 observaciones, entonces primero entrena con las observaciones 1-80 y prueba con las observaciones 81-100; y después entrena con las observaciones 21-100 y prueba con las observaciones 1-20. Esto se repite de tal forma que todos los datos son usados para validar y entrenar en algún momento.

En segundo lugar, habiendo escogido los mejores parámetros para cada modelo, se entrenaron los modelos para comparar los valor f1 resultantes.

5.4 Evaluación de los modelos

La tabla 5.2 resume los valores máximos del F_1 por modelo para las 5 diferencias entre periodos. Visualmente se pueden observar las diferencias en la figura 5.1

Tabla 5.2: Resultados por modelo y diferencia entre periodo

	1	2	3	4	5
LogisticRegression	0.66	0.66	0.68	0.69	0.71
KNeighbors	0.59	0.58	0.60	0.60	0.62
RandomForest	0.66	0.66	0.69	**0.70**	0.71
ExtraTreesClassifier	0.62	0.64	0.67	0.69	0.70
XGBClassifier	**0.67**	**0.67**	**0.69**	0.69	**0.72**

Los modelos con mejor desempeño, de acuerdo al valor F_1, fueron los modelos basados en árboles. De forma similar a resultados de competencias de clasificación, el

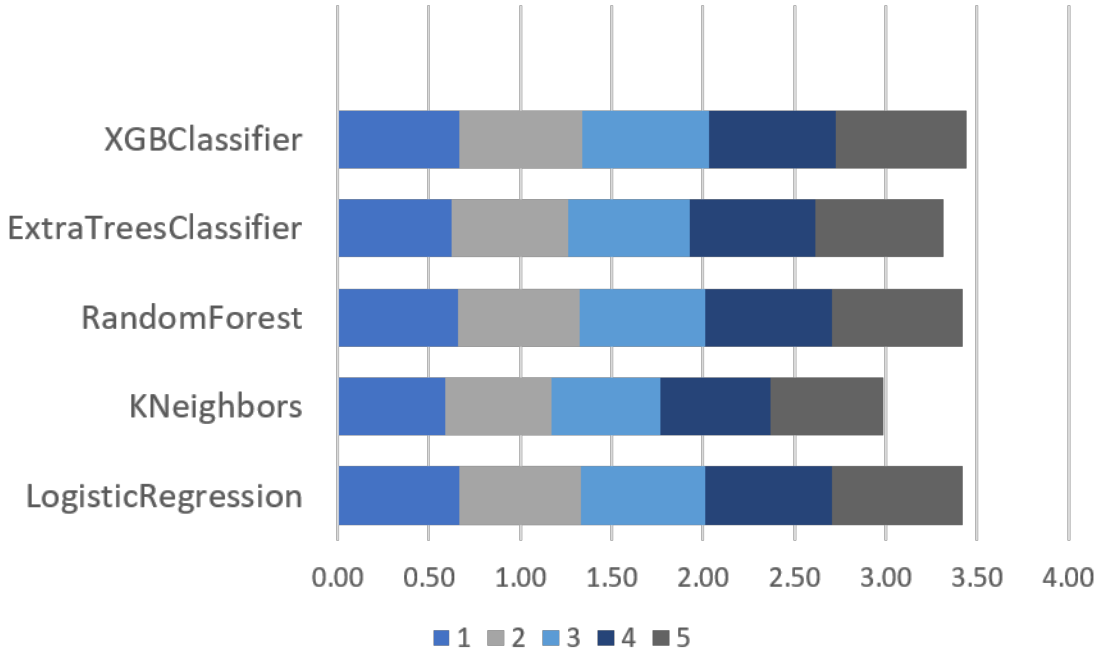


Figura 5.1: Valor F_1 de los modelos

clasificador XGBoost obtuvo los valores más altos en la mayoría de los casos. Los parámetros óptimos fueron una tasa de aprendizaje de 0.2 (menor a la tasa por omisión), profundidad máxima del árbol de 6 (por omisión), y un valor de submuestra de 1.

La figura 5.2 muestra la importancia de las 10 variables más significativas según el modelo XGBoost con parámetros óptimos.

Una de las mayores desventajas de los métodos basados en árboles es la dificultad de interpretar el efecto de las variables. Es decir, la proporción de alumnos y personal (alum_personal_prop) es muy significativa pero no sabemos si la relación es proporcional o inversamente proporcional.

Como se ve en la figura 5.1, el modelo Bernoulli con liga logística tuvo resultados muy similares a los resultados de los árboles. La ventaja de la regresión es que es muy fácil de interpretar y requiere menor poder de computo y memoria. La figura 5.3 muestra los coeficientes de le regresión. Es interesante que ambas figuras coinciden

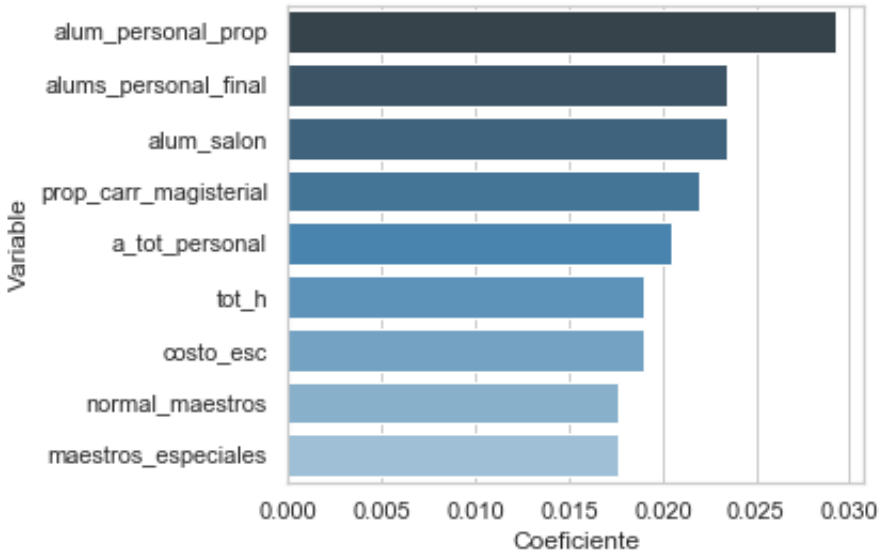


Figura 5.2: Variables importantes según XGBoost

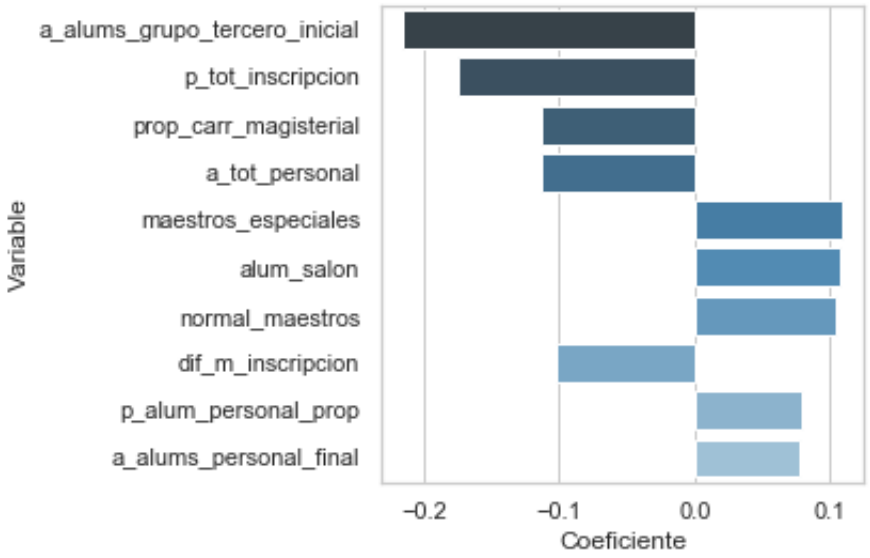


Figura 5.3: Coeficientes de regresión lineal

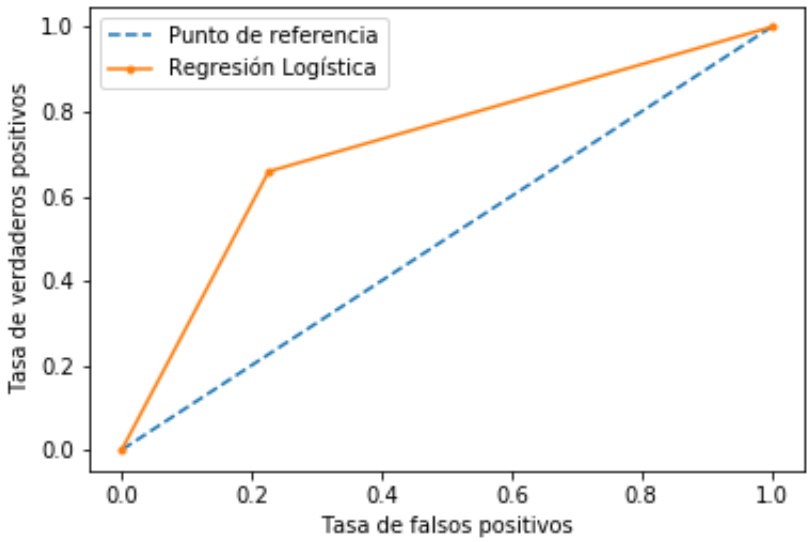


Figura 5.4: Área bajo la curva: 0.72

en la importancia de algunas variables.

Asimismo, la figura 5.4 muestra la curva de Característica Operativa del Receptor (ROC) que gráfica de sensibilidad frente a la especificidad. La diagonal es equivale a haber tomado decisiones aleatorias mientras que abajo de la diagonal son predicciones peores que las aleatorias.

Para determinar si los resultados son satisfactorios o no, vale la pena estimar no sólo en valor F_1 del modelo, si no el valor F_1 de no haber construido ningún modelo. Esto se puede hacer clasificando todas las escuelas con la clase mayoritaria y calcular el valor F_1. Cabe recordar que uno de los requerimientos es que el modelo se pueda aplicar a diferentes estados de la república. Con esto en mente, la tabla 5.3 muestra los valores F_1 por estado y por diferencia de años y la tabla 5.4 muestra la ganancia o pérdida de acuerdo al punto de referencia. Estas tablas se generaron utilizando un modelo de regresión logística con un máximo de 200 iteraciones, regularizador Ridge con una fuerza de regularización de 1.25.

Tabla 5.3: Valores F1 por estado y diferencia entre años

	1	2	3	4	5
Aguascalientes	0.59	0.69	0.76	0.75	0.79
Baja California	0.57	0.67	0.71	0.76	0.76
Baja California Sur	0.62	0.64	0.55	0.58	0.68
Campeche	0.53	0.71	0.78	0.63	0.73
Coahuila	0.61	0.66	0.70	0.68	0.72
Colima	0.62	0.54	0.75	0.68	0.76
Chiapas	0.64	0.66	0.70	0.78	0.74
Chihuahua	0.58	0.62	0.67	0.68	0.71
Distrito Federal	0.68	0.70	0.71	0.71	0.77
Durango	0.61	0.57	0.68	0.67	0.71
Guanajuato	0.66	0.68	0.67	0.73	0.71
Guerrero	0.53	0.66	0.65	0.69	0.67
Hidalgo	0.61	0.67	0.66	0.67	0.68
Jalisco	0.62	0.67	0.67	0.68	0.70
México	0.64	0.67	0.65	0.70	0.75
Michoacán	0.55	0.52	0.64	0.63	-
Morelos	0.56	0.68	0.73	0.76	0.73
Nayarit	0.62	0.61	0.64	0.60	0.67
Nuevo León	0.56	0.62	0.63	0.71	0.72
Oaxaca	-	-	-	-	-
Puebla	0.64	0.65	0.71	0.73	0.73
Querétaro	0.62	0.63	0.69	0.67	0.73
Quintana Roo	0.63	0.59	0.54	0.67	0.76
San Luis Potosí	0.63	0.64	0.68	0.69	0.70
Sinaloa	0.64	0.64	0.65	0.69	0.72
Sonora	0.62	0.66	0.66	0.69	0.68
Tabasco	0.66	0.66	0.64	0.61	0.66
Tamaulipas	0.63	0.66	0.64	0.67	0.62
Tlaxcala	0.64	0.72	0.72	0.65	0.63
Veracruz	0.65	0.67	0.69	0.69	0.73
Yucatán	0.65	0.70	0.69	0.69	0.78
Zacatecas	0.65	0.69	0.71	0.62	0.68
Nacional	0.65	0.66	0.69	0.70	0.72

Tabla 5.4: Margen de ganancia sobre el punto de referencia por estado y diferencia entre años

	1	2	3	4	5
Aguascalientes	0.03	0.20	0.11	0.23	0.00
Baja California	0.03	0.33	0.21	0.31	0.43
Baja California Sur	0.07	0.25	0.21	0.00	0.01
Campeche	0.06	0.15	0.11	0.08	0.30
Coahuila	0.10	0.24	0.37	0.25	0.25
Colima	0.16	0.16	0.37	0.31	0.35
Chiapas	0.09	0.12	0.21	0.15	0.18
Chihuahua	0.15	0.18	0.27	0.33	0.23
Distrito Federal	0.11	0.14	0.34	0.32	0.24
Durango	0.11	0.11	0.24	0.19	0.16
Guanajuato	0.15	0.20	0.29	0.36	0.25
Guerrero	0.09	0.10	0.09	0.17	0.15
Hidalgo	0.09	0.13	0.21	0.24	0.23
Jalisco	0.06	0.23	0.29	0.33	0.36
México	0.15	0.10	0.22	0.29	0.40
Michoacán	0.10	0.17	0.24	0.16	-
Morelos	0.21	0.26	0.38	0.43	0.34
Nayarit	0.09	0.07	0.12	0.26	0.33
Nuevo León	0.09	0.27	0.14	0.29	0.27
Oaxaca	-	-	-	-	-
Puebla	0.11	0.23	0.32	0.40	0.32
Querétaro	0.06	0.21	0.34	0.24	0.36
Quintana Roo	0.05	0.15	0.20	0.17	0.24
San Luis Potosí	0.20	0.20	0.33	0.27	0.30
Sinaloa	0.09	0.13	0.11	0.15	0.19
Sonora	0.15	0.26	0.19	0.24	0.35
Tabasco	0.22	0.24	0.24	0.15	0.26
Tamaulipas	0.17	0.21	0.30	0.20	0.26
Tlaxcala	0.23	0.22	0.30	0.26	0.15
Veracruz	0.25	0.28	0.33	0.31	0.40
Yucatán	0.05	0.15	0.23	0.29	0.37
Zacatecas	0.20	0.21	0.22	0.22	0.27
Nacional	0.17	0.21	0.30	0.31	0.35

CAPÍTULO 6

EVALUACIÓN

6.1 Evaluación de los resultados

Se cumplió el primer sub-objetivo del proyecto de minería ya que se construyó un modelo de clasificación de escuelas con rendimiento decreciente. A través del modelo es posible responder ¿cuáles escuelas están en riesgo de tener bajo desempeño y qué características están relacionadas?. Asimismo, el modelo cumple es flexible ya que se puede adaptar a predicciones de un año a otro o en un intervalo de 5 años y a diferentes entidades federativas. La decisión de utilizar un modelo de regresión logística sobre un modelo basado en árboles permite que los resultados sean interpretados con mayor facilidad.

Se considera el proyecto exitoso por generar nuevo conocimiento. Sin embargo, su éxito real será una vez que sea usado para la toma de decisiones en programas sociales para escuelas primarias en México.

El objetivo del sector educación se cumplirá si se logra hacer una asignación más informada, transparente y específica y como resultado se logra mejorar la calidad educativa del país.

6.2 Proceso de revisión

Por un lado, una de las fortalezas del proyecto fue la sinergia de las dos culturas de modelos estadísticos [24]. Se utilizaron modelos algorítmicos como bosques aleatorios que logran identificar relaciones no-lineales entre los datos e interacciones de varia-

bles para crear nuevas variables y seleccionar las más importantes. Más adelante se seleccionó una regresión logística por ser un modelo más simple.

Por el otro lado, una de las debilidades fue el poco margen de ganancia del modelo sobre el punto de referencia para algunos estados y para algunos periodos de diferencia. En específico, las predicciones son peores entre años consecutivos y mejoran mientras más pasa el tiempo. Esto puede ser consecuencia de acciones de largo plazo que quizá tomen más de un sexenio en poder ser observadas.

Asimismo, es posible que falten variables informativas como cambios en los niveles de violencia de la localidad donde se encuentra la escuela, cambios climáticos o otros cambios externos a la escuela. Vale la pena invertir en nuevas variables de cambio para mejorar el modelo.

Finalmente, cabe resaltar que el modelo no captura información de Oaxaca ya que las escuelas generales no presentaron la prueba en el 2013. Margenes pequeños de ganancia como el del estado de Aguascalientes puede ser resultado del el bajo número de observaciones. Esto a su vez puede ser causa de la falta de control en la aplicación de ENLACE [33]. Puede que no se tenga información suficiente para responder la pregunta con las variables o puede ser que el supuesto de que las calificaciones son verdaderas, informativas y significativas sea falso.

6.3 Determinación de los pasos siguientes

Además de las escuelas generales, en México existen escuelas comunitarias e indígenas. El paso siguiente es incluir las escuelas indígenas y comunitarias en el análisis.

Asimismo, como fue mencionado previamente, en una segunda versión, es deseable incluir características de las localidades y cambios externos a las escuelas.

Como trabajo futuro, también sería interesante construir un modelo de predicción a

nivel alumno utilizando características personales. Se espera que los modelos a nivel alumno, incluyendo características de la escuela expliquen mucho más que los modelos a nivel escuela. Asimismo, podría agregar valor utilizar métodos de aprendizaje de máquina para observar los factores que diferencian a las escuelas y el valor agregado de cada una [26].

Finalmente, dado las limitaciones de la prueba ENLACE, en un futuro se pueden crear modelos que se centren en el la deserción en vez de calificación académica. Esto se puede implementar utilizando el cambio de matricula por ciclo escolar reportado en el formato 911.

CAPÍTULO 7

DISTRIBUCIÓN

Una vez que se han evaluado los resultados, se puede distribuir el conocimiento. En este capítulo se explora el despliegue del producto de datos. Esto con el fin de cumplir el segundo sub-objetivo de crear una aplicación web para hacer disponible el modelo y sus resultados

7.1 Planificación de distribución

La estrategia para la distribución fue a través de una aplicación web. Las ventajas de las aplicaciones web es que pueden ser accedidas desde cualquier lugar geográfico con internet y un navegador. De esta forma se pretende que el proyecto tenga mayor alcance.

La aplicación web permite a los usuarios hacer consultas para estados y diferencia de periodo en específico. De forma que los usuarios pueden realizar las siguientes operaciones:

1. Seleccionar los estados y el periodo de tiempo del cual desee obtener información
2. visualizar en un mapa las escuelas con rendimiento decreciente
3. Obtener la clave de la escuela al dar click sobre un punto en el mapa
4. Descargar una lista con clave de la escuela y su clasificación
5. Visualizar las variables más significativas con sus coeficientes
6. Visualizar el número de observaciones y el valor F_1 de la clasificación

7.1.1 Aplicación web

Se construyó la aplicación Web utilizando la herramienta "Dash". Dash es un entorno de trabajo de Python que está basado en Flask y React. Se escogió este entorno de trabajo por la facilidad de implementar el código previamente escrito en Python de creación de mapas y de aprendizaje de máquina como el bosque aleatorio para la limpieza de datos y modelos de Regresión Logística.

Para la construcción de la aplicación se genero un archivo Çlassifier.pyçon los métodos de limpieza y de modelado y el archivo principal .ªpp.py"que llama a los métodos y despliega el mapa, las tablas y el menú.

7.1.2 Alojamiento web

Para el despligue existen varías alternativas, entre ellas utilizar plataformas como Amazon Web Services, Microsoft Azure o Heroku. En un principio se eligió utilizar Heroku por su facilidad de vincular la aplicación con un repositorio de Github. Sin embargo, en Heroku corren las aplicaciones sobre una máquina Linux sin opción de hacer grandes modificaciones al entorno. Por esa razón, se optó por hacer el despligue en Microsoft Azure desde un contenedor de Docker.

Se construyeron dos imágenes de Docker. Docker es una plataforma para el desarrollo, migración y ejecución de aplicaciones utilizando la tecnología de virtualización de contenedores [48]. Utilizando un archivo Dockerfile es posible crear nuevas imágenes que son utilizadas para crear contenedores.

La primera imagen (paolamedo/dash-sql-azure) está construida sobre una máquina Ubuntu versión 16. Una de las mayores complicaciones del despliegue fue acceder a un servidor SQL de Microsoft desde Ubuntu. Para esto fue necesario instalar drivers y programas especiales como mssql. La instalación de programas, drivers y aplicaciones

como Python es tardada y puede ser utilizada en muchos otros proyectos por eso se eligió separarla del código de este proyecto.

La segunda imagen (paolamedo/sql-azure) está construida sobre la primer imágen y agrega los archivos específicos del proyecto. Asimismo, expone el puerto 8050 (sobre el cual corre la aplicación de Dash) y automáticamente empieza la aplicación.

La ventaja de tener ambas imágenes es que cualquier cambio en el código solo modifica la segunda imagen que se construye en poco tiempo. Ambas imágenes se construyeron de forma local y una vez probadas fueron agragadas a DockerHub, una biblioteca en línea de imagenes.

Finalmente, se creo una aplicación web en Microsoft Azure y se vinculó con la imagen paolamedo/sql-azure. La aplicación se actualiza automáticamente cuando la imagen se actualiza. Al igual que en la construcción de la imágen, fue necesario exponer el puerto 8050 para poder acceder a la aplicacipon. Se escogió un plan de aplicación con 1 GB de memoria y 60 minutos de computo al día por restricción presupuestal. Sin embargo, es posible mejorar la velocidad y memoria en cualquier momento.

Las figuras 7.1 y 7.2 muestran la interfaces de la aplicación recibiendo solicitudes remotas. La figura 7.1 muestra el menú de selección y la explicación introductoria mientras que la figura 7.2 muestra el mapa, la lista de variables importantes, el vínculo para descargar la lista des escuelas y las métricas de resultados. La aplicación está disponible en la siguiente dirección: https://enlace-performance.azurewebsites.net/

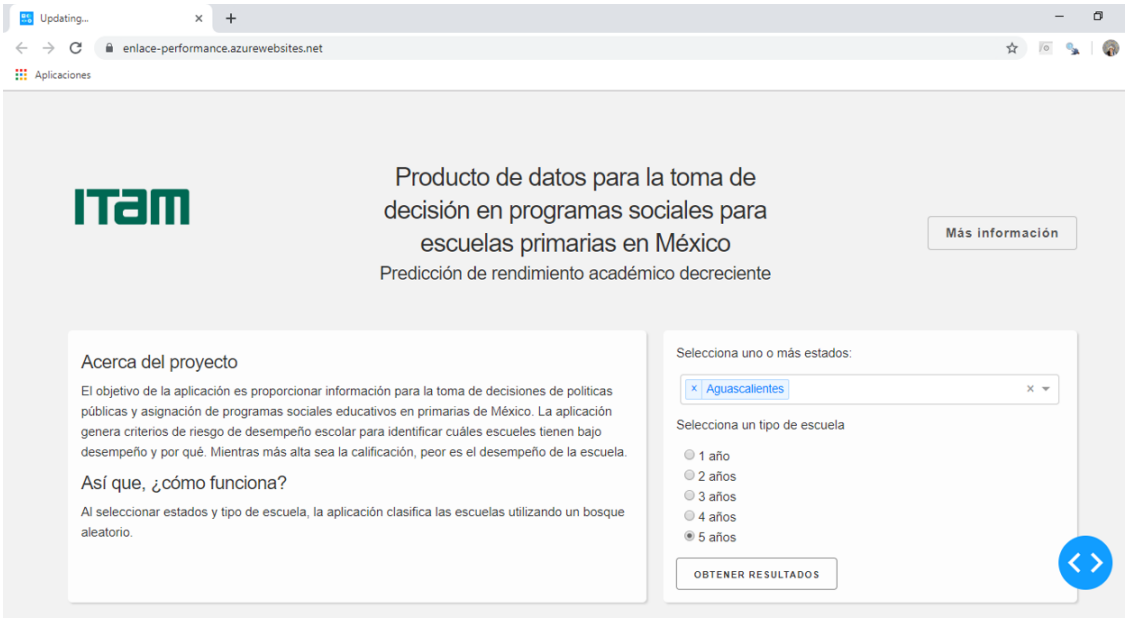


Figura 7.1: Interfaz superior de aplicación web

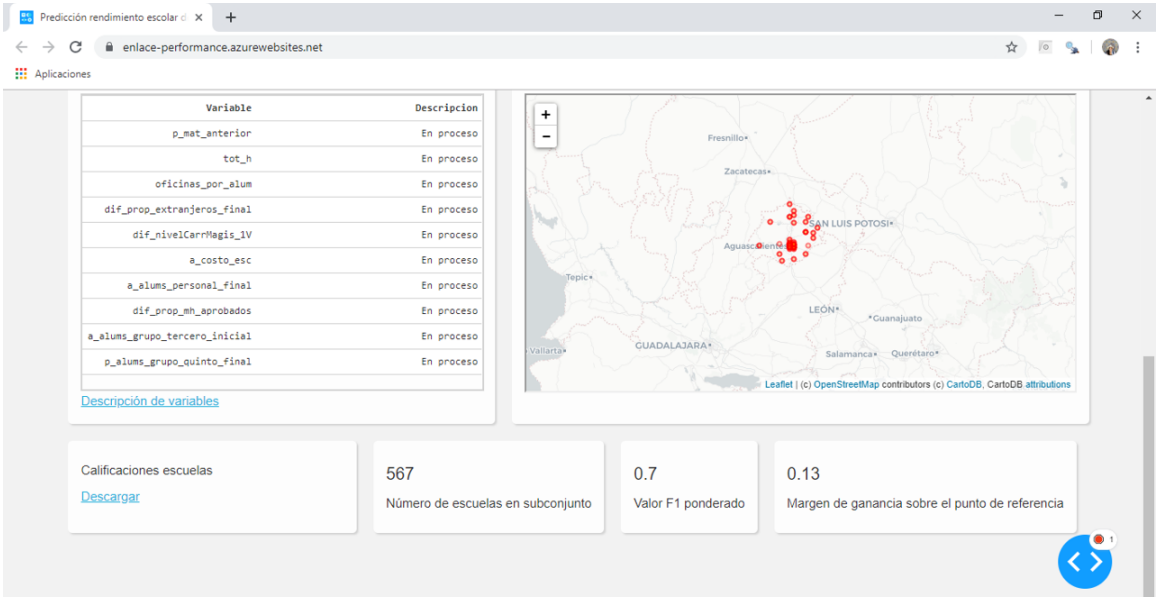


Figura 7.2: Interfaz inferior de aplicación web

7.1.3 Base de datos

Se creo una base de datos y un servidor remoto SQL en Azure. Los datos procesados en formato de texto fueron convertidos en un script SQL. Para poder acceder a la

base de datos desde la aplicación se editó la configuración del firewall de la base para que permitiera solicitudes desde cualquier IP. El almacenamiento tiene un tamaño máximo de 13 GB y un costo mensual aproximado de 37 pesos. Por el momento, el costo de la base de datos está siendo subsidiado por los 100 dólares gratis de prueba como estudiante.

7.2 Planificación de control y mantenimiento

Azure facilita el control y el mantenimiento de la aplicación con estadísticas semanales y notificaciones en caso de que ocurran errores. El mantenimiento será fácil gracias a la buena documentación de código.

Es importante mantener la aplicación vigente actualizando los datos. Actualmente, ENLACE no es vigente. En su lugar está Planea que tiene menor alcance pero información puede ser utilizada para actualizar los datos y los modelos. Asimismo, para darle mantenimiento tiene sentido también actualizar la información de las escuelas con el formato 911 de años futuros.

7.3 Revisión final del proyecto

Hubo tres grandes aprendizajes del proyecto: la importancia de los objetivos del negocio; utilizar modelos como medio, no como fin; y que poco predictibilidad también es un resultado.

En primer lugar, el valor de un proyecto es proporcional al impacto que puede generar. En el caso del proyecto, utilizar una red neuronal suena atractivo pero no responde a los requerimientos de transparencia e interpretabilidad del sector educativo.

Esto está muy ligado al segundo aprendizaje, los modelos algorítmicos y paquetes estadísticos (como Sci-kit Learn) son herramientas que deben ser utilizadas una vez

que los datos han sido comprendidos. En un principio, se intentó construir modelos con todas las variables, en muchos casos, "sucias" y sin relevancia. El aprendizaje fue analizar las datos, buscar correlaciones y utilizar conocimiento del sector educativo para crear nuevas variables más significativas como proporciones o promedios. Asimismo, se contruyó una red recurrente LSTM con varias capas cuyos resultados eran muy similares a los de una regresión lineal. Si bien existe valor en probar con nuevas técnicas, se debe entender el problema primero antes de entrenar un modelo.

Finalmente, no existe una recenta mágica. Se utilizó el mismo modelo para todos los estados y para todas las diferencias entre años. Sin embargo, el modelo fue mejor para ciertos estados y en algunos, el modelo supera el margen por muy poco. En estos caso, lo más sensato es admitir que no se tienen suficientes datos en vez de saltar a conclusiones dudosas.

7.4 Implicaciones éticas

Se está haciendo universal e insistente el clamor que demanda una nueva actitud ética como la única y urgente solución a los graves problemas del mundo. En casi todos los campos de la actividad humana se está agudizando un estado de riesgo y de cercanía a los límites de tolerancia [49] y la ciencia de datos no es excepción.

Por un lado, los modelos de aprendizaje de máquina son criticados por distorsionar los mercados y desfavorecer a los marginados [50]. Por ejemplo, modelos predictivos usados para contratar de empleados leales tienden a favorecen a los hombres sobre las mujeres y a hombres "blancos" sobre hombres "negros". La razón de esto es que los modelos se construyen con datos históricos y la muestra está sesgada ya que en el pasado ha habido más trabajadores "blancos" hombres que mujeres o personas de otras razas [51].

Como consecuencia, las injusticias y los prejuicios de las decisiones humanas histórí-

cas, se han perpetuado a los modelos. Sin embargo, la diferencia entre los modelos y los decisiones humanas retrogradas o mal-informadas es que el pensamiento humano puede evolucionar ha ser más incluyente, mientras que los modelos solo codifican el pasado.

Este trabajo ofrece una herramienta para la toma de decisiones de política pública y asignación de recursos en programas sociales. Sin embargo, no tiene la verdad absoluta. Las variables de la regresión no implican causalidad, si no correlación. El objetivo es orientar e informar de una manera transparente y clara a un grupo de individuos capaz de tomar decisiones. Se espera que este grupo de individuos utilice la herramienta para el bien y tengan la capacidad de discernir y extraer la información valiosa.

Por un lado, el uso puede ser perverso de tres formas. La primera, si se utiliza la información para privilegiar a las escuelas con mayores oportunidades, aumentando la brecha educativa y social.

La segunda, si no se toman en cuenta los sesgos y la poca información de escuelas de estados como Oaxaca y Chiapas o se considera un modelo completamente justo.

Finalmente, la tercera es si se reduce el problema multidimensional de la educación y el aprendizaje en México a una prueba estandarizada. El trabajo utiliza ENLACE por su gran alcance y como una métrica simple de la educación. Sin embargo, existen otros factores que deben ser analizados a profundidad, incluyendo las tasas de deserción y nivel máximo de estudio de los alumnos.

Por otro lado, tomando en cuenta los sesgos y limitaciones de los modelos, el sistema puede ser utilizado como una herramienta para favorecer el desarrollo de un país más prospero, más justo y más libre.

Este trabajo debe ser utilizado para el bien común, traducido en disminuir la brecha escolar y contribuir al contrato social construido sobre los supuestos de dignidad,

igualdad y libertad de todos los seres humanos [49]. Convirtiéndose, de acuerdo al principio de definición de valor, en un sistema valioso por contribuir al desarrollo la humanidad.

Apéndices

APÉNDICES A

INGENIERÍA DE CARACTERÍSTICAS

Utilizando el formato 911 y CEMABE se crearon las siguientes variables

Nombre de variable	Descripción
admin_personal	Proporción de personal administrativo, auxiliar y de servicios del total de personal
alum_especiales_h	Proporción de alumnos hombres con necesidades educativas especiales
alum_personal_prop	Numero de alumnos por personal escolar
alum_salon	Número de alumnos por salón (CEMABE)
alum1	Número total de alumnos en primero de primaria
alumnos_salon	Número de alumnos por salón (F911)
alums_grupo_cuarto_final	Número de alumnos por grupo en cuarto de primaria al final del ciclo escolar
alums_grupo_cuarto_inicial	Número de alumnos por grupo en cuarto de primaria al inicio del ciclo escolar
alums_grupo_quinto_final	Número de alumnos por grupo en quinto de primaria al final del ciclo escolar
alums_grupo_quinto_inicial	Número de alumnos por grupo en quinto de primaria al inicio del ciclo escolar

alums_grupo_sexto_final	Número de alumnos por grupo en sexto de primaria al final del ciclo escolar
alums_grupo_sexto_inicial	Número de alumnos por grupo en sexto de primaria al inicio del ciclo escolar
alums_grupo_tercero_final	Número de alumnos por grupo en tercero de primaria al final del ciclo escolar
alums_grupo_tercero_inicial	Número de alumnos por grupo en tercero de primaria al inicio del ciclo escolar
alums_maestro_final	Número de alumnos entre número de maestros al final del ciclo escolar
alums_personal_final	Número de alumnos entre número total de personal al final del ciclo escolar
anyo_actual	Último año del periodo del cual se calcula el cambio
cambio_alums_grupo	Cambio entre el número de alumnos por grupo al final y al principio del ciclo escolar
cambio_alums_personal	Cambio entre la proporción de alumnos por personal al final y al principio del ciclo escolar
cambio_matricula	Cambio de matricula entre el inicio y el final del ciclo escolar
cambio_prop_mh	Cambio en la proporción de alumnos mujeres y hombres entre el inicio y el final del ciclo escolar
capacidad_alumnos	Total de alumnos que podrían ser atendidos en el inmueble
cct	Clave del Centro de Trabajo. Identifica a las escuelas.
colegiatura	Costo de colegiatura anual

compu_por_alumnos	Proporción de computadoras por alumnos
compu_sirven	Proporción de las computadoras que sí sirven
costo_esc	Costo de la escuelas separado a la colegiatura
diferencia	Diferencia de años entre el periodo actual y el periodo anterior
DIRSERVREG	Numero de delegación regional a la que pertenece
edo	Entidad federativa en la que se encuentra la escuela
extranjeros_alum_h	Proporción de alumnos extranjeros hombres del total de alumnos hombres
h_inscripcion	Proporción de alumnos hombres inscritos del total de alumnos hombres
horas_arte	Cantidad de horas impartidas a la semana por el personal docente especial de arte en el centro de trabajo
horas_idioma	Cantidad de horas impartidas a la semana por el personal docente especial de idiomas en el centro de trabajo
lavamanos_alum	Número de lavamanos por alumnos
m_inscripcion	Proporción de alumnas mujeres inscritas del total de alumnas mujeres
maestros_especiales	Proporción de personal docente especial del personal docente total
p_mat_anterior	Calificación de matemáticas de la escuela en el periodo anterior
muebles_reparacion	Número de muebles que necesitan reparación

nivelCarrMagis_1V	Proporción de profesores que se encuentran en el programa de carrera magisterial en la primera vertiente (profesores frente a grupo) del total de profesores
normal_maestros	Proporción de personal docente con nivel educativo "Normal"del total de personal docente
oficinas_por_alum	Número de oficinas administrativas por alumnos
padres_consejo	Proporción de padres que forman parte del Consejo Escolar de Participación Social del total de miembros
porc_h_aprobados	Proporción de alumnos hombres aprobados al final del ciclo escolar
porc_h_existencia	Proporción de alumnos hombres que se inscribieron y siguen en la escuela al terminar el ciclo escolar
porc_m_aprobados	Proporción de alumnas mujeres que aprobaron al final del ciclo escolar
porc_m_existencia	Proporción de alumnas mujeres que se inscribieron y siguen en la escuela al terminar el ciclo escolar
porc_ocupacion	Cuanto de la capacidad total de la escuela se está utilizando
porc_tot_aprobados	Proporción de alumnos que aprobaron al final del ciclo escolar
porc_tot_existencia	Proporción de alumnos que se inscribieron y siguen en la escuela al terminar el ciclo escolar
prop_carr_magisterial	Proporción de personal en el programa de carrera magisterial del total de personal

prop_extranjeros_final	Proporción de alumnos extranjeros del total de alumnos al inicio del ciclo escolar
prop_extranjeros_inicial	Proporción de alumnos extranjeros del total de alumnos al final del ciclo escolar
prom_edad_3	Promedio de edad de los alumnos de tercero de primeria
prom_edad_4	Promedio de edad de los alumnos de cuarto de primeria
prom_edad_5	Promedio de edad de los alumnos de quinto de primeria
prom_edad_6	Promedio de edad de los alumnos de sexto de primeria
prom_edad_prim	Promedio de edad de los alumnos de primaria de primeria
prom_preescolar_anyos	Promedio de años de pre-escolar cursados por alumnos de primero de primaria
prom_preescolar_anyos_m	Promedio de años de pre-escolar cursados por alumnas mujeres de primero de primaria
prop_mh_aprobados	
prop_mh_final	Proporción de mujeres a hombres de los alumnos inscritos
prop_mh_inicial	Proporción de alumnos que son repetidores sobre el total de alumnos
prop_mh_inscripcion	Número de hombres inscritos sobre el número de mujeres inscritas
prop_repetidores	Proporción de los alumnos repetidores sobre el total de alumnos
prop_usaer_inicial	Proporción alumnos atendidos por la Unidad de Servicios de Apoyo a la Educación Regular del total de alumnos

semaforo_std	Variable objetivo. Indica si en el periodo seleccionado, bajo el desempeño escolar más de 0.2 desviaciones estándar.
tazas_sanitarias_alum	Número de tazas sanitarias por alumnos
tot_h	Número total de alumnos hombres en la escuela
tot_inscripcion	Número total de alumnos que se inscribieron en la escuela
tot_personal	Número total de personal en la escuela
total_banyos	Número total de baños en la escuela
usaer_alum_h	Proporción de alumnos hombres atendidos por la Unidad de Servicios de Apoyo a la Educación Regular del total de alumnos hombres
ZONAESCOLA	Divisiones geográficas de escuelas

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] A. Unanue, *Minería y análisis de datos: Introducción*, Clase ITAM, 2019.

[3] 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).

[4] 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.

[5] 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).

[6] 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.

[7] 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.

[8] IBM, "Manual CRISP-DM de IBM SPSS Modeler", 2012.

[9] 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.

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

[11] 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.

[12] 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.

[13] R. E. d. Hoyos, J. M. Espino y V. García, "Determinantes del logro escolar en México. Primeros resultados utilizando la prueba ENLACE media superior", 2012.

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

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

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

[17] 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`.

[18] M. tu escuela. (2019). Programas de apoyo, dirección: `http://www.mejoratuescuela.org/mejora/programas` (visitado 18-08-2019).

[19] 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.

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

[21] 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).

[22] 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: Ma-

chine Learning in Python", *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[23] 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).

[24] L. Breiman y col., "Statistical modeling: The two cultures (with comments and a rejoinder by the author)", *Statistical science*, vol. 16, n.º 3, págs. 199-231, 2001.

[25] S. M. Dynarski, "For better learning in college lectures, lay down the laptop and pick up a pen", *Washington, DC: The Brookings Institution, August*, vol. 10, 2017.

[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.plataformadetranspare` `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] 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.

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

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

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

[37] 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).

[38] A. Dubey. (). Feature Selection Using Random forest.

[39] 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.

[40] 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.

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

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

[43] 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.

[44] 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.

[45] 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).

[46] RUser4512. (). Random forest vs extra trees.

[47] G. Tseng. (2019). Gradient Boosting and XGBoost, dirección: https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5 (visitado 21-08-2019).

[48] J. S. Mármol, *Intro to Data Science: Docker*, Clase ITAM, 2019.

[49] C. de la Isla, "De esclavitudes y libertades. Ensayos de ética, educación y política", *Miguel Ángel Porrúa*, pág. 297, 2006.

[50] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[51] M. Bogen y A. Rieke, "HELP WANTED", 2018.