***Modelo de Machine Learning para predecir condiciones riesgo en locales educativos en el Perú***

**Curso: Introducción a Machine Learning para las CCSS y Gestión Pública**

**Integrantes:**

-Calvo, Gabriela / código: F0632095

-Ibáñez, Mauricio / código: F0740669

-Núñez, César / código: F1100839

-Soto, Rodrigo / código: F1302466

**Lima, 05 de junio de 2023**

1. **Introducción**

Un pilar fundamental de la educación en el Perú se centra en el desafío de atender la brecha de infraestructura educativa (BIE), la cual actualmente asciende a S/ 152 mil millones. Por lo tanto, es necesario implementar acciones estratégicas y políticas públicas para abordar esta problemática de manera efectiva, en el más breve plazo. Para esta tarea, es imprescindible asignar recursos adecuados en conjunto con una planificación elaborada por el Gobierno Nacional (GN) a través del Ministerio de Educación (MINEDU), la cual requiere de un enfoque integral que involucre la participación activa de múltiples actores, dentro de los cuales se encuentran los tres (03) niveles de gobierno (Gobierno Nacional, Regional y Local).

En ese contexto, el GN, mediante el MINEDU, en conjunto con la participación de diversos actores, elaboró el Plan Nacional de Infraestructura Educativa al 2025 (PNIE). Este plan presenta una estrategia para el cierre de BIE y constituye uno de los instrumentos centrales de planificación de la infraestructura educativa a largo plazo en el Perú.

Cabe precisar que, según el PNIE de 2017, el cálculo de la BIE asciende a la suma de S/ 100 mil millones, lo que representa un incremento de más del 50% en comparación con la situación actual.

Por otro lado, de acuerdo con la Consulta Amigable del MEF, el gasto correspondiente a Inversiones, en la Genérica de Gasto (GG) 2.6 en la Función: Educación a nivel nacional en el año 2022 asciende a S/ 7 mil millones. Es importante destacar que esta cifra excluye el gasto relacionado con la GG 2.4, dado que, para el caso de las inversiones este está destinada a financiar el control concurrente en el marco de la Ley N° 31358.

Ahora bien, al ritmo de ejecución efectuada en el año 2022, se estima que la BIE se cerraría en 22 años, suponiendo que se mantenga el monto actual de la brecha.

Establecido lo anterior, el PNIE realizó un cálculo de la BIE con el objetivo de dimensionar los recursos monetarios necesarios para reducir el riesgo estructural y garantizar las condiciones mínimas requeridas para que el servicio educativo se preste en una infraestructura adecuada. Actualmente, el Grupo de Prioridad: “1: Riesgo”[[1]](#footnote-1) identifica un total 30,160 Locales Escolares (LLEE), las mismas que requieren atención prioritaria.

Por lo expuesto, la presente propuesta de investigación tiene como objetivo implementar un modelo de Machine Learning que nos ayude a predecir, basándonos en el comportamiento de diversas variables, los posibles LLEE que se añadirían al Grupo de Prioridad “1: Riesgo”, con el fin de obtener información oportuna que nos facilite tomar acciones necesarias en el corto plazo, como de: mantenimiento, acondicionamiento, reforzamiento estructural, entre otros.

Sobre el particular, es importante destacar que, si logramos identificar a tiempo las acciones necesarias, es posible evitar un incremento exponencial en la BIE estimada. Esto nos permitiría generar estrategias más efectivas para beneficiar a un mayor número de LLEE y, por ende, a un mayor número de alumnos.

1. **Revisión de la literatura**

**Estado de la cuestión**

La revisión de literatura consistió en situar la problemática que se aborda en el presente trabajo e identificar información (guías y estudios) que nos permitiera construir las variables, y, finalmente, investigaciones que utilizan el modelo predictivo que proponemos. Al respecto, es importante mencionar que nos enfrentamos a un número reducido de estudios vinculados a predicción en cuanto a la calidad de una infraestructura. No obstante, se abordó algunas investigaciones que se vinculan a nivel de modelo predictivo o de algún tipo de infraestructura.

**Sobre el contexto**

La educación cumple un papel fundamental en el desarrollo sostenible de la sociedad. Una inversión en educación de calidad y acceso equitativo no solo impulsa el crecimiento económico, sino que también fortalece los cimientos de una sociedad justa, inclusiva y próspera. Al respecto, se ha identificado que la infraestructura tiene gran influencia sobre la calidad educativa.

La investigación de Michael Bautista y Nicanor Zagaceta (2019) a través de una revisión bibliográfica sobre el tema identificaron que la infraestructura educativa es uno de los principales factores en la construcción del proceso del aprendizaje a nivel funcional y emocional. Su argumento principal gira en torno a los elementos que configuran la infraestructura educativa que favorece el proceso de aprendizaje, a identificar los espacios escolares que promueven el promueven el aprendizaje en las instituciones educativas públicas y describir de qué manera la infraestructura educativa constituye una fuente de motivación para los estudiantes. Sobre ello, se concluyó que los espacios escolares promueven el aprendizaje, comunicando y estimulando el aprendizaje; motivo por el cual, proponen utilizar los espacios educativos de manera estratégica y dinámica para influir sobre el rendimiento académico.

En ese sentido, el MINEDU presenta la Guía General “Parámetros de Mantenimiento de la Infraestructura Educativa” (2022) con el objetivo de orientar las actividades vinculadas al mantenimiento preventivo y contribuir a la mejora de la calidad del servicio educativo, bajo determinados parámetros que permitan la elaboración del diagnóstico del componente de infraestructura educativa, asegurando las condiciones de funcionalidad, habilidad y seguridad. Esta guía identifica los tipos de mantenimientos, estados de conservación y parámetros, así como las causas del deterioro: (i) Desgaste natural o uso cotidiano, (ii) ausencia de mantenimiento, (iii) accidentes, (iv) uso inadecuado, (v) factores ambientales y (vi) fenómenos naturales. Sobre ello, las causas de deterioro nos permiten identificar las variables independientes para la construcción de nuestro modelo predictivo.

**Sobre el modelo predictivo**

A partir de la problemática identificada, se consideró conveniente trabajar con el modelo Random Forest aplicando LASSO como enfoque para la selección de variables de alta dimensionalidad. Sobre ello Vapnik y Lerner (2019) validan experimentalmente la combinación, donde aplican el método LASSO para identificar un subconjunto de variables que posiblemente sean relevantes. Luego de ello, utilizan Random Forest en el conjunto reducido de variables para refinar aún más la selección y evaluar la importancia de cada variable. El método propuesto se evalúa utilizando estudios de simulación y conjunto de datos del mundo real, al respecto, se descubre que la combinación de ambos modelos tiene un buen rendimiento en cuanto a la precisión de la selección de variables y el rendimiento de predicción, en especial con entornos de alta dimensionalidad. En ese sentido, el estudio reconoce que la selección de variables es fundamental en análisis de datos, pues ayuda a reducir la complejidad del modelo y mejorar la interpretabilidad.

Sobre ello, un artículo que utiliza Random Forest como modelo predictivo de la vida útil de una estructura es el trabajo de Kim et al (2020). La investigación se centra en utilizar el algoritmo Random Forest para predecir la vida útil restante de puentes. Para tal objetivo, los autores recopilaron un conjunto de datos que incluían atributos relacionados con las condiciones de los puentes como: (i) factores estructurales, (ii) ambientales y (iii) de mantenimiento. Estos atributos se usaron como predictores del modelo, asimismo, se realizó una selección de características que permitieran lograr el objetivo, para eliminar variables irrelevantes mejorando la precisión y eficiencia del modelo. Para su desarrollo, se entrenó el modelo utilizando las características seleccionadas y luego se evaluó el modelo utilizando diversas métricas de rendimiento para evaluar la precisión y capacidad predictiva. Como resultado, el modelo obtuvo resultados prometedores en la predicción, superando a otros modelos de regresión tradicionales por su robustez y precisión. Finalmente, los autores explican que la estimación precisa de la vida útil restante de los puentes resulta fundamental para la gestión de infraestructuras y toma de decisiones. Asimismo, el modelo permite evaluar el deterioro y necesidades de mantenimiento de los puentes, optimizando la asignación de recursos y mejorando la planificación general de la infraestructura.

Otro estudio que busca trabajar modelos predictivos vinculados a infraestructuras es el de Guo y Hao (2021). Al respecto, el estudio se orienta a identificar áreas específicas del pavimento de asfalto con riesgo potencial a sufrir algún tipo de impacto. Para ello, se recogió información sobre pavimentos de asfalto existentes, donde se tomaron en cuenta variables como el espesor del pavimento, la condición de las capas que lo componen, el tipo de tráfico, el material y las condiciones climáticas. El estudio consistió en realizar observaciones sobre un carril construido específicamente para el estudio de 40 km de largo y se observó el tránsito durante meses de lluvia y calor para incluir las condiciones climáticas. En total, fueron 20 días de prueba en 4 etapas de 5 días cada una. Se utilizaron 34 variables y 29 820 observaciones para el modelo de entrenamiento. La variable dependiente era locación de asfalto dañado y entre las variables de entrada utilizadas fueron: (i) índice internacional de rugosidad, (ii) profundidad de los hoyos, valor de deflexión, entre otras. Finalmente, los resultados fueron favorables debido a que el modelo fue capaz de predecir con precisión los posibles daños.

En ese sentido, ambos estudios nos permiten identificar las variables relevantes al momento de utilizar un modelo que busca predecir daños, mantenimiento o vida útil de infraestructuras. Si bien no se ha encontrado literatura vinculada a edificaciones específicamente, queda claro la pertinencia de utilizar Random Forest y Lasso para el presente trabajo.

1. **Metodología y Data**

**Fuente de Información**

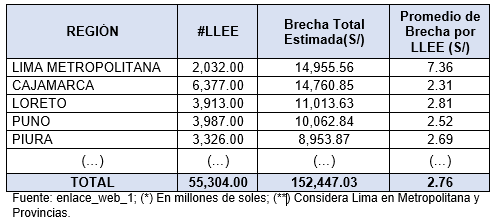
A través de la Resolución Ministerial N° 153-2017-MINEDU, de 06 de marzo de 2017, se aprobó el PNIE al 2025 y se tomó como línea de base para su formulación la información recopilada del Censo de Infraestructura Educativa (2013).

A partir de este Censo, se elaboró un primer diagnóstico del estado de la infraestructura educativa pública. En ese contexto, de  acuerdo a la actualización continua de dicho instrumento, el monto requerido, hasta la fecha, para cerrar la BIE es de S/ 152 mil millones.[[2]](#footnote-2)

En ese sentido, la Dirección de Planificación de Inversiones (DIPLAN) de la Dirección General de Infraestructura Educativa (DIGEIE) del MINEDU, elaboró un Dashboard[[3]](#footnote-3), con la información recopilada para la actualización del instrumento señalado anteriormente. A través de este enlace, se puede extraer la base de datos a nivel LLEE, se adjunta enlace\_web\_1: *https://mineduperu-my.sharepoint.com/:x:/g/personal/diplan15\_minedu\_gob\_pe/EfEe4c9SFPlHlR2xr47F10ABskZMgFUbtBbibde31uttrQ?e=FXr0sk,* el cual, incluye, entre otros, el cálculo de la brecha estimada.

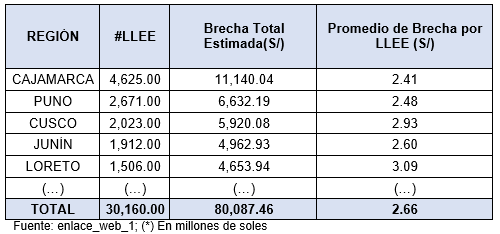
A continuación, se muestra una tabla de la distribución por LLEE y monto de BIE de las principales regiones:

**Cuadro N° 01**

****

Asimismo, de los 30,160 LLEE en situación de “Riesgo”, se muestra una tabla de la distribución por LLEE y monto de BIE de las principales regiones en la citada situación:

**Cuadro N° 02**

****

Con respecto a la información vertida en los Cuadros N° 01 y N° 02, se destaca que el 55% del total de LLEE se encuentran en situación de 'Riesgo', con una brecha total estimada de S/ 80 mil millones.

**Variables**

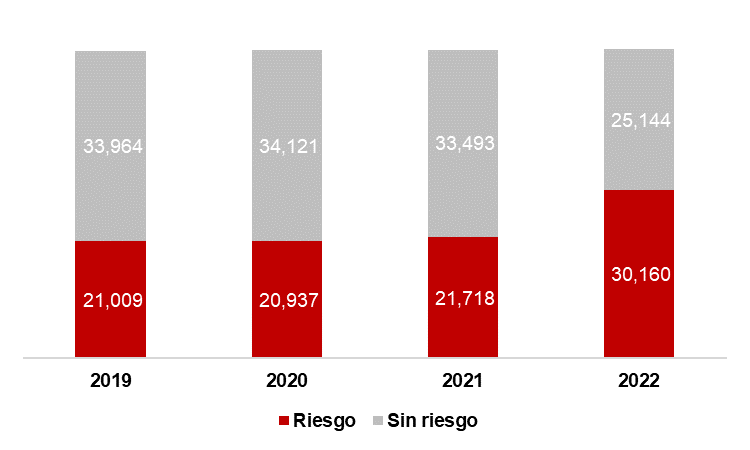
Respecto a las variables de interés del estudio se describe primero la variable dependiente sobre la cual se busca predecir y las variables independientes a partir de las cuales el modelo planteado realizará un aprendizaje. Al respecto, la propuesta de variables se basa tanto en la información disponible así como en la revisión de literatura anteriormente presentada.

La variable a predecir es una dicotómica que toma valor 1 si el LLEE pasará a ser considerado como *“riesgoso”*  y 0 de cualquier otro modo. El Minedu define como LLEE en riesgo si el ratio de demolición es mayor o igual a 70% (es decir, que la infraestructura se encuentra expuesta a un muy alto riesgo de colapso).

Asimismo, el modelo se entrena con un horizonte de rezago de 3 años en consideración a dos motivos. El primero de ellos dado que se evidencian problemas en la calidad de los datos registrados en el periodo 2020 y 2021 principalmente por limitaciones durante la época de pandemia. Segundo, un horizonte de predicción de 3 años permite tomar acciones preventivas para evitar el colapso del LLEE.

El siguiente gráfico muestra la distribución de LLEE según clasificación en riesgo o no entre los años 2019 y 2022. Como se observa, el número de LLEE calificados en riesgo se incrementaron en 9,151en los 3 años de análisis.

**Gráfico N° 1: Número de LLEE por situación de riesgo**



*Fuente: Minedu*

Respecto a las variables independientes se tomaron en cuenta tres grupos de variables: infraestructura del LLEE, servicios del LLEE y contexto a nivel distrital. Las variables independientes se estiman en un periodo t0 sobre un resultado en t1 dado que se busca predecir la vulnerabilidad del LLEE a ser clasificado como “riesgoso”. La metodología de cálculo de las variables se basa en el Informe Técnico: Perú: Estimación de la vulnerabilidad económica a la pobreza monetaria (INEI, 2020)

El primer grupo de variables corresponde a aquellas que describen los ambientes del LLEE, materiales, antigüedad, daños, entre otros. El listado preliminar de variables en este grupo con los que entrenará el modelo son los siguientes:

* propiedad: quien acredita la propiedad del LLEE
* área: área construida del LLEE
* niveles: número de niveles del LLEE
* antigüedad: número de años desde su construcción
* saneado: LLEE se encuentra saneado
* material\_techos: material predominante del techo
* material\_pisos: material predominante del piso
* material\_paredes: material predominante de las paredes
* número\_aulas: número de aulas con las que cuenta el LLEE
* número\_sshh: número de servicios higiénicos en el LLEE
* agua: LLEE cuenta con servicio de agua
* electricidad: LLEE cuenta con servicio de luz eléctrica
* internet: LLEE cuenta con servicio de internet

El segundo grupo de variables contiene información sobre los servicios que se ofrecen en el LLEE y las características del mismo. Las variables propuestas de entrenamiento son las siguientes:

* número\_servicios: número de servicios educativos que se ofrecen en el LLEE
* turnos: si el LLEE atiende mañana, tarde y/o noche
* alumnos: número de alumnos matriculados
* obras: registra actualmente alguna obra en ejecución
* acceso: vías de acceso para llegar al LLEE
* gestión\_riesgos: cuenta con plan de gestión de riesgos

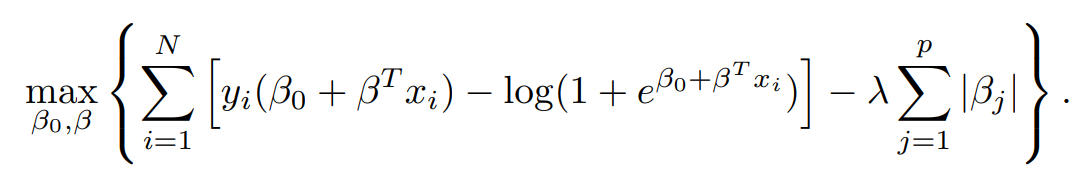
Asimismo, el último grupo hace referencia a variables propias del distrito en el que se encuentra el LLEE. Para ello, se consideran otras bases de datos secundarias como el censo nacional 2017 o el SIRTOD del INEI. La siguiente lista muestra las variables de entrenamiento del modelo:

* rural: indicador si el LLEE se encuentra en un distrito calificado como “rural”
* pobreza: variable continua que señala la incidencia de pobreza en el distrito
* sismo: registro del grado de amenaza sísmica en el distrito
* frontera: señala si el distrito se encuentra en zona de frontera
* vraem: indica si el distrito se encuentra en zona vraem
* heladas: indica si el distrito registra heladas
* FEN: indica si el distrito ha sido afectado por el fenómeno del niño

En total se propone un conjunto de 26 variables independientes. No obstante, ese número se incrementa con los procesos de dicotomización de las variables categóricas y luego de generar interacciones.

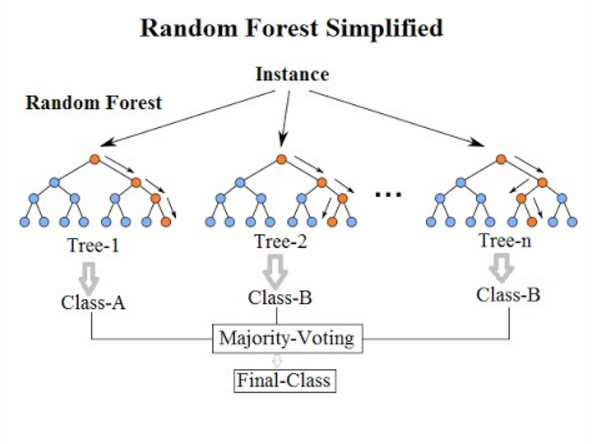
**Método de aprendizaje estadístico**

La metodología de aprendizaje estadístico que proponemos utilizar para predecir los cambios en el estado de riesgo de los LLEE en el Perú comprende la estimación de dos modelos predictivos. Por un lado, se plantea estimar un modelo LASSO (por sus siglas en inglés, *Least absolute shrinkage and selection operator*) Logístico. Así, a partir de las variables descritas en la sección anterior, se espera identificar aquellos predictores que mayor pérdida de error cuadrático medio (MSE, por sus siglas en inglés mean squared error) aporten en la estimación del modelo. Esto se obtiene por la particularidad del modelo LASSO, el cual penaliza la complejidad del modelo (es decir, el número de predictores e interacciones incluidas en el modelo de predicción), atribuyendo un valor 0 a los coeficientes asociados a los predictores de menor importancia o que menor pérdida de MSE aporten al modelo. Esta característica, no solo brinda una mayor interpretabilidad de los predictores (i.e. cuáles son los más relevantes); sino también permite evitar el sobreajuste del modelo y mejorar la capacidad de generalización de este. En particular, el modelo a estimar, siguiendo a Friedman, Hastie & Tibshirani (2008), queda como el descrito a continuación:



Donde, nuestra variable objetivo (yi) es una variable dicotómica que toma valor 1 si el LLEE pasará a ser considerado como *“riesgoso”*  y 0 de cualquier otro modo. Los predictores están representados por xi que incluyen características del LLEE (como el tipo de propiedad, antigüedad, materiales predominantes en la construcción, entre otros), variables de contexto (tales como la ruralidad, el nivel de pobreza del distrito de ubicación del LLEE, y otras variables características del distrito) y variables relacionadas a los servicios que brinda el LLEE (como la cantidad de servicios educativos ofrecidos, turnos de cada servicio, número de alumnos, entre otros). El término de regularización está representado por lambda, el cual permite determinar la complejidad del modelo y ajustar ciertos coeficientes asociados a los predictores con el valor de cero.

Por otro lado, se plantea estimar un Random Forest de clasificación para nuestra variable objetivo la cual, como se describió en la sección anterior, es dicotómica. La lógica detrás de este modelo es estimar distintos árboles de clasificación con muestras obtenidas a partir del *bootstrap* y en cada uno de dichos árboles seleccionar un subconjunto de predictores distintos. Esta estrategia permite mejorar el nivel de predicción del modelo al combinar distintos árboles estimados a partir de diferentes conjuntos de predictores. De manera gráfica, el modelo de predicción de un Random Forest de clasificación se puede resumir o simplificar como en el siguiente gráfico:



1. **Bibliografía**

Bautista, M. S., & Zagaceta, N. (2019). *Rol de la infraestructura educativa en el proceso de Aprendizaje de estudiantes en instituciones educativas públicas.* Lima: PUCP.

Guo, X., & Hao, P. (2021). Using a Random Forest Model to Predict the Location of Potential Damage on Asphalt Pavement. *Applied Sciences*.

Kim, K. H., Nam, M., Hwang, H. H., & Ann, K. Y. (2020). Prediction of Remaining Life for Bridge Decks Considering Deterioration Factors and Propose of Prioritization Process for Bridge Deck Maintenance. *Sustainability*.

MINEDU. (2020). *Guía General “Parámetros de Mantenimiento de la Infraestructura Educativa.* Lima: Ministerio de Educación.

1. Definición de Grupo 1 - Riesgo: Considerando el ratio de demolición, se agrupan los locales en los que el ratio de demolición es mayor o igual a 70% (es decir, que la infraestructura se encuentra expuesta a un muy alto riesgo de colapso). [↑](#footnote-ref-1)
2. De acuerdo a lo señalado en el PNIE al 2025, se entiende por “brecha” al monto consolidado de recursos financieros necesarios para mejorar la condición y ampliar capacidad de la infraestructura existente [↑](#footnote-ref-2)
3. Para mayor detalle, dirigirse al siguiente enlace *https://app.powerbi.com/view?r=eyJrIjoiZWM0NWI0OWQtYjNhYS00YzU2LTljZmEtZTI1ZGVmMzFkMWMyIiwidCI6IjE3OWJkZGE4LWQ5NjQtNDNmZi1hZDNiLTY3NDE4NmEyZmEyOCIsImMiOjR9&pageName=ReportSection25221d34e718ca140075* [↑](#footnote-ref-3)