PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

Mejora de Procesos de Ingeniería mediante Aprendizaje Automático: Un Enfoque en la Fabricación de Antenas y la Colaboración en Equipos de Trabajo en Obras Civiles

Leonardo González Catalán

Profesor Guía: **Leslie Pérez Cáceres**
Profesor Correferente: **Emanuel Vega Mena**
Asignatura: **ICI6541 - Proyecto de Titulo**
Carrera: **Ingeniería Civil en Informática**

Junio 2025

Dedico esta sección a todas las personas que han sido parte de mi proceso formativo, con especial reconocimiento al profe Claudio Cubillos, quien ha sido mi modelo a seguir en el ámbito académico y un pilar fundamental en mi formación, no solo en lo académico, sino también en lo personal. Él confió en mis capacidades en momentos en que ni siquiera yo creía en ellas, y siempre me ha motivado a seguir creciendo.

También me gustaría agradecer al profe Nibaldo Rodríguez, quien ha sido un gran amigo, brindándome consejos siempre que los he necesitado. Por supuesto, también ha sido un guía en mi camino de aprendizaje académico, enseñándome el rigor y la seriedad necesarios para afrontar problemas de ingeniería, y apoyándome en el aprendizaje de las matemáticas, tan necesarias en esta área.

Por otro lado, quisiera agradecer a mis amigos, que me acompañaron a lo largo de todo este trayecto y seguirán ahí sin importar la distancia. Y a mi familia, que si bien no fue parte del proceso académico en sí, de una u otra forma, siempre estuvo presente.

Agradezco también a la profesora Leslie Pérez, quien me guió a lo largo de toda la elaboración de esta tesis y en el desarrollo mismo de la investigación. Por su paciencia, y por darme múltiples herramientas que de otra forma no habría obtenido, como lo son el diseño experimental, las habilidades de escritura, el orden, la constancia y la perseverancia.

Finalmente, ya que no puedo nombrarlos a todos, agradezco a quienes participaron en los proyectos en los que me integré para el desarrollo de esta tesis. Y en general, gracias a todos los que me han acompañado en este largo viaje llamado universidad (¡el cual está lejos de terminar!).

「道草を楽しめ 大いにな。ほしいものより大切なものが きっとそっちに ころがってる」

Resumen

Este trabajo tiene como objetivo, por un lado, mejorar la caracterización y el proceso de manufactura de antenas fabricadas con filamento PLA-$BaTiO_3$, utilizando la impresora 3D desarrollada por el laboratorio de materiales de la Escuela de Ingeniería Química de la PUCV. Se busca optimizar el acabado de las impresiones 3D para lograr una fabricación más rentable sin comprometer las propiedades dieléctricas ni la calidad estructural de las antenas. Para lograr estos objetivos, se utilizarán algoritmos de aprendizaje automático en forma de "clasificación", aplicando también técnicas de selección de características para identificar los parámetros de mayor impacto.

Por otro lado, este trabajo también aborda una problemática crítica en proyectos de arquitectura, ingeniería y construcción (AEC): la calidad de la comunicación y colaboración entre equipos multidisciplinarios. Estos equipos suelen estar altamente fragmentados, lo que puede generar problemas como cuellos de botella en la información, baja cohesión y pérdida de eficiencia, impactando negativamente en los plazos, costos y resultados del proyecto.

Para enfrentar esta problemática, se propone el uso de algoritmos de aprendizaje automático en conjunto con análisis de redes sociales (SNA), con el fin de predecir métricas estructurales clave como el número de comunidades, personas aisladas, cuellos de botella, grado medio, longitud media de caminos y grado de jerarquización de la red. Estas métricas permiten identificar dinámicas ineficientes y proponer estrategias de mejora en la gestión y comunicación de los equipos.

En resumen, este trabajo busca aportar significativamente a la mejora de procesos tanto en entornos materiales como organizacionales dentro de la ingeniería, utilizando aprendizaje automático, análisis de datos y selección de características como herramientas transversales de optimización.

Palabras Clave: Aprendizaje Automático, Impresión 3D, Antenas, PLA-$BaTiO_3$, Dinámicas de Equipo, Análisis de Redes Sociales, Proyectos AEC.

Abstract

This research aims to improve the characterization and manufacturing process of antennas made from PLA-$BaTiO_3$ filament, using the 3D printer developed by the Materials Laboratory of the School of Chemical Engineering at PUCV. The objective is to optimize the surface finish of 3D-printed parts to enable more cost-effective fabrication without compromising the dielectric properties or structural quality of the antennas. To achieve this, machine learning algorithms for classification will be applied, alongside feature selection techniques to identify the most influential parameters.

Additionally, this work addresses a critical issue in architecture, engineering, and construction (AEC) projects: the effectiveness of communication and collaboration among multidisciplinary teams. These teams are often highly fragmented, which can lead to problems such as bottlenecks, poor cohesion, and efficiency losses, negatively impacting project timelines, costs, and outcomes.

To address this problem, we propose the use of machine learning algorithms along with social network analysis (SNA) to predict key structural metrics such as the community count, isolated individuals count, bottlenecks, average degree, average path length and hierarchization degree. These metrics allow us to identify inefficient dynamics and propose strategies to improve team management and communication.

In summary, this work seeks to contribute significantly to process improvement in both material and organizational contexts within engineering, using machine learning, data analysis, and feature selection as transversal optimization tools.

Keywords: Machine Learning, 3D Printing, Antennas, PLA-$BaTiO_3$, Team Dynamics, Social Network Analysis, AEC Projects.

Índice

Índice de figuras

1. Introducción

Un problema recurrente en los países en vías de desarrollo es la falta de desarrollo tecnológico. Esto se debe en gran medida a los altos costos iniciales de la maquinaria de manufactura, lo que a su vez afecta fuertemente la investigación aplicada en estos países. Una de las áreas más afectadas debido a esto es el área de las telecomunicaciones, donde prototipar objetos de alta complejidad, como algunas antenas, tiene un alto costo. Muchas veces, no se puede permitir el lujo de hacer múltiples prototipos para así ir depurando de forma iterativa. Frente a esto, nace la necesidad de generar prototipos de menor costo que sean fácilmente fabricables. En este aspecto, las impresoras 3D basadas en extrusión de termoplásticos juegan un gran rol, ya que permiten una fabricación a bajo costo y fácil de ejecutar. No obstante, los polímeros por sí solos no son suficientemente aptos para la fabricación de antenas, principalmente debido a su baja permitividad.

Por otro lado, otro problema a nivel mundial, y en mayor medida en los países en vías de desarrollo, es lograr una comunicación y colaboración efectiva en los proyectos de arquitectura, ingeniería y construcción, los cuales suelen estar altamente fragmentados debido a la presencia de gran cantidad de profesionales de distintas áreas. Una comunicación deficiente puede llevar a errores, insatisfacción por parte de los trabajadores, y por consiguiente, a una colaboración deficiente, la cual puede resultar en la duplicación de esfuerzos, un bajo rendimiento y la falta de cohesión del equipo. Lo cual en países en vías de desarrollo es problemático ya que finalmente se traduce en mayores costos y peores resultados. Es por esto, que nace la necesidad de entender mejor las interacciones personales en equipos de trabajo, y con esto, determinar qué prácticas, metodologías, y/o tecnologías aportan en mayor medida a una mejor comunicación.

A razón de todo lo expuesto anteriormente, se han planteado distintas soluciones a estos problemas, por un lado, en el ámbito de la fabricación de antenas, se creó un compuesto dieléctrico (PLA-BaTiO3) con el cual se pueden imprimir antenas a un bajo costo, y fácilmente, y por otro lado, con respecto a la comunicación y colaboración en proyectos de arquitectura, ingeniería, y construcción, se han utilizado técnicas de análisis de redes sociales para entender el flujo de la información en los equipos de trabajo, generando un grafo o red social que permite visualizar las interacciones y además analizar patrones de interacción que pueden surgir en equipos. Pese a esto, las soluciones existentes requieren de personas con amplia experiencia y expertiz en sus áreas respectivas, además de ejecutarse de forma manual y ser propensas a errores debido a esto mismo. Comúnmente en el caso de las impresiones 3D, se requieren múltiples iteraciones para lograr un buen acabado en las piezas, el cual permita que estas sean utilizadas. En el caso de la generación de los grafos, este es un proceso que toma bastante tiempo, ya que primero se generan plantillas de redes sociales ideales, con las cuales luego se comparan las redes generadas en base a los equipos de trabajo y se analizan las interacciones.

Es por todo esto que, en este proyecto, se busca apoyar a ambas áreas, en base a datos del mundo real, y algoritmos de aprendizaje automático. Esto se hará, en el caso de los componentes para antenas, prediciendo el acabado de estos basándose en la configuración de impresión de la pieza. Por otra parte, en el caso del análisis de colaboración y comunicación en equipos de trabajo, se utilizarán los datos que solían ser utilizados para generar un grafo de forma manual, y con ello hacer análisis de redes sociales, para ahora estimar directamente las métricas de interacción, mediante algoritmos de aprendizaje automático, y a su vez, determinar qué variables afectan en mayor medida a estas métricas con estos mismos. Para ambos problemas, la técnica de aprendizaje automático a utilizar inicialmente, será un algoritmo de tipo Random Forest (Aunque también se comparará contra modelos clásicos, y una red neuronal superficial para ver que modelo es el mejor en cada caso) con el cual se hará una estimación, en el caso de las antenas, de como será el acabado de la pieza a imprimir, y en el caso del análisis de equipos de trabajo, se estimarán las métricas de N° de Comunidades, N° de Personas Aisladas, N° de Cuellos de botella, Grado Medio de la red, Longitud Media de los caminos, y Grado de Jerarquización de la red. Con el apoyo brindado, se espera aportar significativamente a ambos procesos, mejorando la eficiencia y reduciendo el costo asociado a estos.

2. Objetivos

2.1. Objetivo General

Desarrollar herramientas basadas en inteligencia computacional que permitan predecir resultados de procesos en el ámbito de la ingeniería, con especial énfasis en los procesos de fabricación de antenas y la colaboración en proyectos de ingeniería civil.

2.2. Objetivos Específicos

- Comprender las variables asociadas a los casos de estudio (impresión 3D y colaboración en proyectos de ingeniería civil) y reconocer tareas de predicción asociadas a estos.
- Construir, limpiar y estandarizar un set de datos de los casos de estudio (impresión 3D y colaboración en proyectos de ingeniería civil) para el entrenamiento de modelos predictivos.
- Estudiar y seleccionar modelos predictivos y técnicas asociadas para aplicarlos a los casos de estudio (impresión 3D y colaboración en proyectos de ingeniería civil).
- Entrenar y evaluar modelos predictivos en base al análisis realizado y los datos generados.

3. Marco Teórico

Los avances en inteligencia computacional y aprendizaje automático han permitido su aplicación transversal en múltiples áreas de la ingeniería. Este proyecto aborda dos problemáticas relevantes en contextos distintos: la mejora del diseño y fabricación de antenas mediante impresión 3D (en el ámbito de la ingeniería química), y la optimización de la comunicación en equipos de trabajo en obras civiles (en el ámbito de la ingeniería civil). En ambas líneas de trabajo, se hace uso de técnicas de aprendizaje automático para modelar procesos complejos, predecir métricas clave, y proponer soluciones más eficientes. A continuación, se presentan los conceptos fundamentales que sustentan cada una de estas problemáticas.

3.1. Conceptos Químicos y de Manufactura Relevantes

La **Manufactura Aditiva (AM)** es un proceso de fabricación de objetos tridimensionales mediante la adición sucesiva de capas de un material (de ahí su nombre "manufactura aditiva"). Una técnica de manufactura aditiva es la ***impresión 3D***, siendo esta relevante ya que es la técnica utilizada para crear las piezas de ***PLA-BaTiO3***. Para nuestro caso de estudio, la ***impresora*** utilizada es una ***impresora 3D basada en la extrusión de termoplásticos***, que como dice su nombre, se basa en el fundido y posterior extrusión (salida de material por un cabezal o extrusor), de un material termoplástico (en nuestro caso de estudio, ***PLA***, que actua como ***carrier*** del compuesto dielectrico de interés).

Al fabricar piezas multi-material mediante ***impresión 3D*** se utilizan materiales llamados ***Carrier***, que actúan como medio auxiliar para otros materiales que dadas sus características podrían ser difíciles de manipular, ya sea porque acortan la vida útil de las impresoras, dejan residuos difíciles de limpiar, o simplemente son difíciles de manipular en estado puro. Estos ***Carrier*** suelen ser polímeros debido a su fácil manipulación, bajo costo y facilidad de eliminación de este mismo. El proceso de eliminación del ***Carrier*** para así obtener una pieza del material puro se conoce como ***Debinding*** y puede realizarse tanto de forma física (separando el polímero del compuesto dieléctrico), o bien de forma química (eliminando el polímero mediante degradación térmica, o en palabras mas simples para dar contexto, "quemándo"). Una vez eliminado el polímero, se requiere dar un tratamiento térmico al material *"puro"*, este proceso se llama **Sinterización**, y es de suma importancia, ya que maximiza las propiedades térmicas, eléctricas y mecánicas del material, cambiando el tamaño de la impresión, y la distribución de las partículas. En este tratamiento el material se calienta a temperaturas cercanas a su punto de fusión, esto con el objetivo de eliminar la mayor porosidad posible del material, y por consiguiente, aumentar su densidad en la pieza, además de brindar todas las características previamente descritas.

En resumen, el ***PLA-BaTiO3*** es un multi-material compuesto por ácido poliláctico (***PLA***) y titanato de bario (BaTiO3). El ***PLA*** actúa como *carrier* para el titanato de bario, gracias a su propiedad termoplástica, que le permite depositarse en conjunto al titanato mediante su fusión aproximadamente a los 240 °C. El ***BaTiO3*** por otro lado, actúa como cerámica piezoeléctrica, dicho de otra manera, cambia su estructura al someterle a un campo eléctrico, y en el caso de la fabricación de antenas, es de interés debido a que es aislante de ondas de alta frecuencia debido a su alta ***constante dieléctrica***. El ***BaTiO3*** resultante del ***debinding*** realizado al ***PLA-BaTiO3*** puede ser sinterizado a 1000°C gracias a un aditivo utilizado al generar los pellets del ***PLA-BaTiO3***, ya que el ***BaTiO3*** por si solo, se sinteriza a 1200°C. Por otro lado, el ***PLA*** se degrada a 320°C. Estas propiedades son importantes para el post-procesamiento de las piezas impresas.

3.2. Conceptos de Comunicación y Colaboración Relevantes en Proyectos de Ingeniería Civil

Los proyectos ***AEC*** (***arquitectura, ingeniería y construcción***) se caracterizan por ser altamente multidisciplinarios, ya que requieren la colaboración de profesionales de diversas áreas para llevarse a cabo de manera exitosa. Esta diversidad, si bien es una fortaleza en términos de especialización, también representa un desafío: la comunicación entre profesionales de distintas disciplinas no siempre ocurre de forma fluida, lo que puede derivar en múltiples problemas durante la ejecución de los proyectos, tales como retrasos, errores, aumento de costos e insatisfacción por parte de los trabajadores.

Para mitigar estos problemas, se han desarrollado diversos enfoques y metodologías colaborativas, siendo exitosos durante años ***Building Information Modeling (BIM)*** y ***Lean Construction*** [8]. Ambos frameworks buscan mejorar la coordinación y la toma de decisiones dentro de los equipos, fomentando una cultura de colaboración e integración desde las etapas tempranas del proyecto.

Sin embargo, incluso al aplicar estos enfoques, la alta fragmentación de los equipos puede seguir siendo un obstáculo. Por ello, surge la necesidad de abordar los proyectos ***AEC*** desde una perspectiva más analítica, con el fin de entender mejor las dinámicas de interacción entre los participantes y así identificar oportunidades de mejora.

En este contexto, el ***Análisis de Redes Sociales (SNA, por sus siglas en inglés)*** se presenta como una herramienta poderosa [1]. Esta técnica permite modelar las interacciones entre personas como un grafo, donde los nodos representan individuos (miembros del equipo, especialistas, ***stakeholders***, etc.), y los enlaces representan relaciones o flujos de información. A partir de esta representación, es posible calcular métricas que reflejan el grado de cohesión, eficiencia y centralidad de la red de trabajo.

Las métricas más utilizadas en este tipo de análisis incluyen:

- **Número de comunidades**: indica el nivel de fragmentación de la red. Un número elevado puede reflejar silos de información o una débil integración entre subgrupos dentro del equipo.
- **Número de cuellos de botella**: identifica nodos clave cuya eliminación (o ausencia) dividiría la red en componentes desconectados. En este contexto, un cuello de botella suele ser una persona que actúa como puente entre equipos distintos; si esta persona no está disponible, la comunicación entre esos equipos podría verse severamente afectada.
- **Grado medio**: representa el promedio de conexiones por nodo, es decir, cuántas personas, en promedio, se comunican directamente con cada miembro del equipo.
- **Longitud media de los caminos**: mide la cantidad promedio de nodos que debe atravesar la información para llegar de un punto A a un punto B. En otras palabras, estima cuántas personas, en promedio, median la comunicación entre dos individuos cualquiera dentro del proyecto.
- **Número de nodos aislados**: refleja la cantidad de personas que no presentan conexiones con otros miembros del equipo, lo que puede ser un indicio de falta de inclusión o problemas de integración.

Estas métricas permiten diagnosticar de forma cuantitativa el estado de la colaboración en un proyecto, y son la base sobre la cual este trabajo propone entrenar modelos de aprendizaje automático que permitan predecir y mejorar estas dinámicas de forma automatizada.

3.3. Modelos Propuestos

3.3.1. Regresión Lineal

Modelo predictivo simple y ampliamente utilizado en problemas de regresión. Su objetivo es modelar la relación entre una variable dependiente y y una o más variables independientes $\mathbf{x} = [x_1, x_2, ..., x_n]$ mediante una función lineal [13]. Esta técnica se basa en la hipótesis de que existe una relación lineal entre las variables, y busca ajustar los coeficientes que mejor representen dicha relación.

Modelo Matemático

El modelo general de regresión lineal múltiple se expresa como:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{1}$$

donde $\hat{y}$ es el valor estimado de la variable dependiente, x_i son las variables independientes, β_i es la pendiente asociada a cada variable, β_0 es el término intercepto.

Entrenamiento del Modelo

Consiste en encontrar los coeficientes β_i que minimicen la diferencia (o error) entre las predicciones del modelo y los valores reales observados. El modelo se puede resolver analíticamente utilizando el método de los mínimos cuadrados ordinarios (OLS), o numéricamente mediante métodos iterativos como el descenso por gradiente.

Supuestos y Limitaciones

Para que el uso de un modelo de regresión lineal sea válido y útil, deben cumplirse las siguientes condiciones:

- **Linealidad:** La relación entre las variables debe ser lineal.
- **Independencia:** Las observaciones deben ser independientes entre sí.
- **Homoscedasticidad:** La varianza del error debe ser constante.
- **Normalidad del error:** Los errores deben seguir una distribución normal.
- **No multicolinealidad:** Las variables independientes no deben estar altamente correlacionadas entre sí.

El cumplimiento de estas condiciones se puede verificar mediante análisis estadísticos como la inflación de varianza (VIF), gráficos de dispersión del error y mapas de correlación entre variables.

Aplicación en esta Investigación

En este estudio, la regresión lineal fue utilizada con el fin de compararla contra técnicas mas complejas como lo son otros tipos de regresión, Random Forest y Redes Neuronales. A pesar de su simplicidad, mostró un rendimiento competitivo en varias métricas en comparación con modelos más complejos como Random Forest y redes neuronales, especialmente cuando las relaciones entre variables eran predominantemente lineales. Además, su interpretabilidad facilita el análisis del peso relativo de cada variable predictora.

3.3.2. Árbol de Decisión

Algoritmos de aprendizaje supervisado utilizados en la toma de decisiones. Se basan en la subdivisión de una decisión compleja en decisiones más sencillas, que se pueden representar de forma binaria (sí o no, existe o no existe), de forma jerárquica. La subdivisión de las decisiones (**Nodos del Árbol**) se llama **Split**. El mecanismo de Split funciona de la siguiente forma [6]:

- **En caso de tener un arbol de clasificación:** Primero, se obtienen los datos completos. Luego, se calcula la entropía de cada variable predictora. Una vez calculadas las entropías, se determina la ganancia de información de cada variable. El corte o Split se realiza en la variable con la mayor ganancia de información. En otras palabras, se hace el corte en la variable que disminuye en mayor medida la incertidumbre. A continuación se adjuntan la formula de la entropía y la ganancia de información.

$$H(X) = -\sum_{i=1}^{n} P_i \log_2(P_i) \tag{2}$$

La ecuación 2 muestra la entropía ($H(X)$) donde P_i es la probabilidad de la variable i y n es el numero de variables.

$$IG = H(X_{inicial}) - \sum_{i=1}^{n} W_i \cdot H(X_i) \tag{3}$$

La ecuación 3 muestra la ganancia de información (IG) donde $H(X_{inicial})$ es la entropía de los datos iniciales, W_i es el peso del split respecto al tamaño inicial de la muestra y $H(X_i)$ es la entropía de los datos en el i-esimo split.

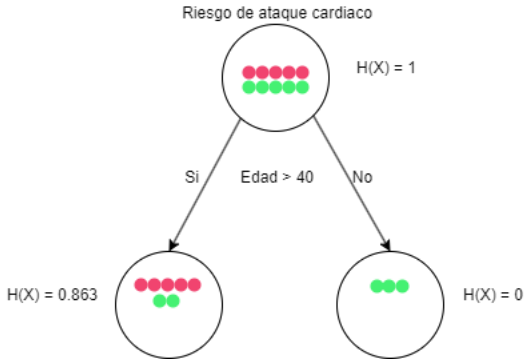

Figura 1: Imagen a modo de ejemplo de un de split de clasificación.

- **En caso de tener un árbol de regresión:** Primero, se tienen los datos completos. Luego, se calcula la media de cada par de puntos consecutivos en una variable predictora, y estos valores serán los posibles puntos de corte para esa variable. Con cada punto de corte en la variable predictora, se genera un punto de corte en la variable objetivo, el cual será la media de todas las respuestas correspondientes a los valores de la variable predictora que están por debajo del punto de corte. Para seleccionar el punto de corte óptimo, se calcula el error cuadrático medio de las predicciones para cada candidato a punto de corte. Este proceso se repite para cada variable predictora, y el corte o 'split' se realiza en la variable cuyo candidato a punto de corte tenga el menor error. A continuación se adjunta la formula del error cuadrático medio (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4}$$

Donde Y_i es el valor objetivo real, e $\hat{Y}_i$ es el valor predicho.

Con cada **Split** que se hace, nacen nuevas **Ramas**, y se agrega profundidad al modelo. A medida que se va profundizando en el árbol, se disminuye el espacio de soluciones, y en el caso ideal, se converge a una única solución. Sin embargo, no siempre se llega a una única solución. A las posibles soluciones se les llama **Nodos Hoja**, o **Nodos Terminales**. En los **Nodos Hoja** se realiza el cálculo de la salida.

En resumen, el algoritmo funciona de manera tal que, dado un conjunto de datos inicial, se "entrena" o "ajusta" el modelo al problema, esto significa, que se realizan los **Split** basados en el conjunto de datos de entrada, y se forma el árbol con tal de maximizar el criterio de **Clasificación** o **Regresión**. Luego de este proceso, se puede ingresar nuevos datos para obtener la predicción para estos dado el entrenamiento previamente realizado.

Parámetros de un Árbol de Decisión

- **Máxima Profundidad:** Cantidad máxima de splits que se pueden generar. De no fijarse una cantidad máxima, se hacen splits hasta que todos los nodos son hojas, o bien hasta llegar al mínimo de variables por split en caso de estar fijado.
- **Criterio de Calidad:** Función con la cual se seleccionan los cortes óptimos. En el caso de arboles de clasificación puede ser la entropía, ganancia de información o el índice de Gini. En el caso de los arboles de regresión puede ser el error cuadrático medio, el error cuadrático, o el error absoluto.
- **Número mínimo de variables por Split:** Cantidad mínima de variables para hacer un split. Su valor por defecto es 2 y es el valor mínimo que puede tomar debido a que con menos de 2 variables no se puede hacer split.
- **Número mínimo de variables por nodo hoja:** Numero mínimo de variables para considerar un nodo como hoja, si un split genera nodos con un numero de variables menor al fijado, ese corte se desecha. Su valor por defecto es 1.

3.3.3. Random Forest

Algoritmos basados en el ensamblaje de múltiples **Árboles de decisión**, de los cuales cada uno posee una muestra aleatoria de los datos involucrados en la toma de decisión (de ahí su nombre). Al ser un ensamblaje de **Árboles de decisión**, gran parte de sus características son heredadas de estos. Dentro de las características de **Random Forest** se encuentran, la posibilidad de abordar tanto problemas de **Regresión** como de **Clasificación** definiendo el tipo de árbol que se usará para el problema abordado [4], por otro lado, su estructura le permite trabajar con datos incompletos o ruidosos, lo que lo hace robusto frente a diferentes tipos de datos [5], finalmente, su estructura es jerárquica, lo que permite estudiar la importancia de las variables involucradas en el problema [2]. La ventaja que ofrece frente al uso de un único **Árbol de decisión** es que tiene menor **overfitting** al tener múltiples arboles trabajando con distintas porciones de los datos, lo cual hace que sea un modelo más generalizable.

Respecto al **overfitting**, el sobre-ajuste de un modelo de toma de decisiones, **Clasificación** o **Regresión**, respecto a sus datos de entrenamiento. Es algo que se busca evitar, ya que da una falsa impresión de que el modelo es robusto, sin embargo a la hora de trabajar con nuevos datos no generaliza bien debido a lo mismo.

Respecto a los conjuntos de datos para **Random Forest**, estos consisten en N datos, donde estos tienen variables de predicción (entrada al modelo) y variables a predecir (salida del modelo). Según el tipo de salida del modelo, se dice que este puede ser de predicción o de **Regresión**, siendo de **Clasificación** si su salida es una variable categórica, y de **Regresión** si su salida es una variable continua.

Respecto al funcionamiento del algoritmo, dado un conjunto de datos inicial, se "entrena" o "ajusta" el modelo al problema, esto significa, que se realizan los **Split** basados en el conjunto de datos de entrada, y se forman los árboles con tal de maximizar el criterio de **Clasificación** o **Regresión**. Luego de este proceso, se puede ingresar nuevos datos para obtener la predicción para estos dado el entrenamiento previamente realizado. En el caso de **Random Forest** para regresión, la predicción

corresponde al promedio de predicciones en los arboles y en el caso de **Random Forest** para **Clasificación** la predicción corresponde al resultado del **Majority Voting** de los árboles, lo que significa que se selecciona la variable que se repite mas veces como solución.

Parámetros de un Modelo Random Forest

Para instanciar un modelo de tipo Random Forest es necesario ajustar ciertos parámetros, estos suelen ser:

1. **Numero de árboles o estimadores del modelo (N):** Cantidad de arboles del modelo, mientras mayor sea su valor, menor es la varianza de las predicciones. El aumentar de forma innecesaria el valor de este parámetro aumenta de forma lineal la complejidad computacional, sin aportar mejoras significativas a la predicción [7].

2. **Bootstrap:** Parámetro booleano que indica si el modelo de tipo Random Forest utiliza Bootstrapping o también llamado Bagging. El Bootstrapping es el muestreo aleatorio del conjunto de datos para la conformación de los arboles. De no utilizarse el Bootstrapping, los arboles se conforman con todo el conjunto de datos.

    a) **Máximo de muestras:** De utilizarse Bootstrapping, se debe especificar un numero máximo de muestras para cada árbol. El valor de este valor es un numero flotante entre 0 y 1, ya que este valor se multiplica por la cantidad de muestras del conjunto de datos original.

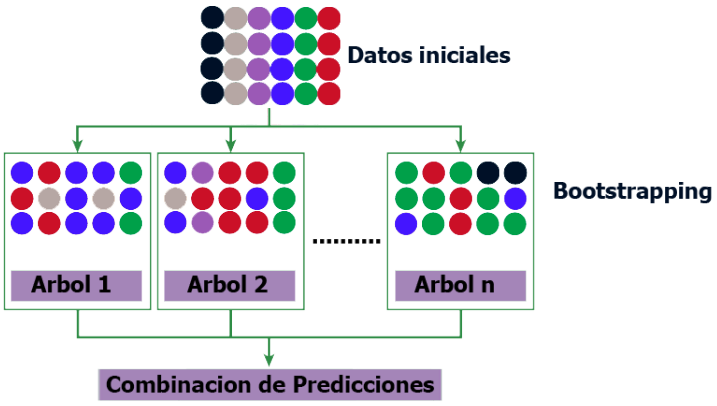

Figura 2: Imagen a modo de ejemplo de un modelo de tipo Random Forest con bootstrapping.

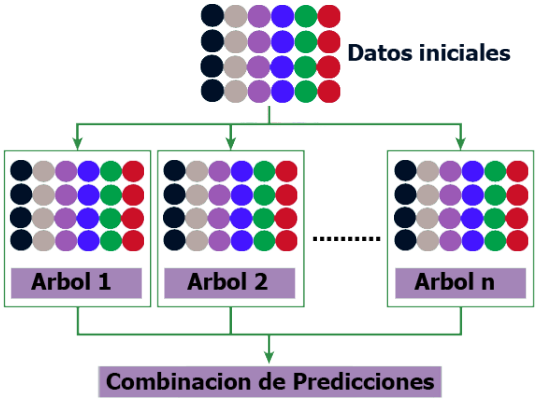

Figura 3: Imagen a modo de ejemplo de un modelo de tipo Random Forest sin bootstrapping.

3.3.4. Redes Neuronales Artificiales

Modelos de aprendizaje automático diseñados para reconocer patrones complejos y relaciones no lineales entre variables. Dentro de estos modelos existen muchas topologías de red, sin embargo, por efectos prácticos y de relevancia en la investigación, se explicaran en detalle las **redes neuronales superficiales** o *shallow neural networks*.

Perceptrón monocapa

El perceptrón monocapa es la unidad básica de las redes neuronales artificiales. Este modelo fue propuesto por Rosenblatt en 1958 [16]. Consiste en un conjunto de entradas $\mathbf{x} = [x_1, x_2, ..., x_n]$ que se combinan linealmente con un vector de pesos $\mathbf{w} = [w_1, w_2, ..., w_n]$ y un sesgo b. La salida del modelo se define como:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{5}$$

donde f es una función de activación, la cual típicamente es no lineal, y que su finalidad es dar la capacidad de modelar relaciones complejas. Algunas funciones de activación mas usadas son la sigmoide, ReLU (Unidad Lineal Rectificada) y la tangente hiperbólica.

Limitaciones y Extensión a Redes Multicapa

A pesar de su gran utilidad, el perceptrón monocapa por si solo no es suficiente. De hecho, Minsky y Papert demostraron en 1969 que este modelo por si solo, no puede resolver problemas que no sean linealmente separables, como lo es la función lógica XOR [12].

La solución a estas limitaciones, fue la introducción de mas capas a las redes neuronales artificiales, conocidas como perceptrones multicapa (MLP, por sus siglas en inglés). Estos incorporan una o más capas ocultas entre las capas de entrada y salida, permitiendo modelar relaciones mucho mas complejas. La topología de la red se suele ajustar según el problema, cambiando el número de capas, nodos o neuronas por capa y las funciones de activación.

Una red neuronal con una sola capa oculta (o Shallow Neural Network) se puede representar como:

$$\hat{y} = f^{(2)}\left(\sum_{j=1}^{m} w_j^{(2)} \cdot f^{(1)}\left(\sum_{i=1}^{n} w_{ji}^{(1)} x_i + b_j^{(1)}\right) + b^{(2)}\right) \tag{6}$$

donde:

- $f^{(1)}$ es la función de activación de la capa oculta,
- $f^{(2)}$ es la función de activación de la capa de salida,
- $w_{ji}^{(1)}$ son los pesos de la capa oculta,
- $w_j^{(2)}$ son los pesos de la capa de salida,
- $b_j^{(1)}$ y $b^{(2)}$ son los sesgos correspondientes.

Entrenamiento y Etapas

El entrenamiento de una red neuronal consiste en ajustar los pesos y sesgos para minimizar una función de pérdida, típicamente mediante el algoritmo backpropagation combinado con un optimizador como pueden ser el descenso del gradiente estocástico (SGD) o Adam.

Las etapas del proceso de entrenamiento son:

1. **Inicialización:** Los pesos se inicializan aleatoriamente.
2. **Forward:** Se calcula la salida de la red utilizando los pesos actuales.
3. **Cálculo del error:** Se mide la diferencia entre la salida real y la predicha.
4. **Backpropagation:** Se calcula el gradiente del error respecto a los pesos actuales.
5. **Actualización:** Se actualizan los pesos utilizando el optimizador seleccionado.
6. **Iteración:** Se repite el proceso durante un número definido de épocas o bien hasta alcanzar convergencia.

Aplicación en esta Investigación

En esta tesis se utilizaron **Shallow Neural Networks**, esto ya que presentan una menor complejidad computacional en comparación con las redes profundas y son apropiadas para conjuntos de datos de tamaño moderado o bajo como los utilizados en este trabajo. Según Zhang et al. [21], las **Shallow Neural Networks** son capaces de lograr buenos niveles de rendimiento en tareas de regresión y clasificación sin requerir arquitecturas profundas, siempre y cuando se cuente con una selección adecuada de características y una arquitectura bien ajustada.

3.4. Métricas de Evaluación de Modelos

3.4.1. Métricas de Evaluación de Modelos de Regresión

- **R^2:** Métrica que determina la calidad de un modelo de regresión, para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo. Esta métrica toma valores reales entre 0 y 1 la mayor parte del tiempo, si la varianza residual es cero, el modelo predictivo explica el 100 % del valor de la variable a estimar, por otro lado si coincide con la varianza de la variable dependiente, el modelo no explica nada y esta métrica toma el valor de 0. En otras palabras, cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que se intenta explicar, por otro lado, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será. Como se dijo anteriormente, esta métrica toma valores entre 0 y 1, sin embargo, en algunos casos, puede tomar valores negativos. Esto sucede cuando el modelo de regresión es completamente inadecuado y realiza predicciones peores que si simplemente se hubiera utilizado la media de los valores observados como estimación. La formula de este es la siguiente:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Error Absoluto Medio (MAE):** Métrica que determina la magnitud media de la diferencia entre el valor estimado por el modelo y el valor real de la variable dependiente. Es útil para entender cuánto se desvían las predicciones del modelo de los valores reales. La formula de este es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Error Cuadrático Medio (MSE):** Métrica que determina la media de los cuadrados de la diferencia entre el valor estimado por el modelo y el valor real de la variable dependiente. Es útil para entender cuánto se desvían las predicciones del modelo de los valores reales. A diferencia del MAE, el MSE da más peso a los errores grandes debido al cuadrado en la fórmula. La fórmula de este es la siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Raíz del Error Cuadrático Medio (RMSE):** Métrica que determina la raíz cuadrada del promedio de los cuadrados de las diferencias entre el valor estimado por el modelo y el valor real de la variable dependiente. Es útil para entender cuánto se desvían las predicciones del modelo de los valores reales. A diferencia del MSE, el RMSE está en las mismas unidades que la variable dependiente, lo que puede facilitar su interpretación. Por otro lado, a diferencia del MAE, este penaliza los errores muy grandes de igual manera que el MSE. La fórmula de este es la siguiente:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Donde para todas las formulas, y_i es el valor observado, $\hat{y}_i$ es el valor predicho, $\bar{y}$ es el valor medio de las observaciones, n es el número total de observaciones.

3.4.2. Métricas de Evaluación de Modelos de Clasificación

- **Accuracy:** Medida que determina la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el número total de casos examinados. La formula de esta es la siguiente:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Medida que determina la proporción de verdaderos positivos entre todas las predicciones positivas (tanto verdaderos positivos como falsos positivos). La formula de esta es la siguiente:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** Medida que determina la proporción de verdaderos positivos que se identificaron correctamente. La formula de esta es la siguiente:

$$Recall = \frac{TP}{TP + FN}$$

- **Specificity:** Medida que determina la proporción de verdaderos negativos que se identificaron correctamente. La formula de esta es la siguiente:

$$Specificity = \frac{TN}{TN + FP}$$

- **F-Score (también conocido como F1):** Medida que combina Precision y Recall. Se calcula como la media armónica entre Precision y Recall. La formula de este es la siguiente:

$$F - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Donde para todas las formulas, TP son los Verdaderos Positivos, TN son los Verdaderos Negativos, FP son los Falsos Positivos, FN son los Falsos Negativos.

3.5. Selección de características

Etapa fundamental para el correcto desempeño de modelos predictivos, especialmente en contextos donde se dispone de un alto número de variables. Su propósito es identificar el subconjunto más relevante de variables explicativas, reduciendo así la dimensionalidad del problema sin perder poder predictivo, esto no solo con la finalidad de mejorar el rendimiento del modelo, sino que también se busca incrementar la interpretabilidad y reducir el sobreajuste [9].

Este proceso puede realizarse principalmente de 3 formas:

- **Métodos de filtro:** Descartando características basándose en propiedades estadísticas como la correlación, inflación de varianza o la información mutua.
- **Métodos wrapper:** Usando el desempeño de un modelo para evaluar el desempeño con distintos subconjuntos de variables.
- **Métodos embedded:** Realizando la selección en base al entrenamiento del modelo (por ejemplo, basándose en la feature importance por descenso de impureza en random forest).

3.5.1. Selección de Características por Información Mutua

Uno de los métodos mas utilizados para seleccionar características es la **información mutua** (mutual information, MI), la cual mide cuánta información aporta una variable X respecto a otra variable Y. A diferencia de otras métricas como la correlación de Pearson, la información mutua capta tanto relaciones lineales como no lineales entre variables.

Matemáticamente, la información mutua entre dos variables aleatorias discretas X y Y está definida como:

$$MI(X;Y) = H(X) - H(X|Y) \tag{7}$$

donde:

- $H(X)$ es la entropía de X. (Ecuación 2)
- $H(X|Y)$ es la entropía condicional de X e Y, la cual se define a continuación.

$$H(X|Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x|y) \tag{8}$$

La selección de características por MI implica calcular $I(X_i;Y)$ para cada variable independiente X_i, y luego seleccionar aquellas con mayor valor (ya que se correlacionan en mayor medida con la variable dependiente). Esta técnica ha demostrado ser bastante efectiva en dominios donde las relaciones entre variables no son estrictamente lineales [19].

3.5.2. Selección de características basada en random forest:

La selección de características basada en random forest implica entrenar un modelo de tipo random forest y basándose en alguna métrica como lo son Impurity Decrease y la importancia por permutacion, seleccionar las variables significativas y eliminar el resto. A continuación se definen ambas métricas.

- **Impurity Decrease:** Esta métrica se basa en el hecho de que cada vez que un nodo de un árbol se divide, la impureza (en este caso, la entropía, aunque también puede ser la ganancia de información, el índice de Gini, entre otras métricas) del nodo disminuye. La importancia de una característica se determina como la suma total de las reducciones de impureza que resultan de las divisiones donde se utiliza dicha característica, promediada a través de todos los árboles del modelo. Puede ser engañosa con características de alta cardinalidad (muchos valores únicos).

- **Importancia por Permutacion:** Esta métrica se calcula después de que el modelo ha sido entrenado. Para cada característica en el conjunto de datos, los valores se permutan de manera aleatoria y se mide la disminución en el rendimiento del modelo. Si el rendimiento del modelo disminuye de manera significativa, esto indica que la característica es importante. Esta métrica tiene la ventaja de no estar sesgada hacia características de alta cardinalidad y puede calcularse en un conjunto de datos de prueba independiente.

Como se puede ver, ambas métricas son complementarias y permiten interpretar el modelo entrenado y entender el peso relativo de cada variable de entrada. La métrica por disminución de impureza es rápida de calcular y útil durante el entrenamiento, aunque puede sesgarse hacia variables con alta cardinalidad. En cambio, la métrica basada en permutaciones es más robusta, ya que mide el efecto real de cada variable en el rendimiento del modelo al ser alterada de manera aislada. Por lo mismo, para efectos de la investigación se utilizan ambas métricas en conjunto para así realizar un análisis más completo.

3.5.3. Eliminación de características por Inflación de Varianza (VIF)

La **inflación de varianza** (Variance Inflation Factor, VIF) es una métrica usada para detectar multicolinealidad entre las variables independientes. Es decir, una VIF elevada indica que una variable independiente tiene dependencia de una o mas variables independientes, por lo que no aporta información real al modelo y solo añade redundancia, lo cual compromete la estabilidad y la interpretabilidad del modelo de regresión.

Para una variable X_j, su VIF se calcula como:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{9}$$

donde R_j^2 es el coeficiente de determinación al predecir X_j usando las demás variables como regresores.

Dicho esto, la eliminación de características por VIF se realiza de forma iterativa siguiendo los siguientes pasos:

1. Calcular el VIF de todas las variables.
2. Eliminar la variable con mayor VIF si este excede un umbral (usualmente 5 o 10).
3. Repetir hasta que todos los VIF estén bajo el umbral.

Esta técnica altamente útil en modelos de regresión lineales donde la multicolinealidad infla la varianza de los coeficientes estimados y afecta su significancia estadística [10].

3.6. Métodos de muestreo

El muestreo de datos es una técnica utilizada en análisis de datos para seleccionar, manipular y analizar un subconjunto representativo de datos para identificar patrones y tendencias.

- **Oversampling:** Es una técnica utilizada para corregir el desequilibrio en los datos al aumentar la cantidad de instancias en la clase minoritaria. Se puede hacer duplicando instancias de la clase minoritaria o generando nuevas instancias sintéticas. Es útil en problemas de clasificación donde hay una gran disparidad en las clases.
- **Undersampling:** Es una técnica que reduce la cantidad de instancias en la clase mayoritaria para equilibrar la distribución de clase. Puede ser útil cuando tienes una gran cantidad de datos y el entrenamiento del modelo es computacionalmente costoso.

En problemas de clasificación, estas técnicas son muy útiles para tratar con conjuntos de datos desequilibrados donde una clase tiene muchas más instancias que la otra. En problemas de regresión, se pueden utilizar técnicas de muestreo como SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise)[3] y SMOTER (Synthetic Minority Over-Sampling Technique for Regression)[18] , para problemas donde los valores de interés para predecir son raros o poco comunes [20].

1. **SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise):** Esta técnica genera ejemplos sintéticos en las regiones del espacio donde los valores de la variable objetivo son raros o extremos. Utiliza vecinos más cercanos para interpolar nuevos puntos y les añade ruido gaussiano para evitar la creación de duplicados. Es especialmente útil cuando se desea mejorar el aprendizaje del modelo en zonas de difícil representación (por ejemplo, valores muy bajos o muy altos de la variable objetivo).
2. **SMOTER (Synthetic Minority Over-sampling Technique for Regression):** Similar a SMOGN, SMOTER genera nuevos puntos sintéticos mediante interpolación lineal entre ejemplos minoritarios, pero sin añadir ruido. Está diseñada para mejorar la capacidad predictiva del modelo en regiones escasamente representadas, y puede ser útil cuando el ruido adicional podría afectar negativamente el rendimiento.

4. Marco de Trabajo

4.1. Herramientas a Utilizar en el Estudio

Las herramientas a utilizar para el desarrollo este proyecto de investigación serán:

- **LaTeX:** Sistema de composición de textos poderoso, utilizado gracias a la versatilidad que provee al poder fijar el formato del documento, sin sufrir mayores inconvenientes de formato a medida que el documento crece, lo que permite enfocarse exclusivamente en el contenido y no en la forma. Por otro lado provee una amplia variedad de símbolos matemáticos.
- **Overleaf:** Editor de LaTeX en linea, utilizado para la redacción del informe de proyecto en la asignatura Seminario de Titulo, y de su continuación en la asignatura de Proyecto de Titulo.
- **Google Drive:** Servicio de Almacenamiento en la nube utilizado para el almacenamiento de documentos del proyecto, literatura revisada, conjuntos de datos, y respaldos del código fuente.
- **Python:** Lenguaje de programación multiparadigma, utilizado por su facilidad de uso, y sencillez sintáctica, lo que se traduce en una codificación rápida en comparación a otros lenguajes, y una fácil interpretación del código al leerlo. Se utiliza mediante Jupyter notebooks.
- **DeepNote:** Servicio de Jupyter Notebooks utilizado para el desarrollo y pruebas de la solución.
- **GitHub:** Servicio de control de versiones utilizado como respaldo de los conjuntos de datos. También se utiliza por su integración con DeepNote que realiza automáticamente la carga de archivos en el espacio de trabajo utilizado para desarrollar, sin necesidad de cargar manualmente los archivos en cada sesión.
- **SciKit Learn:** Librería de aprendizaje automático para el lenguaje Python, utilizada para instanciar el modelo Random Forest.
- **Imbalanced Learning Regression:** Librería de tecnicas de muestreo para regresion [20].
- **MatPlotLib y Seaborn:** Librerías utilizadas para la generación de visualizaciones (Mapas de calor, ScatterPlots,Gráficos de Lineas, etc).
- **Web of Science:** Portal utilizado para la búsqueda de literatura científica.

4.2. Metodología de Trabajo

La metodología de trabajo será una **basada en la investigación aplicada en la mejora de la calidad y eficiencia de procesos** [15].

Se eligió basarse en esta metodología debido a que el fin del proyecto es justamente la mejora en la calidad de procesos, como lo son el diseño e impresión 3D de antenas en filamento PLA-BaTiO3, y las etapas de construcción y diseño de proyectos AEC.

La metodología se compondrá de 10 pasos, los cuales no necesariamente se realizan todos de forma secuencial, ni de forma única, siendo estos:

- **Familiarización con el contexto:** Se identifica el problema a resolver. Se define de forma clara los aspectos a resolver, mejorar u optimizar.
- **Revisión de literatura:** Se revisa literatura relacionada al problema identificado. Esto con el fin de conocer en mayor medida la problemática, ver las soluciones existentes, y los resultados de estas (si las hubiese).
- **Propuesta de solución e hipótesis:** Se diseña y plantea una posible solución al problema identificado. Se plantea una hipótesis que se comprobará con la solución a implementar.
- **Recopilación de Datos:** Se recopilan los datos necesarios para el correcto desarrollo de la solución. Se consulta a expertos en el tema respecto a posibles dudas que surjan.
- **Análisis de datos:** Se hace un análisis exploratorio a los datos, se generan transformaciones a estos de ser necesario. De vital importancia para la correcta utilización de estos en el desarrollo de la solución.
- **Desarrollo de la solución:** Se desarrolla la solución propuesta, habiendo iteraciones de testeo y mejora de esta si fuese necesario.
- **Evaluación de solución:** Se monitorea el funcionamiento de la solución y se reportan los resultados.
- **Interpretación de resultados:** Se interpretan los resultados reportados a lo largo de la investigación (no solo los reportados en la puesta en marcha). Se generan visualizaciones de los datos.
- **Elaboración de conclusiones:** Dadas las interpretaciones realizadas en base a los datos y las visualizaciones, se generan conclusiones a la hipótesis planteada inicialmente.
- **Publicación de resultados:** Se comunican los resultados de la investigación con la entrega del informe final de proyecto de titulo, así como también con la presentación final de este. De ser posible, se genera una publicación en revista o conferencia.

4.3. Planificación

A continuación se tiene la calendarización del proyecto, se consideran todos los entregables y tareas de la asignatura de seminario de proyecto y ademas de esto, las tareas hasta la entrega de avance de la asignatura de proyecto de titulo.

Semana	**Descripción**
26 de Enero de 2024	Pre-inscripción de Seminario de Proyecto.
30 de Enero de 2024	Recepción de Conjunto de Datos Inicial.
11 de Marzo de 2024	Entrega de Formulario de Inscripción de Seminario de Proyecto.
12 de Marzo de 2024	Recepción de información relevante para la investigación (Papers profesora Dreidy).
13 al 20 de Marzo de 2024	Codificación de Prototipo Inicial con modelo Random Forest. Transformaciones básicas a los datos (Pasar a csv, ordenar datos, transformar observaciones en categorías).
21 de Marzo de 2024	Reunión con Profesora Dreidy Vasquez (EIQ) para recibir detalles del proyecto, dar a conocer inquietudes respecto a los datos, además de solicitar mayor cantidad de datos.
26 de Marzo de 2024	Reunión con Profesor Francisco Pizarro (EIE) y Profesora Dreidy Vasquez (EIQ) para aclarar dudas respecto a los datos, recibir mayores detalles del proyecto y fijar compromiso de entrega de datos para día 15 de Abril.
1 al 12 de Abril de 2024	Redacción de primera versión del Informe de Avance de Seminario de Proyecto.
15 al 19 de Abril de 2024	Correcciones del Informe de Avance según indicaciones de Profesora Leslie Pérez.
16 de Abril de 2024	Recepción de nuevos datos por parte de Lucas Ruz (Tesista Profesora Dreidy Vasquez).
19 de Abril de 2024	Entrega del Informe de Avance de Seminario de Proyecto.
22 de Abril al 3 de Mayo de 2024	Revisión y modificaciones a códigos. Pruebas con modelo Random Forest.
2 de Mayo de 2024	Presentación de Avance de Seminario de Proyecto.
6 de Mayo al 2 de Junio de 2024	Correcciones preliminares al Informe Final. Modificaciones al código existente y codificación de prototipo civil.
13 de Mayo de 2024	Recepción de datos de Ingeniería Civil.
3 al 21 de Junio de 2024	Ajustes al modelo civil. Finalización del Informe Final.
21 de Junio de 2024	Entrega Informe Final de Seminario de Proyecto.
4 de Julio de 2024	Presentación Final de Seminario de Proyecto.
Julio 2024 - Inicios Febrero 2025	Pausa por intercambio a Japón.
Inicios Febrero - Inicios Marzo 2025	Pruebas con modelos distintos a Random Forest.
10 de Marzo de 2025	Inscripción de Proyecto de Título.
13 al 31 de Marzo de 2025	Pruebas y tabulación de modelos con Feature Selection. Reorganización de código para mayor claridad.
1 al 14 de Abril de 2025	Pruebas estadísticas a los modelos (inflación de varianza, eliminación de variables, correlación entre variables independientes, validación de datos y rendimiento).
15 de Abril de 2025	Solicitud de nuevos datos al profesor Rodrigo Herrera por medio de Profesora Leslie Pérez.

Semana	Descripción
15 al 25 de Abril de 2025	Trabajo en el Informe de Avance de Proyecto de Título y su presentación.
25 de Abril de 2025	Entrega del Informe de Avance de Proyecto de Título.
6 de Mayo de 2025	Presentación de Avance de Proyecto de Título.
7 de Mayo al 2 de Junio de 2025	Modificaciones al código existente, realización de experimentos dataset problema impresión de antenas.
2 al 16 de Junio de 2025	Análisis de resultados de todos los experimentos. Finalización de Informe Final de Proyecto de Titulo
16 al 20 de Junio de 2025	Correcciones del Informe Final de Proyecto de Titulo según indicaciones de Profesora Leslie Pérez.
20 de Junio de 2025	Entrega Informe Final de Proyecto de Titulo.
9 de Julio de 2025	Presentación Final de Proyecto de Titulo.

Se adjunta gantt en los anexos debido al tamaño A.2, y en caso de requerirse, se subió a la nube (click acá para ir al enlace) para poder hacer zoom sin problemas.

5. Descripción de los datos

5.1. Problema Impresión de Antenas

La Escuela de Ingeniería Química desarrolló un filamento de impresión 3D con propiedades dieléctricas (PLA-BaTiO3). Además de esto, adaptaron una impresora 3D para el correcto manejo de este material. Esta impresora debe ser configurada para cada impresión que se realiza, lo cual actualmente se hace de forma manual. Como se explicó previamente también, esto requiere de experiencia y un proceso experimental de ensayo-error. Actualmente, se han impreso 29 piezas cilíndricas de manera experimental, a las cuales se les hizo un análisis de distintas características. Estas piezas fueron realizadas en 2 tandas: una primera tanda de 9 piezas, que fue el primer conjunto de datos, recibido el 30 de Enero de 2024, y una segunda tanda de 20 piezas, recibida el 16 de Abril. Cabe destacar que ambas tandas difieren en cuanto a cómo se registraron los datos, esto debido a que la 2^{da} tanda de datos, fue registrada posterior a las reuniones del 21 y 26 de Marzo, donde se hicieron sugerencias de cambios a cómo registraban los datos, en pos de estandarizarlos.

N°Pieza	Temperatura Impresión [°C]	Flujo	Altura de capa [mm]	Velocidad Impresión [mm/s]	% Ventilador	Tiempo de Impresión [min]	Peso [g]	Observaciones
1	250	500	0,32	4	10	22	1,39	Buen acabado de pieza pero presenta capas desiguales al inicio y final de pieza
2	237	450	0,32	8	10	17	1,04	La pieza presenta capas desiguales al comienzo(pie de elefante) y falta de material entre capas
3	240	450	0,32	4	10	22	0,97	Buen acabado de pieza pero presenta falta de material entre capas
4	237	500	0,32	2	10	32	1,31	Buen acabado de pieza pero presenta capas desiguales
5	250	500	0,32	1	20	52	1,51	Capas Desiguales boquilla repasa la pieza leve warping
6	250	500	0,32	2	30	32	1,42	Buen acabado delaminacion en las ultimas capas por falta de material
7	250	500	0,32	2	30	32	1,68	Capas Desiguales
8	250	500	0,32	2	50	32	1,03	Capas Desiguales con delaminacion de las capas
9	250	500	0,16	2	20	36	2,36	Leve desigualdad de capas superiores en la pieza pero logrando un buen acabado

Figura 4: Conjunto de datos Cilíndricos de la primera tanda

La variable a predecir en este conjunto de datos son las observaciones. Sin embargo, estas no pueden ser introducidas en un modelo en su estado bruto, por lo que es necesario establecer una escala para transformar las observaciones en categorías, y así poder introducir los datos a un modelo de clasificación. En este contexto, se sugiere una clasificación en tres clases, de acuerdo con algún criterio que permita a quienes proporcionan los datos, evitar la pérdida de sus datos existentes y además prevenir que estos generen ruido. La escala seleccionada por ellos fue la siguiente:

Categoría	
1	**Sinterizable**
2	**Defectos Menores Apta Para sinterizar**
3	**No Apta para Sinterizar**

Utilizando la escala seleccionada para clasificar las antenas producidas, el conjunto de datos queda de la siguiente manera:

N°Pieza	Temperatura Impresión [°C]	Flujo	Altura de capa [mm]	Velocidad Impresión [mm/s]	% Ventilador	Tiempo de Impresión [min]	Diametro	Altura	Categoria
1	250	500	0.32	4	10	22	11.889	4.012	1
2	237	450	0.32	8	10	17	12.234	3.789	2
3	240	450	0.32	4	10	22	12.102	3.869	2
4	237	500	0.32	2	10	32	12.201	4.201	2
5	250	500	0.32	1	20	52	11.83	3.795	3
6	250	500	0.32	2	30	32	12.145	3.962	2
7	250	500	0.32	2	30	32	12.21	4.19	2
8	250	500	0.32	2	50	32	12.305	3.84	2
9	250	500	0.16	2	20	36	12.089	4.087	1

Figura 5: Conjunto de datos Cilíndricos de la primera tanda, recibidos con correccion

Los datos restantes del conjunto de datos se encuentran para su consulta en el Anexo A.1.

5.2. Problema Comunicación Eficiente en Equipos

En los proyectos de arquitectura, ingeniería y construcción, la alta fragmentación requiere altos niveles de colaboración entre equipos. Sin embargo, estos niveles de colaboración a menudo no son cubiertos por la tecnología existente. Aquí es donde entra en juego el trabajo en equipo, que se basa en la comunicación, la coordinación y el respeto entre los miembros de este.

El análisis de redes sociales (SNA) es una técnica que se utiliza para estudiar el flujo de información en estos proyectos, ya que permite analizar la fortaleza de las relaciones dentro de un equipo, que son críticas para el éxito organizacional y la satisfacción de los trabajadores. En este contexto, una red social se puede visualizar como un grafo compuesto por "actores" (nodos) y "vínculos sociales" (enlaces). En este caso particular, los nodos son los miembros del equipo de trabajo de un proyecto, los stakeholders o los especialistas. La topología de estas redes permite analizar patrones de interacción.

Cuando se evalúan las relaciones de trabajo, las métricas del SNA son útiles. El grado de cohesión entre los individuos se mide por el grado medio, el diámetro, la densidad, el número de componentes y el coeficiente de clustering promedio.

Aquí es donde la importancia de BIM (Building Information Modeling) y Lean se hace evidente. Ambos se plantean como soluciones para los problemas de comunicación, colaboración y coordinación en los equipos de trabajo. Los estudios revelan que en los equipos de trabajo que utilizan Lean y BIM hay menos cuellos de botella[22]. Es decir, sus nodos están altamente conectados, por lo que son menos propensos a separarse si se ausenta parte del equipo.

Además, los equipos de trabajo altamente conectados tienen un mayor rendimiento, lo que demuestra la importancia de la colaboración y la comunicación efectiva en los proyectos de arquitectura, ingeniería y construcción. En resumen, la combinación de SNA, BIM y Lean pareciese ser una herramienta poderosa para formar equipos altamente comunicativos y colaborativos. Sin embargo, no son suficientes, por lo que se busca estimar las métricas de los grafos con datos disponibles sobre equipos de trabajo en obras civiles de distintos edificios. Con esto, también se pretende tener una mejor noción de qué variables dependen en mayor medida de la variable objetivo.

El conjunto de datos para esto, tiene las siguientes características.

Cuadro 2: Variables independientes

Variable Independiente	**Descripción**	**Rango de Valores**
Tipo de proyecto	Edificación, Infraestructura.	{0,1}
Fase de proyecto	Diseño, construcción.	{0,1}
Cliente	Tipo de mandante (público, privado).	{0,1}
Estrategia de contrata	Diseño-Licita-Constr, Cost-Oper-Transf	{0,1}
GL - Relación Stakeholders	-	{0,1,2,3}
GL - Planificación	-	{0,1,2,3}
GL - Toma Decisión	-	{0,1,2,3}
Aplicación BIM	-	{0,1,2,3}
Tipo de Red	Información, Interacción, Colaboración, Planificación	{0,1,2,3}
Número de Personas	Número entero	$[0,\infty)$
% Gestores	Proporción de personas con rol de gestión.	[0,1]
% de ejecutantes	Proporción de personas con rol de ejecución.	[0,1]

Cuadro 3: Variables dependientes

Variable	**Descripción**	**Rango de Valores**
Número de comunidades	Subgrupos desconectados o débilmente conectados.	$[0,\infty)$
Número de personas aisladas	Nodos sin conexiones (grado = 0).	$[0,\infty)$
Número de cuellos de botella	Personas que dividen la red si se eliminan.	$[0,\infty)$
Grado medio	Promedio de conexiones por persona.	$[0,\infty)$
Longitud media de camino	Promedio de pasos entre nodos.	$[0,\infty)$
Grado de jerarquización	Nivel de centralización en torno a ciertos nodos.	[-1,1]

Cabe destacar que se utiliza notación de intervalos [17] al describir los rangos de valores, es decir, los rangos de valores entre corchetes son rangos continuos (pudiendo ser abiertos, semiabiertos, semicerrados, o cerrados), por otro lado, los valores que se encuentren entre llaves representan categorías, por lo que corresponden únicamente a los valores listados, y no a rangos. Además de esto, el conjunto de datos contiene 126 registros o filas, y debido a su tamaño no se adjuntará en el informe.

6. Experimentos

Esta sección describe los experimentos realizados a lo largo del proyecto. Estos experimentos evaluarán la calidad del conjunto de modelos seleccionado para predecir las variables dependientes en los dataset estudiados. Se evaluará el uso de estos modelos junto a técnicas de selección de características y oversampling.

Ademas se incluye la descripción del proceso utilizado para preparar los datos, entrenar los modelos y evaluar su rendimiento en ambos casos de estudio: la predicción del acabado de las impresiones 3D (problema de las antenas) y el análisis de desempeño comunicacional en equipos de trabajo en obras civiles (problema de las comunicaciones). Se aplicaron distintos enfoques de preprocesamiento y modelado para evaluar el impacto de la selección de características y del oversampling.

6.1. Preparación de los Datos

Para ambos conjuntos de datos, el dataset de antenas y el dataset de proyectos, se aplicó una rutina de limpieza y transformación basada en los siguientes pasos:

1. **Eliminación de columnas no informativas:** Se descartaron columnas identificadoras (ID), ya que no aportan valor predictivo.

2. **Sanitización de variables porcentuales:** Todas las variables que contenían porcentajes estan acotadas al intervalo cerrado $[0, 1]$.

3. **Conversión explícita de variables categóricas:** Debido a que la biblioteca `pandas` no reconocía automáticamente todas las variables categóricas, se forzó su tipificación a tipo `category`. Esta operación fue necesaria para garantizar un tratamiento consistente en las librerías `scikit-learn` y `PyTorch`, las cuales manejan de forma diferente los datos categóricos y numéricos.

4. **Normalización de variables numéricas:** Finalmente, todas las variables numéricas continuas (no categóricas) fueron normalizadas utilizando la función `MaxAbsScaler` de la librería `scikit-learn`.
   Esta normalización consiste en escalar cada característica (X) dividiendo sus muestras por el valor absoluto del máximo valor que toma esta misma. A continuación se presenta la fórmula:

$$X' = \frac{X}{\text{máx}(|X|)} \tag{10}$$

   A consecuencia de esto, las variables (X') quedan en el rango $[-1, 1]$ y conservan el signo de la variable original. Además mantiene la proporcionalidad entre variables y no desplaza la media de las variables a cero, lo que preserva la interpretación de las magnitudes originales y es consistente con la naturaleza de los datos.

6.2. Enfoques de Evaluación

Cada modelo fue evaluado bajo tres escenarios distintos manteniendo las mismas proporciones de training (80 % de los datos) y testing (20 % de los datos) para todos los escenarios:

1. **Configuración Vanilla:** Los modelos fueron entrenados y evaluados directamente sobre los datos preprocesados sin aplicar selección de características ni balanceo.

2. **Selección de Características (FS):** Se aplicó una combinación de técnicas de filtrado, incluyendo información mutua y eliminación por inflación de varianza (VIF), para reducir la dimensionalidad antes del entrenamiento, para mayor información referirse a la sección 3.5.

3. **Sobre-muestreo (SMOTER):** Se eligió aplicar sobre-muestreo sintético mediante `SMOTER` (Synthetic Minority Over-sampling Technique for Regression, mayor informacion en 2) después de la selección de características, con el objetivo de equilibrar la distribución de los valores objetivo y mejorar el ajuste en regiones con baja densidad de datos.

6.3. Modelos Entrenados

Los modelos utilizados, fueron seleccionados para representar un conjunto relevante dentro del área de aprendizaje automático. Las tablas 4 y 5 listan los modelos seleccionados, incluyendo la librería utilizada para su ejecución y los parámetros asociados a los modelos de este proyecto:

Cuadro 4: Resumen de parámetros utilizados por modelos (Experimento de las Comunicaciones)

Experimento de las Comunicaciones (Regresión)	
Modelo	**Hiperparámetros**
Regresión Lineal (scikit-learn)	Por defecto.
Regresión Lasso (scikit-learn)	Por defecto.
Regresión Ridge (scikit-learn)	Por defecto.
Random Forest (scikit-learn)	`Numero de Arboles (mayor información en 1)=100`, `Bootstrap (mayor información en 2)=True`, `Máximo de muestras (mayor información en 2a)=0.8`
Red Neuronal Superficial (PyTorch)	`Nodos de Entrada=10`, `Nodos Ocultos=14`, `Nodos de Salida=6`, `Función de Pérdida: MSE`, `Optimizador: Adam` con `Tasa de Aprendizaje =0.1`

Cuadro 5: Resumen de parámetros utilizados por modelo (Experimento Antenas)

Experimento Antenas (Clasificación)	
Modelo	**Hiperparámetros**
Regresión Logística (scikit-learn)	Optimizador: `SAGA`
R. Logística con Penalización L2 (Ridge) (scikit-learn)	Optimizador: `SAGA`
R. Logística con Penalización L1 (Lasso) (scikit-learn)	Optimizador: `SAGA`
Random Forest (scikit-learn)	`Numero de Arboles (mayor información en 1)=100`, `Bootstrap (mayor información en 2)=True`, `Máximo de muestras (mayor información en 2a)=0.8`
Red Neuronal Superficial (PyTorch)	`Nodos de Entrada=5`, `Nodos Ocultos=10`, `Nodos de Salida=1`, `Función de Pérdida: MSE`, `Optimizador: Adam` con `Tasa de Aprendizaje=0.1`

Cada modelo fue entrenado y evaluado de forma independiente para predecir cada una de las variables dependientes del problema en cuestión. Se utilizaron set de entrenamiento y validación (no se aplicó validación cruzada), manteniendo consistencia en la comparación entre algoritmos. El rendimiento de los modelos fue evaluado mediante las métricas presentadas en la sección 3.4.

7. Experimento de las Comunicaciones en Proyectos AEC

En esta sección se aborda el problema de dinámicas de comunicación y colaboración en proyectos AEC. Se explora la estructura del conjunto de datos, se identifican redundancias y se evalúa el rendimiento de distintos modelos de regresión bajo tres configuraciones: *vanilla*, selección de características y selección con sobre-muestreo. Finalmente, se discuten los resultados obtenidos.

7.1. Análisis exploratorio de los datos

El análisis exploratorio de los datos reveló inicialmente una correlación inversa muy marcada entre las variables `% Gestores` y `% Ejecutantes`. Esto es consistente con el hecho de que, por definición, un mayor porcentaje de gestores implica necesariamente un menor porcentaje de ejecutantes, y viceversa. Al ser linealmente dependientes, ambas variables aportan la misma información al modelo, por lo que se decidió eliminar la variable `% Gestores` para evitar redundancias.

Otro hallazgo importante fue que las variables asociadas a la gestión Lean presentaron una alta correlación entre sí, así como una relación inversa considerable con la fase del proyecto. Esto podría generar problemas de multicolinealidad al momento de ajustar los modelos.

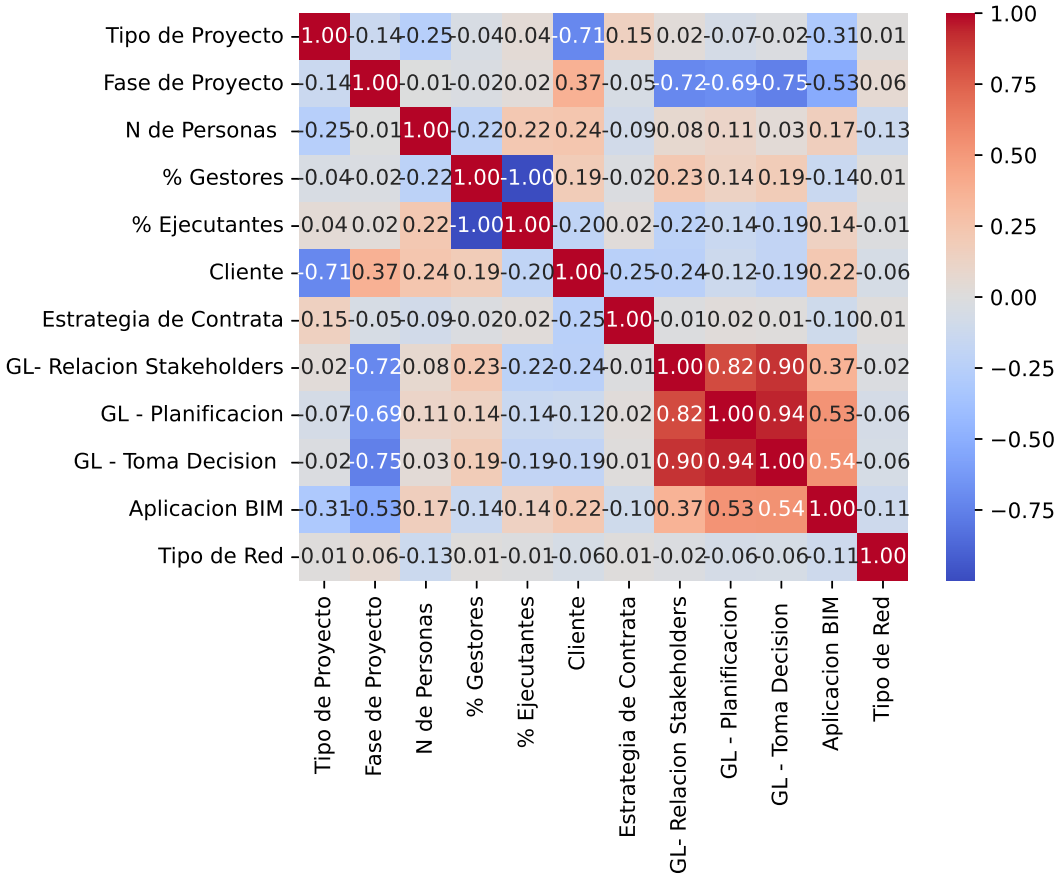

Figura 6: [Problema Comunicaciones] Mapa de Correlación de Variables Independientes

Para abordar la multicolinealidad sin eliminar excesivamente variables de interés, se procedió a calcular el factor de inflación de varianza (VIF) para las variables independientes. El uso del VIF permite detectar colinealidad lineal que no es necesariamente evidente mediante la inspección directa de correlaciones bivariadas. Este indicador permite identificar aquellas variables que están altamente correlacionadas linealmente con el resto de las variables explicativas.

Cuadro 6: Valores de Inflación de Varianza (VIF) para las variables independientes (primera iteración)

Variable Independiente	**VIF**
Tipo de Proyecto	1.85
Fase de Proyecto	6.96
Número de Personas	7.00
% Ejecutantes	6.13
Cliente	10.48
Estrategia de Contrata	1.09
GL - Relación Stakeholders	9.86
GL - Planificación	14.47
GL - Toma Decisión	**28.49**
Aplicación BIM	3.31
Tipo de Red	2.03

Se decidió eliminar la variable `GL - Toma Decisión` por presentar el VIF más elevado, reduciendo así la multicolinealidad general en el modelo. Tras su eliminación, se recalculó el VIF para el conjunto de variables restantes.

Cuadro 7: Valores de Inflación de Varianza (VIF) para las variables independientes (segunda iteración)

Variable Independiente	**VIF**
Tipo de Proyecto	1.78
Fase de Proyecto	6.96
Número de Personas	6.72
% Ejecutantes	6.13
Cliente	**10.41**
Estrategia de Contrata	1.09
GL - Relación Stakeholders	5.31
GL - Planificación	5.62
Aplicación BIM	3.02
Tipo de Red	2.03

Como se observa, el VIF disminuyó notablemente para la mayoría de las variables. Sin embargo, la variable `Cliente` mantuvo un VIF superior a 10, lo que podría indicar colinealidad moderadamente alta [11]. A pesar de ello, se decidió mantener esta variable debido a que no presenta una correlación lineal significativa con las otras variables de forma directa (lo que podría sugerir colinealidad estructural más que correlación empírica directa). Además, tal como advierte O'Brien (2007), eliminar variables automáticamente al superar un umbral arbitrario de VIF puede ser inapropiado en ciertos contextos donde la interpretación teórica es relevante [14].

7.2. Resultados obtenidos

Esta sección presenta los resultados obtenidos tras aplicar distintos modelos de regresión sobre las variables dependientes definidas en los casos de estudio. Se reportan las métricas de rendimiento obtenidas para cada configuración experimental: ejecución directa del modelo (*vanilla*), selección de características (FS), y selección de características con sobre-muestreo mediante SMOTER (FS+SMOTER).

A continuación se presentan los resultados obtenidos para cada configuración experimental.

7.2.1. Resultados sin Selección de Características (*Vanilla*)

En esta configuración, los modelos fueron entrenados y evaluados utilizando la totalidad de las variables disponibles (previo preprocesamiento), sin aplicar técnicas de reducción de dimensionalidad ni balanceo.

- Se observan diferencias significativas en el rendimiento entre modelos lineales y no lineales.
- El modelo **Shallow Neural Network (SNN)** presentó los mejores desempeños en múltiples variables.
- Random Forest superó a los modelos lineales en algunas métricas, pero en ciertos casos su rendimiento fue bajo (e.g., cuellos de botella).

Cuadro 8: [Problema Comunicaciones] Resultados Obtenidos con las Técnicas *vanilla*

Variables Dependientes	**Random Forest**		**Regresión Lineal**		**Regresión Ridge**		**Regresión Lasso**		**Shallow Neural-Net**	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Número de Comunidades	0.303	2.055	0.493	1.404	0.448	1.531	-0.024	2.837	**0.924**	**1.139**
Número de Personas Aisladas	-0.153	2.830	0.181	1.270	-0.042	1.617	-0.002	1.555	**0.951**	**0.818**
Número de Cuellos de Botella	-0.288	0.922	-0.848	0.637	0.007	0.342	-0.056	0.364	**0.986**	**0.136**
Grado Medio	0.357	11.965	0.618	10.964	**0.654**	**9.912**	0.222	22.295	0.309	20.649
Longitud Media Camino	0.498	**0.068**	0.420	0.077	0.383	0.082	-0.018	0.135	**0.981**	**0.113**
Grado de Jerarquización	0.018	0.020	0.128	0.022	**0.231**	**0.020**	-0.039	0.026	-0.564	14.016

7.2.2. Resultados con Selección de Características (FS)

Posteriormente, se aplicaron técnicas de selección de características, reduciendo el número de variables independientes a aquellas con mayor relevancia según información mutua y con baja colinealidad según VIF.

- Se evidenció una mejora general en los modelos lineales, lo que sugiere que la eliminación de ruido y redundancia benefició a estas arquitecturas.
- El rendimiento de SNN se mantuvo competitivo en general, la predicción del grado de jerarquización empeoró en mas de un 100 %, sin embargo se observa una mejora sustancial en la predicción del grado medio.
- Random Forest mostró resultados más consistentes en comparación con la configuración anterior.

Cuadro 9: [Problema Comunicaciones] Resultados Obtenidos con las Técnicas + Feature Selection

Variables Dependientes	**Random Forest**		**Regresión Lineal**		**Regresión Ridge**		**Regresión Lasso**		**Shallow Neural-Net**	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Número de Comunidades	0.351	1.797	0.462	1.490	0.463	1.489	-0.024	2.837	**0.915**	**1.271**
Número de Personas Aisladas	0.203	**1.236**	0.265	1.140	0.150	1.319	-0.002	1.555	**0.817**	**3.026**
Número de Cuellos de Botella	0.029	0.335	-0.176	0.405	0.062	**0.323**	-0.056	0.364	**0.928**	**0.674**
Grado Medio	0.591	11.716	0.584	11.932	**0.599**	**11.486**	0.222	22.295	0.568	12.922
Longitud Media Camino	-0.024	0.136	0.036	0.128	0.121	**0.116**	-0.018	0.135	**0.978**	**0.129**
Grado de Jerarquización	**0.383**	**0.016**	0.344	0.017	0.348	0.017	-0.039	0.026	-2.425	30.695

7.2.3. Resultados con FS + Sobre-muestreo (SMOTER)

Finalmente, se aplicó la técnica SMOTER sobre los datos previamente seleccionados, con el objetivo de equilibrar la distribución de los valores objetivo en el entrenamiento.

- Se observaron mejoras moderadas en el desempeño de random forest.
- Los modelos basados en regresión lineal presentaron inestabilidad y baja de rendimiento, reflejando sobreajuste al conjunto de datos de entrenamiento o sensibilidad al ruido sintético.
- El modelo de red neuronal superficial (SNN) no fue considerado en esta configuración debido a limitaciones técnicas. Esto debido a que el problema corresponde a una regresión multiobjetivo, y tras una revisión exhaustiva de literatura, no se identificaron métodos publicados que permitieran aplicar técnicas de sobremuestreo en tareas de regresión multiobjetivo.

Cuadro 10: [Problema Comunicaciones] Resultados Obtenidos con Feature Selection + Oversampling

Variables Dependientes	**Random Forest**		**Regresión Lineal**		**Regresión Ridge**		**Regresión Lasso**		**Shallow Neural-Net**	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Número de Comunidades	**0.454**	**1.512**	0.297	1.947	0.265	2.037	-0.458	4.041	–	–
Número de Personas Aisladas	0.100	1.396	**0.146**	**1.325**	0.124	1.360	-0.132	1.756	–	–
Número de Cuellos de Botella	**0.190**	**0.279**	-1.640	0.910	-0.300	0.448	-0.002	0.345	–	–
Grado Medio	0.591	11.727	0.615	11.026	**0.627**	**10.679**	0.276	20.766	–	–
Longitud Media Camino	**0.357**	**0.085**	-0.533	0.203	-0.408	0.186	-1.155	0.285	–	–
Grado de Jerarquización	**0.126**	**0.022**	-0.355	0.034	-0.081	0.027	-0.191	0.030	–	–

7.3. Análisis de Resultados

A pesar del número limitado de observaciones (126 registros), los modelos entrenados demostraron una capacidad para capturar tendencias relevantes en las dinámicas colaborativas de equipos de trabajo en el contexto de proyectos AEC. El análisis exploratorio dejó al descubierto redundancias y colinealidad entre variables, las cuales fueron mitigadas mediante selección de características, beneficiando especialmente a los modelos lineales.

A continuación se presentan gráficos radar que permiten identificar visualmente lo previamente descrito. En estos, cada vértice representa una de las variables dependientes, y los valores corresponden a los coeficientes de determinación (R^2). Con esta visualización, el comparar las fortalezas y debilidades de cada enfoque sobre los distintos indicadores de desempeño comunicacional es más intuitivo.

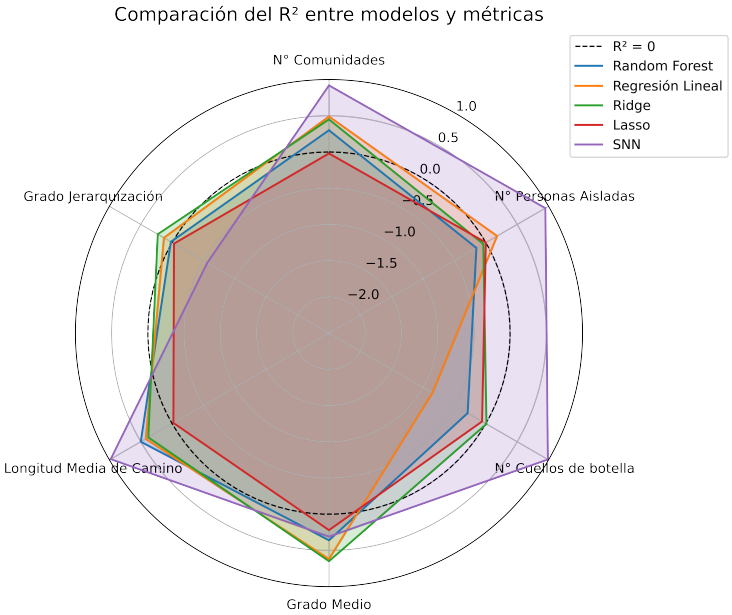

Figura 7: [Problema Comunicaciones] Radar Chart R^2 – Configuración Vanilla

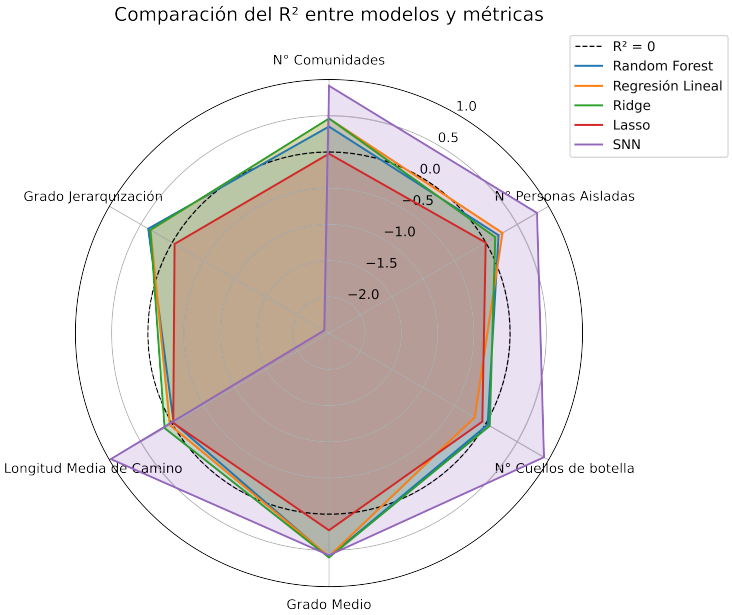

Figura 8: [Problema Comunicaciones] Radar Chart R^2 – Configuración FS

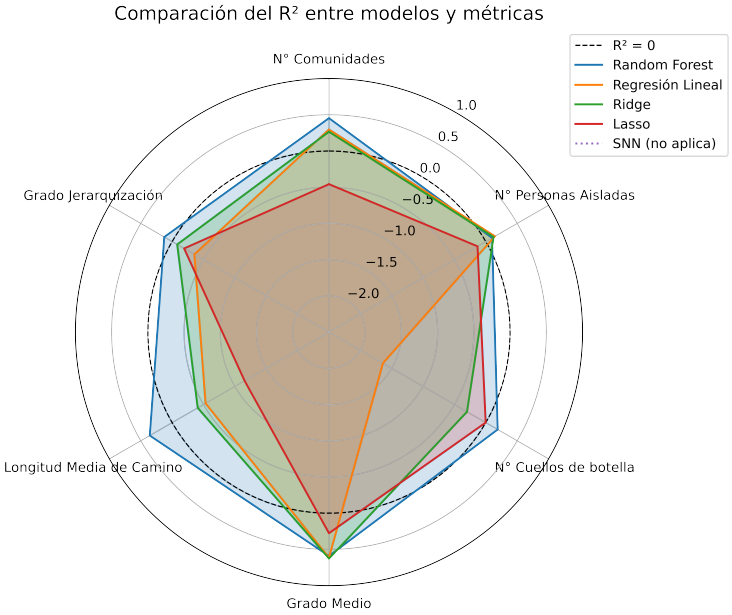

Figura 9: [Problema Comunicaciones] Radar Chart R^2 – Configuración FS+SMOTER

Como se puede observar, la red neuronal superficial (SNN) destacó como el modelo con mejor rendimiento general en múltiples variables, mostrando una alta capacidad para modelar relaciones no lineales. Además, los modelos lineales (especialmente la regresión Ridge) mostraron resultados sólidos y consistentes, en especial luego de aplicar selección de características. Sin embargo, Random Forest presentó un desempeño irregular, y en general, no logró superar a los modelos lineales salvo en el escenario con sobre-muestreo, donde mostró mejoras moderadas.

Estos resultados respaldan el uso de aprendizaje automático como herramienta para el monitoreo automatizado de la colaboración en proyectos de ingeniería. La posibilidad de predecir métricas estructurales de red abre la puerta a un seguimiento proactivo de la dinámica comunicacional, permitiendo detectar y abordar tempranamente problemas de fragmentación, aislamiento o ineficiencia en equipos de trabajo multidisciplinarios.

8. Experimento de Impresiones 3D en Antenas

En esta sección se aborda el problema de predecir el acabado superficial de piezas impresas con filamento PLA-$BaTiO_3$, en función de su configuración de impresión. Se explora la estructura del conjunto de datos, y se evalúa el rendimiento de distintos modelos de clasificación bajo las mismas tres configuraciones experimentales empleadas anteriormente: *vanilla*, selección de características y selección con sobremuestreo. Finalmente, se discuten los resultados obtenidos

8.1. Análisis exploratorio de los datos

El conjunto de datos correspondiente al problema de impresión 3D en antenas presenta un volumen extremadamente reducido, con solo 20 observaciones distribuidas en tres clases: piezas sinterizables (6 muestras), piezas con defectos menores pero aptas para sinterizar (8 muestras), y piezas no aptas para sinterizar (6 muestras) (Ver 10). Esta distribución relativamente equilibrada entre clases permitió aplicar técnicas de sobremuestreo sin requerir un preprocesamiento adicional de balanceo previo.

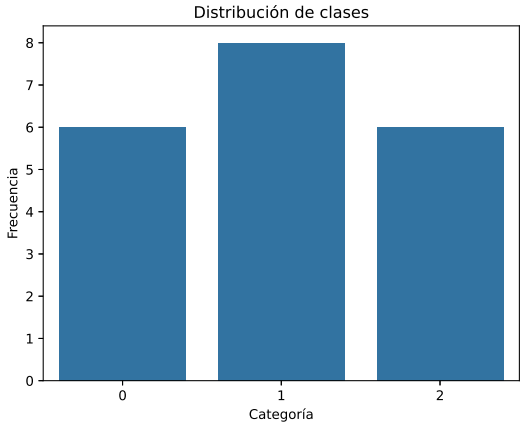

Figura 10: [Problema Antenas] Distribución de Clases

A nivel descriptivo, los boxplots revelan algunos patrones preliminares entre las variables de impresión y las categorías de calidad superficial:

Altura de capa: Las piezas no sinterizables tienden a presentar alturas de capa más bajas, mientras que las sinterizables muestran valores más altos (ver fig. 11). Esto sugiere una posible relación entre altura de capa y calidad del acabado. Esta variable podría tener un rol importante en la diferenciación de clases.

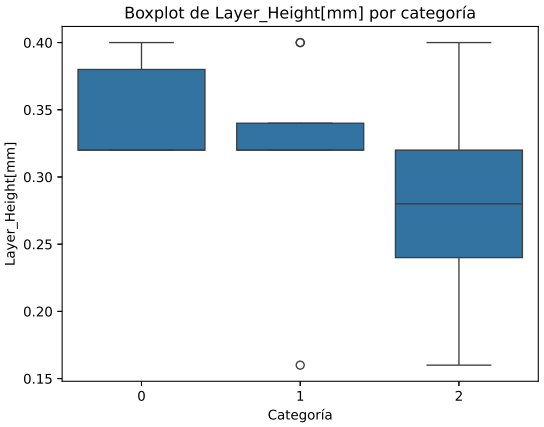

Figura 11: [Problema Antenas] Box Plot Altura de capa vs Categoría

Velocidad de impresión: Las piezas con defectos menores muestran una alta variabilidad en la velocidad de impresión, abarcando casi todo el rango observado. En contraste, las piezas sinterizables tienden a fabricarse a mayores velocidades, mientras que las no sinterizables tienden a fabricarse a velocidades más bajas (ver fig. 12).

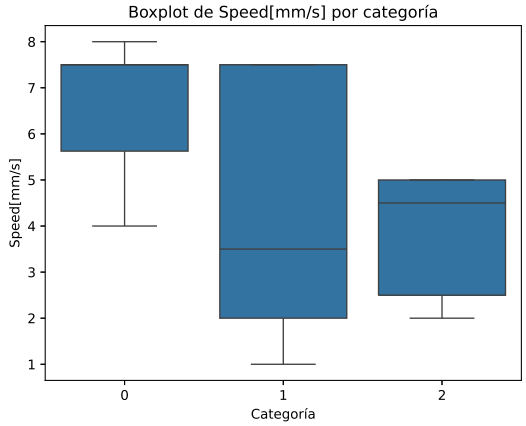

Figura 12: [Problema Antenas] Box Plot Velocidad de Impresión vs Categoría

Temperatura de impresión: Las piezas no sinterizables y con defectos menores abarcan un rango más amplio de temperaturas, mientras que las piezas sinterizables tienden a estar dentro de un rango más estrecho e inferior (ver fig. 13). Esto podría indicar que un exceso de temperatura afecta negativamente el acabado de las piezas.

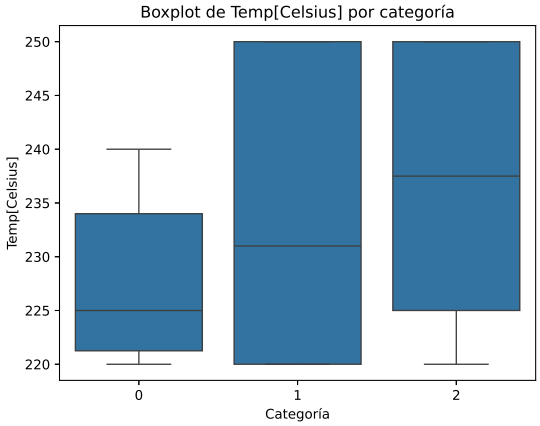

Figura 13: [Problema Antenas] Box Plot Temperatura de Impresión vs Categoría

Flujo de material: No se observa una diferenciación clara entre clases en términos del flujo (ver fig. 14).

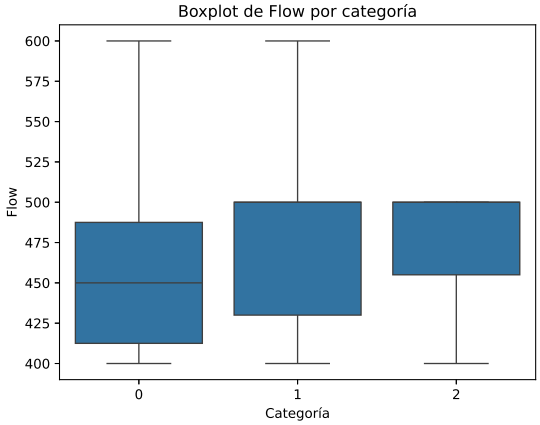

Figura 14: [Problema Antenas] Box Plot Flujo vs Categoría

Forma de la impresión: Hay un leve sesgo respecto a las piezas sinterizables, ya que la mayoría de estas son rectangulares (ver fig. 15).

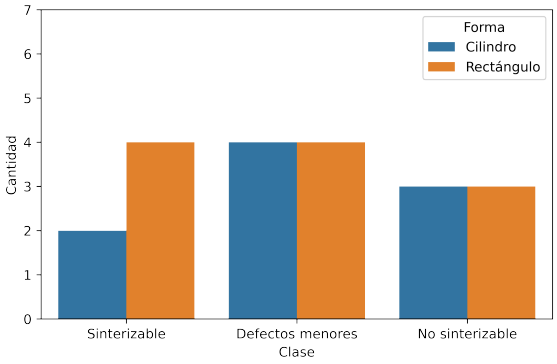

Figura 15: [Problema Antenas] Diagrama forma vs Categoría

Dado el volumen extremadamente reducido de datos y la baja dimensionalidad del conjunto, el análisis exploratorio tuvo un carácter meramente ilustrativo, evitando el análisis de colinealidad y aplicar reducciones de dimensionalidad que pudieran comprometer la ya limitada variabilidad.

8.2. Resultados obtenidos

Esta sección presenta los resultados obtenidos tras aplicar distintos modelos de clasificación sobre la variable dependiente. Se reportan las métricas de rendimiento obtenidas para cada configuración experimental: ejecución directa del modelo (*vanilla*), selección de características (FS), y selección de características con sobre-muestreo mediante SMOTE (FS+SMOTE).

A continuación se presentan los resultados obtenidos para cada configuración experimental.

8.2.1. Resultados sin Selección de Características (*Vanilla*)

En esta configuración, los modelos fueron entrenados y evaluados utilizando la totalidad de las variables disponibles, sin aplicar técnicas de reducción de dimensionalidad ni balanceo.

- Se observó un rendimiento general deficiente: todos los modelos fueron capaces de identificar únicamente una clase. En el caso de *Random Forest*, se logró detectar la clase no apta para sinterizar, mientras que los demás modelos solo reconocieron la clase con defectos menores. En ambos casos, la precisión alcanzada fue baja. Este comportamiento podría atribuirse al reducido volumen de muestras disponibles, lo que dificulta una representación adecuada de todas las clases en el entrenamiento.

A continuación, en la tabla 11, se presentan los resultados obtenidos bajo la configuración *vanilla*.

Cuadro 11: [Antenas] Resultados con Configuración *Vanilla*

Clase	**Random Forest**			**Regresión Logística**			**Regresión Ridge (Logística)**			**Regresión Lasso (Logística)**			**SNN**		
	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.
1	0.25	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
2		0	0		1	0.67		1	0.67		1	0.67		1	0.5
3		1	0.33		0	0		0	0		0	0		0	0

8.2.2. Resultados con Selección de Características (FS)

Posteriormente, se aplicó un proceso de selección de características, reduciendo el número de variables independientes a las cuatro con mayor relevancia según información mutua.

- No se observaron mejoras en el rendimiento de los modelos. Se identificó una leve degradación en la *precisión* de las regresiones logística y Lasso al predecir piezas con defectos menores. Este comportamiento podría atribuirse a la combinación entre la reducción de dimensionalidad y el escaso volumen de muestras disponibles, lo que dificulta la correcta representación de las clases.

A continuación, en la tabla 12, se presentan los resultados del experimento bajo la configuración con selección de características. Las celdas en rojo indican una degradación del rendimiento respecto a la configuración original.

Cuadro 12: [Antenas] Resultados con Configuración *Feature Selection (FS)*

Clase	**Random Forest**			**Regresión Logística**			**Regresión Ridge (Logística)**			**Regresión Lasso (Logística)**			**SNN**		
	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.
1	0.25	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
2		0	0		1	0.5		1	0.67		1	0.5		1	0.5
3		1	0.33		0	0		0	0		0	0		0	0

8.2.3. Resultados con FS + Sobre-muestreo (SMOTE)

Finalmente, se aplicó la técnica SMOTE sobre los datos previamente seleccionados, con el objetivo de equilibrar la distribución de las clases en el conjunto de entrenamiento.

- Se observaron mejoras sustanciales en el desempeño del modelo **Random Forest**, el cual alcanzó un *recall* del 100 % tanto para las piezas con defectos menores como para aquellas no aptas para sinterizar. Esto indica que el modelo logra detectar correctamente ambas clases. No obstante, la clase de piezas con defectos menores presentó una *precisión* del 67 %, es decir, una de cada tres predicciones positivas para esta clase fue un falso positivo.
- Exceptuando el caso de *Random Forest*, que mejoró su rendimiento, y de la regresión *Ridge* logística, que mantuvo un comportamiento estable a lo largo de las tres configuraciones experimentales, el resto de los modelos mostró una caída en el rendimiento. Esto sugiere una posible sensibilidad al ruido introducido por el sobremuestreo sintético o un sobreajuste al conjunto de entrenamiento balanceado artificialmente.

A continuación, en la tabla 13 , se presentan los resultados del experimento correspondiente a la configuración de selección de características con sobremuestreo (FS + SMOTE). Las celdas en verde indican métricas que mejoraron en comparación a las configuraciones anteriores, mientras que las celdas en rojo señalan una degradación en el rendimiento.

Cuadro 13: [Antenas] Resultados con Configuración *Feature Selection + SMOTER*

Clase	**Random Forest**			**Regresión Logística**			**Regresión Ridge (Logística)**			**Regresión Lasso (Logística)**			**SNN**		
	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.	Acc.	Recall	Prec.
1		0	0		1	0.25		0	0		1	0.25		0	0
2	**0.75**	1	0.67	0.25	0	0	0.5	1	0.67	0.25	0	0	0.25	0	0
3		1	1		0	0		0	0		0	0		1	0.25

8.3. Análisis de Resultados

A pesar del reducido número de observaciones (21 muestras), los modelos evaluados lograron capturar parcialmente patrones en la relación entre los parámetros de impresión y el acabado de las piezas fabricadas con PLA-$BaTiO_3$. Sin embargo, el desbalance de las clases y la baja dimensionalidad del dataset impusieron limitaciones significativas en el desempeño general.

Bajo la configuración *vanilla*, los modelos fueron capaces de identificar únicamente una clase, lo que evidenció problemas de representatividad y cobertura de las clases minoritarias. La selección de características no aportó mejoras sustanciales, e incluso produjo leves degradaciones en el rendimiento de algunas arquitecturas, probablemente debido a la pérdida de variabilidad informativa en un contexto ya limitado en muestras.

Por otro lado, la incorporación de sobremuestreo mediante SMOTE, aplicada sobre los datos filtrados, permitió una mejora notable en el modelo *Random Forest*, que logró capturar correctamente dos de las tres clases con altos niveles de *recall*. El resto de los modelos, sin embargo, mostró una caída en el rendimiento o inestabilidad, posiblemente atribuida a sobreajuste o sensibilidad al ruido sintético introducido.

En conjunto, los resultados evidencian que la predicción del acabado en piezas impresas con materiales experimentales como el PLA-$BaTiO_3$ requiere de conjuntos de datos más amplios y balanceados. En su ausencia, técnicas como el sobremuestreo pueden ser útiles, pero solo ciertos modelos —como Random Forest— parecen robustos frente a las limitaciones impuestas por este tipo de escenarios.

9. Conclusiones y Trabajo Futuro

A lo largo de esta tesis se abordaron dos problemáticas distintas pero complementarias, ambas enmarcadas en contextos donde las limitaciones tecnológicas y de recursos dificultan la toma de decisiones. Por un lado, se exploró el uso de aprendizaje automático para predecir el acabado superficial de piezas impresas en 3D con materiales dieléctricos de bajo costo. Por otro, se analizó la aplicación de modelos de aprendizaje automático para estimar métricas de interacción en equipos de trabajo en proyectos AEC, con el objetivo de facilitar el monitoreo de la colaboración.

Los resultados obtenidos en ambos experimentos demuestran que el uso de técnicas de aprendizaje automático puede ser efectivo para abordar problemas de ingeniería caracterizadas con datos limitados o altamente variables. A pesar de las restricciones en el tamaño de las muestras, los modelos aplicados lograron capturar patrones relevantes, ofreciendo una base sólida para desarrollos futuros.

De forma transversal, se identifican aprendizajes relevantes sobre el comportamiento de los modelos evaluados. En primer lugar, los modelos de tipo *Random Forest* requieren una mayor cantidad de datos para alcanzar un rendimiento competitivo. En condiciones de escasez de datos, modelos lineales o incluso redes neuronales superficiales pueden ofrecer mejores resultados. En segundo lugar, si bien el uso de sobremuestreo puede ser beneficioso en contextos de alto desbalance, su aplicación debe ser evaluada caso a caso. Tal como se observó en el experimento de las comunicaciones, su aplicación no mejoró los resultados y, en algunos casos, los degradó.

En el caso del experimento de las comunicaciones en proyectos AEC, los modelos lograron predecir con éxito métricas estructurales derivadas del análisis de redes sociales, lo cual respalda el enfoque propuesto para realizar monitoreos automáticos de colaboración en equipos multidisciplinarios. La red neuronal superficial (SNN) se destacó en la mayoría de los escenarios, superando a modelos más tradicionales, especialmente en contextos no lineales. No obstante, los modelos lineales también mostraron buen rendimiento tras la selección de características, lo que sugiere que la reducción de ruido y redundancia puede ser crítica en dominios con alta colinealidad. En este sentido, una mejora natural sería explorar técnicas como PCA, que permitirían eliminar colinealidad sin necesidad de descartar variables, aunque a costa de menor interpretabilidad.

En el caso del experimento de las antenas, asociado al acabado de impresiones 3D con filamento PLA-$BaTiO_3$, el reducido volumen de datos limitó el rendimiento de los modelos. Aun así, se evidenció que el uso de sobremuestreo, en combinación con selección de características, permitió que algunos modelos como *Random Forest* alcanzaran un desempeño considerablemente superior, identificando correctamente todas las piezas no aptas para sinterizar. Estos resultados muestran que incluso en contextos de escasa información, el uso estratégico de técnicas de muestreo puede marcar diferencias importantes.

La selección de características, por su parte, se confirma como una herramienta valiosa, aunque su aplicación debe ser cautelosa y guiada por el conocimiento del dominio. En el experimento de las comunicaciones, el análisis del factor de inflación de varianza permitió identificar variables con alta colinealidad. Inicialmente, este análisis sugería la eliminación de múltiples variables; sin embargo, al remover únicamente aquella con el VIF más elevado, se logró una reducción significativa de la multicolinealidad sin comprometer la dimensionalidad del conjunto. En contraste, en el experimento de las antenas, la exclusión de una sola variable derivó en una caída del *recall* en algunos modelos, sin evidenciar mejoras en el rendimiento global. Estos resultados refuerzan la idea de que la selección de características no debe aplicarse de forma automática o genérica, sino basada en una comprensión profunda del dominio y la estructura de los datos.

En resumen, los resultados de esta tesis validan el potencial del aprendizaje automático como herramienta de apoyo en la toma de decisiones en ingeniería, brindando herramientas prácticas para anticipar problemas de colaboración y mejorar procesos de fabricación, todo ello sobre una base técnica robusta, accesible y con potencial de escalabilidad.

Asimismo, se identifica como área de mejora la validación estadística de los modelos en ambos casos. En particular, el análisis sistemático de los residuos mediante pruebas como Durbin-Watson y Shapiro-Wilk permitiría confirmar el cumplimiento de supuestos clásicos, fortaleciendo la robustez de las conclusiones.

Finalmente, como trabajo futuro, se propone:

- Recolectar un mayor volumen de datos, especialmente en el caso del problema de impresión 3D, lo cual permitiría aplicar modelos más complejos y evitar sobreajuste.
- Explorar otras arquitecturas de redes neuronales o bien modelos ensamblados (como *Gradient Boosting*), que podrían superar las limitaciones observadas en los modelos actuales.
- Incorporar nuevas variables explicativas, tanto en el ámbito de las comunicaciones en proyectos AEC (e.g., características organizacionales, dinámicas temporales), como en el caso de impresión 3D con PLA-$BaTiO_3$ (e.g., humedad del ambiente, formato del filamento).

Referencias

[1] Alarcón, D.M., Alarcón, I.M., Alarcón, L.F.: Social network analysis: A diagnostic tool for information flow in the aec industry. In: Proceedings for the 21st Annual Conference of the International Group for Lean Construction. pp. 947–956 (2013)

[2] Biau, G., Scornet, E.: A random forest guided tour. Test **25**, 197–227 (2016)

[3] Branco, P., Torgo, L., Ribeiro, R.P.: Smogn: a pre-processing approach for imbalanced regression. In: First international workshop on learning with imbalanced domains: Theory and applications. pp. 36–50. PMLR (2017)

[4] Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

[5] Breiman, L., Cutler, A.: Setting up, using, and understanding random forests V4.0. University of California, Berkeley. Department of Statistics (2003)

[6] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. CRC Press (1984)

[7] Díaz-Uriarte, R., Alvarez de Andrés, S.: Gene selection and classification of microarray data using random forest. BMC bioinformatics **7**, 1–13 (2006)

[8] El Mounla, K., Beladjine, D., Beddiar, K., Mazari, B.: Lean-bim approach for improving the performance of a construction project in the design phase. Buildings **13**(3), 654 (2023)

[9] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**, 1157–1182 (2003)

[10] Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-Hill/Irwin, 5 edn. (2004)

[11] Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-Hill/Irwin, New York, 5th edn. (2005), véase pp. 409–412 para Factor de Inflación de Varianza (VIF)

[12] Minsky, M., Papert, S.: Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge, MA, expanded edn. (1988). https://doi.org/10.7551/mitpress/11301.001.0001, chapter 3

[13] Montgomery, D.C., Peck, E.A., Vining, G.G.: Simple linear regression. In: Introduction to Linear Regression Analysis, chap. 2, p. 12. John Wiley & Sons, 5th edn. (2012)

[14] O'Brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Quality & Quantity **41**(5), 673–690 (2007). https://doi.org/10.1007/s11135-006-9018-6

[15] Ortega, C.: Questionpro - investigacion aplicada (-), https://www.questionpro.com/blog/es/investigacion-aplicada/, consultado el 19 de Abril de 2024

[16] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)

[17] Scheinerman, E.R.: Mathematical Notation: A Guide for Engineers and Scientists. CreateSpace Independent Publishing Platform (2011)

[18] Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P.: Smote for regression. In: Portuguese conference on artificial intelligence. pp. 378–389. Springer (2013)

[19] Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural computing and applications **24**(1), 175–186 (2014)

[20] Wu, W., Kunz, N., Branco, P.: Imbalancedlearningregression-a python package to tackle the imbalanced regression problem. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 645–648. Springer (2022)

[21] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)

[22] Zhang, L., Ashuri, B.: Bim log mining: Discovering social networks. Automation in construction **91**, 31–43 (2018)

A. Anexos

A.1. Anexo 1

Altura de capa [mm]	Velocidad Impresión [mm/s]	Tiempo de Impresión [min]	Diametro [mm]	Altura [mm]	%aditivo	Categoria
0.16	4	34	14.68	5.142	2%	3
0.16	2	38	15.24	5.562	2%	1
0.16	2	42	15.342	6.434	2%	1
0.32	8	16	15.025	N.D	2%	3
0.16	2	38	14.783	5.192	2%	2
0.16	4	37	15.099	5.26	2%	2
0.24	4	32	14.726	5.365	2%	3
0.16	2	38	14.833	4.838	2%	3
0.24	4	36	14.766	4.683	2%	3
0.32	4	28	15.062	5.559	2%	3
0.16	2	38	14.709	5.204	2%	3
0.32	4	32	15.709	5.944	2%	1
0.16	2	36	14.896	5.39	3%	2
0.16	2	36	14.92	4.5	3%	1
0.32	4	32	17.321	6.292	3%	3
0.16	3	52	128.75	10.135	1%	2
0.16	4	48	127.45	10.484	1%	2
0.16	8	44	128.22	10.8	2%	3
0.16	4	48	127.15	10.11	2%	3
0.16	4	48	128.74	10.114	3%	2

Figura 16: Conjunto de datos segunda tanda de piezas cilíndricas

A.2. Anexo 2

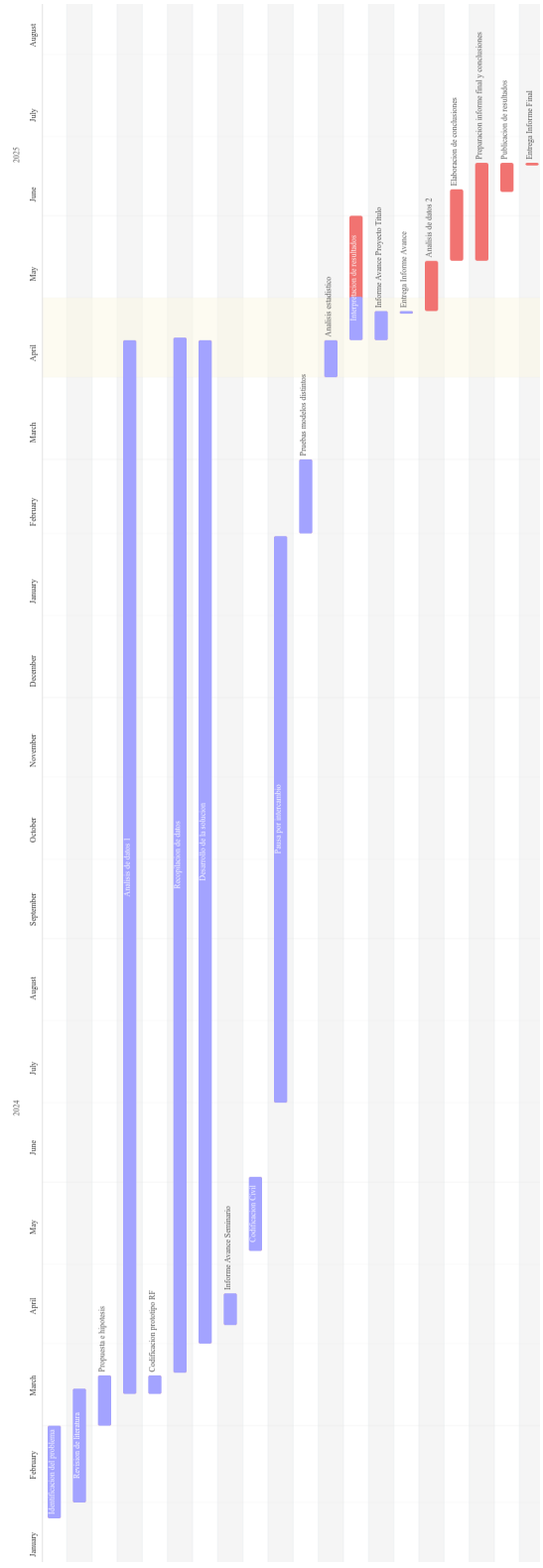

Figura 17: Carta gantt de la planificación con sus etapas