UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías, Politécnico

Predicción de Actividades Productivas Empresariales a través de un Enfoque Basado en Relatedness Density y Registros administrativos a nivel de empresa del Servicio de Rentas Internas (SRI)

Gustavo René Terán Mina

Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito

para la obtención del título de

Ingeniero en Ciencias de la Computación

Quito, 1 de mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías, Politécnico

HOJA DE CALIFICACIÓN

DE TRABAJO DE FIN DE CARRERA

Predicción de Actividades Productivas Empresariales a través de un Enfoque Basado en Relatedness Density y Registros administrativos a nivel de empresa del Servicio de Rentas Internas (SRI)

Gustavo René Terán Mina

Nombre del profesor, Título académico **Pablo Andrés Astudillo Estévez, PhD**

Quito, 1 de mayo de 2025

© DERECHOS DE AUTOR

Nombres y apellidos: Gustavo René Terán Mina

Código: 00324422

Cédula de identidad: 1003998570

Lugar y fecha: Quito, 1 de mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La predicción de actividades productivas empresariales es fundamental para la planificación económica y la formulación de políticas públicas basadas en datos. Este trabajo propone un enfoque basado en Relatedness Density y técnicas avanzadas de Machine Learning para anticipar la aparición de nuevas actividades económicas pioneras en los cantones del Ecuador, como una forma de modelar la evolución de la productividad territorial. Se emplean datos administrativos del Servicio de Rentas Internas (SRI) correspondientes al período 2007-2015. Los modelos desarrollados, entre ellos XGBoost y Continuous Projection Space (CPS), permiten identificar patrones complejos y predecir oportunidades de diversificación productiva. Los resultados destacan la utilidad de estos enfoques para anticipar transformaciones en la estructura empresarial y orientar decisiones estratégicas. Finalmente, se discuten las implicaciones del estudio en términos de desarrollo económico y posibles líneas de mejora futura.

Palabras clave: Predicción económica, Relatedness Density, Machine Learning, XGBoost, Continuous Projection Space, especialización productiva, diversificación empresarial, minería de datos, análisis económico.

ABSTRACT

The prediction of productive business activities is essential for economic planning and evidence-based policy-making. This study proposes an approach based on Relatedness Density and advanced Machine Learning techniques to anticipate the emergence of new pioneering economic activities across Ecuadorian cantons, as a way to model the evolution of local productivity. Administrative data from Ecuador's Internal Revenue Service (SRI) covering the period 2007–2015 is employed. The developed models, including XGBoost and Continuous Projection Space (CPS), enable the identification of complex sectoral patterns and prediction of diversification opportunities. Results demonstrate the effectiveness of these methods in anticipating changes in business structures and supporting strategic decision-making. The study concludes with implications for economic development and future directions to improve predictive performance.

Key words: Economic prediction, Relatedness Density, Machine Learning, XGBoost, Continuous Projection Space, productive specialization, business diversification, data mining, economic analysis.

TABLA DE CONTENIDO

Introducción **10**

Estado del Arte **12**

Descripcion de la propuesta **15**

Desarrollo del modelo predictivo **18**

Experimentos y análisis de resultados **25**

Conclusiones y trabajo futuro **34**

Referencias bibliográficas **36**

ÍNDICE DE TABLAS

Tabla 1: Resultados de F1-score y precisión para cada modelo probado 31

ÍNDICE DE FIGURAS

Ilustración 1: Matriz de Tax total correspondiente al año 2007 17

Ilustración 2: Matriz de descripción de actividades económicas.....17

Ilustración 3: Matriz Binaria de RCA correspondiente al año 2007.....18

Ilustración 4: Matriz de Relatedness Density correspondiente al año 2010 18

Ilustración 5: Matriz de productos nuevos entre el año 2012 y 2013 19

Ilustración 6: Distribución de instancias antes y después del balanceo.....20

Ilustración 7: Distribución de clases antes vs después de usar SMOTE.....21

Ilustración 8: Entrenamiento vs Validacion XGBoost sin SMOTE 23

Ilustración 9: Entrenamiento vs Validación de XGBoost con UMAP 28

Ilustración 10: Entrenamiento vs validación XGBoost + SMOTE.....33

INTRODUCCIÓN

En un mundo cada vez más dinámico y competitivo, la capacidad de anticipar cambios en la estructura productiva de un país se ha vuelto una necesidad clave tanto para la planificación económica como para la toma de decisiones empresariales. En este contexto, la predicción de actividades productivas empresariales surge como una herramienta fundamental para mejorar la competitividad y diseñar estrategias de crecimiento sostenibles.

Este estudio busca aprovechar la métrica de Relatedness Density y técnicas avanzadas de Machine Learning para predecir la aparición de nuevas actividades económicas en los cantones del país, como una forma de observar y modelar la evolución de la productividad.

El interés en este tema proviene de la creciente necesidad de modelos más precisos y detallados que permitan comprender la dinámica empresarial a partir de datos históricos. La técnica de Relatedness Density, utilizada ampliamente en estudios de complejidad económica, facilita la identificación de oportunidades de diversificación y crecimiento al analizar la proximidad entre sectores productivos. En combinación con algoritmos de aprendizaje automático, se pueden capturar relaciones complejas y mejorar la capacidad predictiva de los modelos tradicionales.

En el contexto ecuatoriano, este estudio tiene un impacto relevante, ya que aporta herramientas analíticas para la formulación de políticas públicas basadas en evidencia y la toma de decisiones estratégicas en el sector privado. Como una economía emergente, Ecuador enfrenta retos significativos en la diversificación de su estructura productiva y la implementación de estrategias basadas en análisis de datos. A través del uso de registros administrativos del Servicio de Rentas Internas (SRI) entre 2007 y 2015, este estudio busca generar información clave para identificar patrones de especialización y diversificación productiva en el país. La variable que se busca predecir es binaria: indica si un determinado

cantón adopta una nueva actividad económica (CIIU4) en un año específico. Las variables explicativas utilizadas derivan exclusivamente de los registros administrativos del SRI e incluyen Relatedness Density y transformaciones reducidas de las matrices RCA. Esto permite capturar indirectamente la trayectoria productiva y el grado de afinidad entre sectores.

Para una mejor comprensión del trabajo, es importante definir algunos conceptos fundamentales:

- **Relatedness Density**: Métrica que mide la afinidad entre actividades económicas y permite predecir sectores con mayor potencial de crecimiento en un territorio.
- **XGBoost**: Algoritmo de aprendizaje automático basado en árboles de decisión que utiliza técnicas de bagging y boosting para mejorar su capacidad predictiva.
- **Continuous Projection Space (CPS)**: Técnica de aprendizaje automático que transforma los modelos en un espacio continuo de proyección, facilitando su interpretación y mejorando su precisión de predicción.

Este documento se estructura de la siguiente manera: en la sección **Estado del Arte**, se presenta una revisión de estudios previos sobre complejidad económica y modelos de predicción. Posteriormente, en la **Descripción de la Propuesta**, se detallan los métodos y modelos empleados en este estudio. Luego, en la sección **Desarrollo del Modelo Predictivo**, se explican los procedimientos seguidos para implementar las técnicas propuestas. A continuación, en **Experimentos y Análisis de Resultados**, se presentan los experimentos realizados y la evaluación de los modelos. Finalmente, en la sección **Conclusiones y Trabajo Futuro**, se discuten los hallazgos obtenidos y se sugieren líneas de investigación para mejorar el enfoque propuesto.

ESTADO DEL ARTE

En esta sección se desarrolla el tema elegido para este trabajo final, asegurando el rigor académico mediante referencias a la bibliografía utilizada. Se presentan los antecedentes conceptuales y metodológicos de la predicción de actividades productivas empresariales, destacando su relevancia en la planificación económica y la formulación de políticas públicas.

A. Complejidad Económica y Relatedness Density

La Teoría de Complejidad Económica ha sido ampliamente utilizada para analizar la evolución productiva y la especialización económica de diferentes países. En particular, el concepto de Relatedness Density permite cuantificar la proximidad entre actividades económicas, lo que facilita la identificación de sectores con mayor potencial de crecimiento [1]. Hidalgo (2021) resalta que la complejidad económica no solo mide la diversificación productiva de una región, sino que también permite identificar oportunidades estratégicas basadas en capacidades productivas latentes [1].

El estudio de Tacchella et al. [4] introduce un nuevo enfoque basado en aprendizaje automático para medir la Relatedness, incorporando modelos basados en árboles de decisión y técnicas de proyección continua como el Continuous Projection Space (CPS), que han demostrado una mejora significativa en la predicción de patrones de diversificación industrial y especialización económica.

B. Modelos de Predicción Económica

Las técnicas de aprendizaje automático han sido implementadas en diversos estudios para la predicción del desempeño empresarial y la planificación económica. Afolabi et al. [2] desarrollaron un modelo basado en Naïve Bayes y J48 para predecir el éxito empresarial, demostrando que los modelos de clasificación pueden mejorar la toma de decisiones

estratégicas. Por su parte, Provost y Fawcett [3] enfatizan la importancia del data-driven decision making (DDD) en la planificación económica, resaltando cómo el análisis de big data permite descubrir patrones ocultos en los mercados.

Tacchella et al. [4] proponen el uso de XGBoost como una alternativa más avanzada, basada en técnicas de boosting y bagging, que permite capturar patrones complejos de interacción entre sectores productivos. Además, presentan el método CPS, el cual permite representar las relaciones entre actividades económicas en un espacio continuo, mejorando la capacidad de predicción y la interpretabilidad de los resultados.

Además, Birchha et al. [7] aplicaron XGBoost en el análisis predictivo empresarial, logrando una precisión del 91.5% en la predicción de la pérdida de clientes, lo que evidencia su eficacia en contextos empresariales diversos. Este estudio destaca cómo el uso de XGBoost permite a las empresas adoptar estrategias proactivas basadas en datos para mejorar la retención de clientes y la toma de decisiones estratégicas.

C. Impacto de la Complejidad Económica en el Desarrollo Empresarial

La aplicación de la teoría de complejidad económica en distintos contextos ha demostrado su relevancia para el desarrollo empresarial. Hidalgo [1] señala que la complejidad económica es un predictor clave del crecimiento económico a largo plazo. Además, Provost y Fawcett [3] destacan que la proximidad entre industrias influye significativamente en la probabilidad de diversificación exitosa. Afolabi et al. [2] analizan el uso de técnicas de aprendizaje profundo para modelar la transición entre industrias en economías emergentes.

Tacchella et al. [4] refuerzan estos hallazgos, destacando que el uso de modelos avanzados de aprendizaje automático, como XGBoost y CPS, permite mejorar la precisión en la predicción de cambios en la estructura productiva, ofreciendo nuevas herramientas para la formulación de políticas de desarrollo industrial.

El desarrollo de este trabajo final se fundamenta en estos antecedentes, proporcionando un marco metodológico sólido para la implementación de modelos predictivos basados en datos administrativos de empresas ecuatorianas. A continuación, en la sección Descripción de la Propuesta, se detallarán los enfoques metodológicos empleados en este estudio.

DESCRIPCION DE LA PROPUESTA

Este estudio propone un modelo de predicción de actividades productivas empresariales basado en la métrica de Relatedness Density y el uso de técnicas avanzadas de Machine Learning. La implementación se realiza utilizando **registros administrativos del Servicio de Rentas Internas (SRI) del Ecuador**, los cuales abarcan el período **2007 a 2015**. Estos datos contienen información detallada sobre la facturación anual de **todas las empresas y personas naturales con actividad económica en el país**, permitiendo construir matrices anuales de productos por cantón. El objetivo es predecir la aparición de actividades económicas pioneras (nuevas en el cantón) como una expresión de evolución productiva local.

La base de datos original del año 2015, por ejemplo, contiene más de 13 millones de registros (aproximadamente 13.8 millones de filas y 6 columnas), lo que evidencia el nivel de granularidad disponible para el análisis. Para fines ilustrativos, se incluyen capturas representativas del formato y estructura de estas bases de datos en el documento.

El enfoque metodológico consta de las siguientes fases:

- **Análisis y procesamiento de datos:** Se recopilarán y limpiarán los registros administrativos del SRI, identificando variables clave que impactan la evolución de las actividades productivas. Se aplicarán técnicas de minería de datos y preprocesamiento avanzado para garantizar la calidad de los datos utilizados. La variable dependiente se define como binaria (0/1), representando si un producto aparece por primera vez en un cantón determinado respecto al año anterior. Las variables independientes se derivan exclusivamente de información generada a partir de los registros administrativos del SRI, incluyendo la Relatedness Density y componentes reducidos de las matrices RCA.

- **Desarrollo del modelo predictivo:** Se implementarán modelos basados en Relatedness Density y aprendizaje automático. Los principales modelos escogidos serán XGBoost y Continuous Projection Space (CPS). XGBoost, un algoritmo basado en árboles de decisión con técnicas de bagging y/o boosting, permitirá capturar relaciones complejas entre actividades productivas. CPS, por su parte, permitirá representar las relaciones entre actividades económicas en un espacio continuo, facilitando la interpretación de los resultados.
- **Evaluación y validación de los modelos:** Se utilizarán técnicas de validación cruzada y estudios de caso para evaluar el desempeño de los modelos predictivos. Se medirán métricas de precisión, recall y F1-score para determinar su capacidad de predicción. Además, se analizará el impacto potencial de los resultados en la toma de decisiones estratégicas y en la formulación de políticas públicas.

El principal aporte de este trabajo radica en la integración de técnicas de complejidad económica con modelos avanzados de aprendizaje automático para mejorar la planificación económica y empresarial en Ecuador. Al predecir la incorporación de actividades productivas nuevas en cada cantón, se busca anticipar la evolución de la productividad regional y proporcionar herramientas valiosas para la toma de decisiones estratégicas en el ámbito empresarial y gubernamental.

	A	B	C
1	id_inf	id_prov	total_tax
2	1000135	1797029	414.85
3	1000135	1798069	19.76
4	1000135	1923186	302.73
5	1000135	2502927	1262.86
6	1000135	2504288	155.04
7	1000135	2516113	76.7
8	1000135	2546937	5690.52
9	1000135	2588864	209.15
10	1000135	2615928	111.65
11	1000135	2733484	169
12	1000135	2835267	675.83
13	1000135	2839067	5633.71
14	1000135	2844573	300.28
15	1000135	2844615	42.2
16	1000135	2844709	3.88
17	1000135	2847162	2742.46
18	1000135	2847661	151.43
19	1000135	2848491	337.98
20	1000135	2848965	890.93

Ilustración 1: Matriz de Tax total correspondiente al año 2007

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	oblig	clase_con	contrib_a	ciiu4n4	city_code	anio	ciiu_4n1	descrip_n	ciiu_4n2	descrip_n	ciiu_4n3	descrip_n	ciiu_4n4
2	174290	S	ESP	22204	O8411	20101		O	Public ad	O84	Public ad	O841	Administr	O8411
3	174291	S	ESP	222021	O8411	20101		O	Public ad	O84	Public ad	O841	Administr	O8411
4	174292	S	OTR	222021	O8411	20102		O	Public ad	O84	Public ad	O841	Administr	O8411
5	174293	S	ESP	222021	O8411	20103		O	Public ad	O84	Public ad	O841	Administr	O8411
6	174294	S	ESP	222021	O8411	20105		O	Public ad	O84	Public ad	O841	Administr	O8411
7	174295	S	ESP	222021	O8411	20108		O	Public ad	O84	Public ad	O841	Administr	O8411
8	174296	S	ESP	222021	O8411	20109		O	Public ad	O84	Public ad	O841	Administr	O8411
9	174297	S	OTR	222021	O8411	20107		O	Public ad	O84	Public ad	O841	Administr	O8411
10	174298	S	ESP	222021	O8411	20104		O	Public ad	O84	Public ad	O841	Administr	O8411
11	174299	S	OTR	22401	E3600	20101		E	Water sup	E36	Water coll	E360	Water coll	E3600
12	174300	S	ESP	223012	P8530	20101		P	Education	P85	Education	P853	Higher ed	P8530
13	174301	S	OTR	223013	S9499	20101		S	Other serv	S94	Activities	S949	Activities	S9499
14	174302	S	OTR	223	O8411	20101		O	Public ad	O84	Public ad	O841	Administr	O8411
15	174303	S	OTR	222021	O8411	20106		O	Public ad	O84	Public ad	O841	Administr	O8411
16	174304	S	OTR	223	O8411	20101		O	Public ad	O84	Public ad	O841	Administr	O8411
17	174305	S	OTR	223	Q8620	20101		Q	Human he	Q86	Human he	Q862	Medical a	Q8620
18	174306	S	OTR	222021	O8411	20110		O	Public ad	O84	Public ad	O841	Administr	O8411
19	174307	S	OTR	222021	O8411	20111		O	Public ad	O84	Public ad	O841	Administr	O8411
20	174308	S	OTR	222021	O8411	20112		O	Public ad	O84	Public ad	O841	Administr	O8411

Ilustración 2: Matriz de descripción de actividades económicas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	location	A0111	A0112	A0113	A0114	A0115	A0116	A0119	A0121	A0122	A0123	A0126	A0127	A0129
2	10701	0	0	0	0	0	0	0	0	0	0	0	0	0
3	10702	0	0	0	0	0	0	0	0	0	0	0	0	0
4	10703	0	0	0	0	0	0	0	0	0	0	0	0	0
5	10704	0	0	0	0	0	0	0	0	0	0	0	0	0
6	10705	0	0	0	0	0	0	0	0	0	0	0	0	0
7	10706	0	0	0	1	0	0	0	0	1	0	0	0	0
8	10707	0	0	0	0	0	0	0	0	0	0	0	0	0
9	10708	0	0	0	0	0	0	0	0	0	0	0	0	0
10	10709	0	0	0	0	0	0	0	0	1	0	0	0	0
11	10710	0	0	0	0	0	0	0	0	0	0	0	0	0
12	10711	0	0	0	0	0	0	0	0	0	0	0	0	0
13	10712	0	0	0	0	0	0	0	0	1	0	0	0	0
14	10713	0	0	0	0	0	0	0	0	0	0	0	0	0
15	10714	0	0	0	0	0	0	0	0	0	0	0	0	0
16	10801	0	0	0	0	0	0	0	0	0	0	0	0	0
17	10802	0	0	0	0	0	0	0	0	0	0	1	1	0
18	10803	0	0	0	0	0	0	0	0	0	0	0	0	1
19	10804	0	0	0	0	0	0	0	0	0	0	1	0	0
20	10805	0	0	0	0	0	0	0	0	0	0	1	0	0

Ilustración 3: Matriz Binaria de RCA correspondiente al año 2007

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	activity	A0111	A0112	A0113	A0114	A0115	A0116	A0119	A0121	A0122	A0123	A0125	A0126
2	A0111	0	0.076923	0.076923	0	0.1538461	0	0.2307692	0.076923	0.125	0	0	0.3571428
3	A0112	0.076923	0	0	0.1666666	0	0	0	0	0.1666666	0	0	0
4	A0113	0.076923	0	0	0	0	0	0.3333333	0	0	0	0	0
5	A0114	0	0.1666666	0	0	0.083333	0	0	0	0.25	0	0	0.1428571
6	A0115	0.1538461	0	0	0.083333	0	0	0	0	0.2083333	0	0	0.2142857
7	A0116	0	0	0	0	0	0	0	0	0.0416666	0	0	0.071429
8	A0119	0.2307692	0	0.3333333	0	0	0	0	0.083333	0	0	0	0
9	A0121	0.076923	0	0	0	0	0	0.083333	0	0	0	0	0.071429
10	A0122	0.125	0.1666666	0	0.25	0.2083333	0.0416666	0	0	0	0	0	0.1666666
11	A0123	0	0	0	0	0	0	0	0	0	0	0	0
12	A0125	0	0	0	0	0	0	0	0	0	0	0	0
13	A0126	0.3571428	0	0	0.1428571	0.2142857	0.071429	0	0.071429	0.1666666	0	0	0
14	A0127	0.1111111	0.1666666	0	0.055556	0.055556	0	0.055556	0	0.125	0	0	0.1111111
15	A0128	0	0	0	0	0	0	0.083333	0	0	0	0	0
16	A0129	0.1538461	0	0.3	0.083333	0.2	0.1	0.25	0	0.125	0	0	0.2142857
17	A0130	0.076923	0	0.4285714	0	0	0	0.3333333	0.1666666	0	0	0	0
18	A0141	0.1212121	0	0.090909	0.060606	0	0.060606	0.1515151	0.0303030	0.090909	0	0	0.1212121
19	A0142	0	0	0	0	0	0	0.083333	0	0	0	0	0
20	A0143	0	0	0	0	0	0	0	0	0	0	0	0

Ilustración 4: Matriz de Relatedness Density correspondiente al año 2010

DESARROLLO DEL MODELO PREDICTIVO

El presente trabajo se centra en la construcción de un modelo predictivo capaz de anticipar la aparición de nuevos productos (clase 1) a partir de bases de datos binarias. La

problemática surge del desbalance inherente en los datos, ya que la mayoría de las observaciones corresponden a la clase 0 (ausencia de nuevos productos).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	location	A0111	A0112	A0113	A0114	A0115	A0116	A0119	A0121	A0122	A0123	A0125	A0126
2	10701	0	0	0	0	0	0	0	0	0	0	0	0
3	10702	0	0	0	0	0	0	0	0	0	0	0	0
4	10703	0	0	0	0	0	0	0	0	0	0	0	0
5	10704	0	0	0	0	0	0	0	0	0	0	0	0
6	10705	0	0	0	0	0	0	0	0	0	0	0	0
7	10706	0	0	0	0	0	0	0	0	0	0	0	0
8	10707	0	0	0	0	0	0	0	0	0	0	0	0
9	10708	0	0	0	0	0	0	0	0	0	0	0	0
10	10709	0	0	0	0	0	0	0	0	0	0	0	0
11	10710	0	0	0	0	0	0	0	0	0	0	0	0
12	10711	0	0	0	0	0	0	0	0	0	0	0	0
13	10712	0	0	0	0	0	0	0	0	0	0	0	0
14	10713	0	0	0	0	0	0	0	0	0	0	0	0
15	10714	0	0	0	0	0	0	0	0	0	0	0	0
16	10801	0	0	0	0	0	0	0	0	0	0	0	0
17	10802	0	0	0	0	0	0	0	0	0	0	0	1
18	10803	0	0	0	0	0	0	0	0	0	0	0	0
19	10804	1	0	0	0	0	0	0	0	0	0	0	0
20	10805	0	0	0	0	0	0	0	0	0	0	0	0

Ilustración 5: Matriz de productos nuevos entre el año 2012 y 2013

Esta investigación se fundamenta en la teoría de la complejidad económica [1] y en métodos de machine learning que permiten la extracción de relaciones (relatedness) entre actividades, como se expone en *"Relatedness in the era of machine learning"* [4]. Es importante destacar que, a diferencia del enfoque global utilizado en [4] –que opera con datos a nivel de país– nuestro estudio se basa en datos a nivel cantonal de un único país, situación que exacerba el problema del desbalance de clases.

Preprocesamiento de Datos

El preprocesamiento de los datos incluyó la imputación de valores faltantes y la normalización de las variables para asegurar la comparabilidad entre características. Además, se implementó la técnica de reducción de dimensionalidad UMAP (Uniform Manifold Approximation and Projection) con 10 componentes como método exploratorio para mitigar el sobreajuste en uno de los modelos evaluados. UMAP es una técnica no lineal que preserva la estructura local y global de los datos, siendo eficaz en la reducción de dimensiones para

conjuntos de datos complejos [7]. Sin embargo, en este estudio, su aplicación no mejoró significativamente el rendimiento del modelo, por lo que se optó por otras estrategias para abordar el problema de sobreajuste.

Estrategias de Balanceo de Clases

Dado el desbalanceo observado en el conjunto de datos, con una predominancia de la clase 0, se exploraron diversas técnicas para equilibrar las clases y mejorar el desempeño de los modelos predictivos. Inicialmente, se aplicaron métodos de submuestreo en modelos sensibles al costo. Sin embargo, la técnica que mostró mejores resultados fue el sobremuestreo mediante SMOTE (Synthetic Minority Over-sampling Technique), que genera nuevas instancias sintéticas de la clase minoritaria basándose en sus vecinos más cercanos [8]. Esta estrategia permitió incrementar la representatividad de la clase minoritaria y mitigar el sesgo del modelo, traduciéndose en mejores métricas de desempeño para ambas clases.

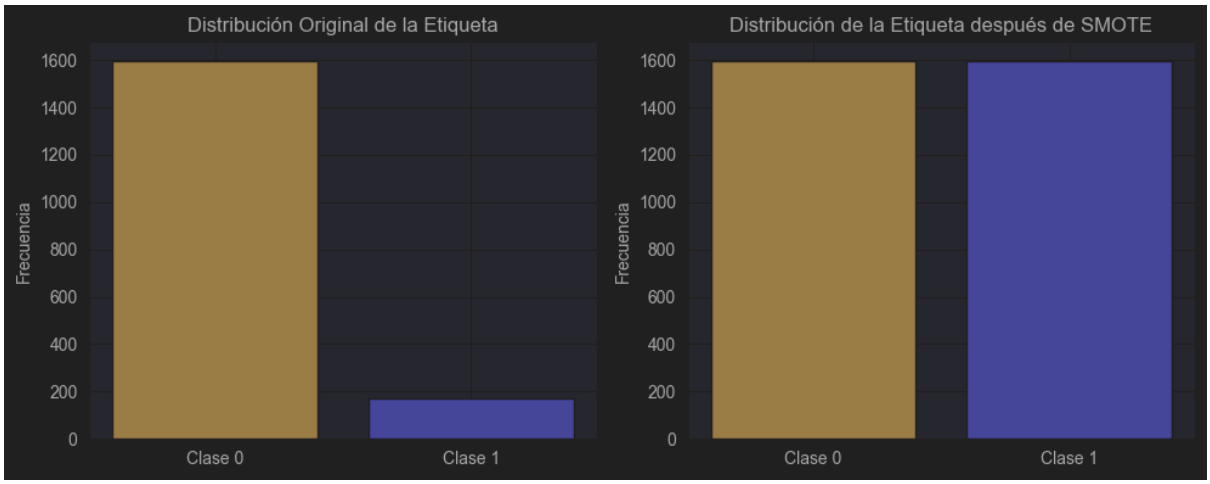

Ilustración 6: Distribución de instancias antes y después del balanceo

Para evaluar la credibilidad de las instancias sintéticas generadas por SMOTE, se aplicó una visualización basada en t-SNE, que permite representar la distribución espacial de las clases en un espacio bidimensional.

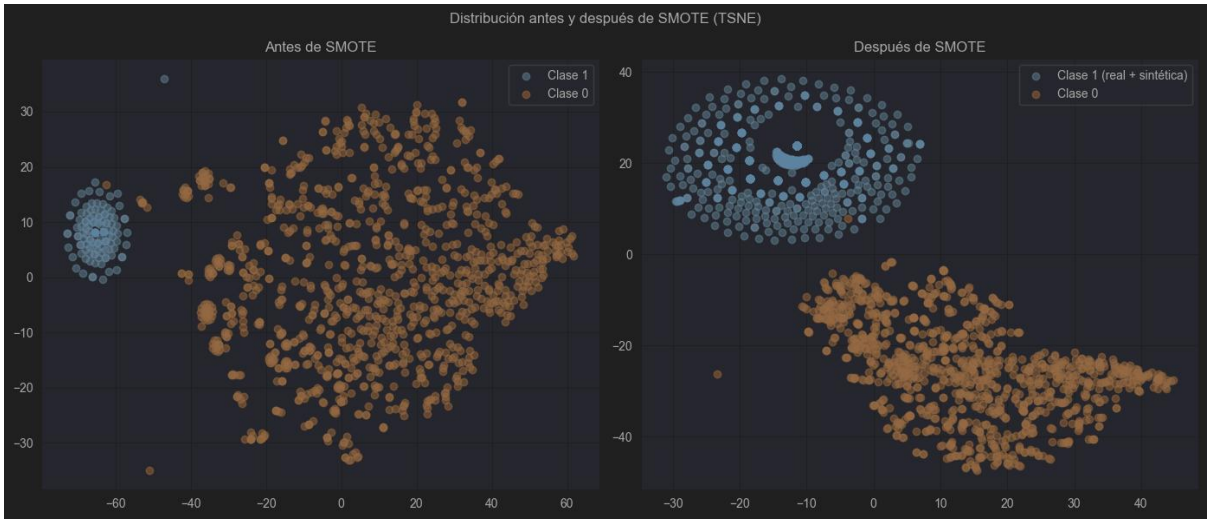

Ilustración 7: Distribución de clases antes vs después de usar SMOTE

En el gráfico, se observa que antes de aplicar SMOTE, la clase minoritaria (clase 0) está fuertemente agrupada y claramente separada de la clase mayoritaria (clase 1). Tras aplicar SMOTE con ***k_neighbors=3***, las nuevas instancias sintéticas de la clase 1 se posicionan de forma coherente dentro del espacio ocupado originalmente por dicha clase, manteniendo su estructura y sin invadir regiones pertenecientes a la clase 0.

Esta separación bien definida entre ambas clases sugiere que las muestras generadas no distorsionan el espacio de representación, sino que refuerzan la densidad de la clase minoritaria. Adicionalmente, se introdujo un leve ruido a las instancias sintéticas para evitar una sobreconcentración artificial, lo que contribuye a una distribución más realista.

Estos resultados respaldan la validez de los datos generados por SMOTE y su utilidad como complemento fiable para el entrenamiento.

Desarrollo de los Modelos Predictivos

1. **Modelos Basados en Árboles de Decisión**

   1.1. XGBoost con SMOTE (Modelo Base): Para mitigar el sesgo derivado del predominio de la clase 0, se combinó el algoritmo XGBoost con la técnica de oversampling SMOTE. Se realizó una optimización de hiperparámetros mediante RandomizedSearchCV, ajustando parámetros como el número de estimadores, la profundidad máxima, la tasa de aprendizaje, y parámetros de regularización. Este modelo se diseñó para capturar relaciones complejas entre las características reducidas (tras aplicar PCA) y la variable objetivo, logrando que la representación de la clase minoritaria sea adecuada.

   1.2. XGBoost sin Reducción de Dimensionalidad ni Oversampling: En este enfoque se utilizó el mismo algoritmo XGBoost, pero sin aplicar la técnica de reducción de dimensionalidad UMAP. Este modelo se entrenó directamente sobre las características originales de los datos, lo que permite evaluar la importancia de la reducción de dimensionalidad en la mejora del desempeño predictivo. La optimización de hiperparámetros se realizó mediante GridSearchCV, ajustando parámetros críticos para maximizar la precisión en la predicción de las clases.

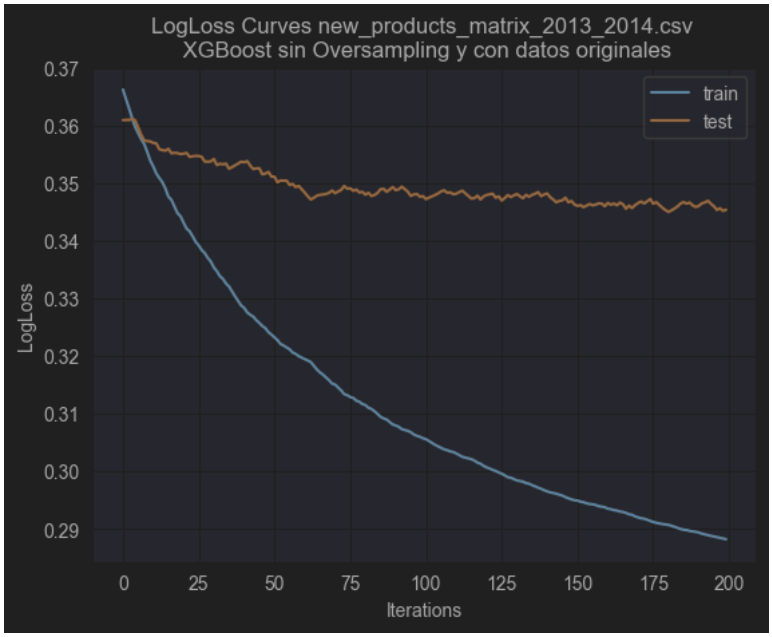

Ilustración 8: Entrenamiento vs Validacion XGBoost sin SMOTE

1.3. **XGBoost con Datos Preprocesados (UMAP):** Se exploró además un modelo XGBoost utilizando datos preprocesados que integran reducción de dimensionalidad mediante UMAP. Este modelo pretende capturar la estructura intrínseca de los datos mediante un mapeo no lineal, que puede favorecer la detección de patrones complejos en contextos de alta variabilidad. La optimización de hiperparámetros se realizó con RandomizedSearchCV.

Nota: Este modelo, pese a ser técnicamente avanzado, presenta desafíos relacionados con el sobreajuste y la interpretación, aspecto que se profundizará en la sección de resultados.

1.4. **Balanced Random Forest (Modelo CPS):** Este modelo se basa en el uso de un clasificador Random Forest cost-sensitive, que internamente balancea la distribución de clases mediante submuestreo de la clase mayoritaria. Se realizó la optimización de hiperparámetros mediante RandomizedSearchCV, ajustando parámetros como el

número de árboles, la profundidad máxima y el método de selección de características. La fortaleza de este modelo radica en su capacidad para evitar el sobreajuste sin necesidad de recurrir a técnicas de oversampling externas, lo que lo hace especialmente adecuado en escenarios con datos altamente desbalanceados.

2. **Modelos Basados en Métodos Lineales y de Gradiente**

   2.1. ElasticNet Logistic Regression: Se implementó un modelo de regresión logística con regularización ElasticNet, que combina penalizaciones L1 y L2 para mejorar la selección de variables. Este enfoque busca optimizar la capacidad predictiva del modelo en contextos de alta dimensionalidad y colinealidad, aunque en este caso se observó una tendencia al sobreajuste que afecta principalmente la clasificación de la clase 0.

   2.2. HistGradientBoostingClassifier: Este modelo utiliza técnicas de boosting adaptadas para conjuntos de datos grandes y de alta complejidad. Se optimizaron hiperparámetros como la profundidad máxima, el número de hojas y la regularización L2. Aunque permite capturar relaciones no lineales complejas, su implementación reveló problemas de sobreajuste en la clasificación de la clase 0, lo que sugiere la necesidad de ajustes adicionales en escenarios de desbalance extremo.

3. **Detección de Anomalías**

   3.1. Modelo Isolation Forest: Se incorporó el algoritmo Isolation Forest como técnica no supervisada para la detección de anomalías en el conjunto de datos. Isolation Forest identifica anomalías aislando observaciones mediante particiones aleatorias en árboles de decisión, siendo eficaz en conjuntos de datos de alta dimensión y con ruido [9]

Consideraciones Metodológicas Comunes

En todos los modelos se aplicó una estrategia de validación cruzada para garantizar la robustez de los modelos ante la alta autocorrelación temporal. Asimismo, se utilizaron métricas de evaluación como el **F1-score, precisión y mAP@20**, que permiten una comparación objetiva sin entrar en detalles de los resultados, ya que estos serán presentados en la siguiente sección de Experimentos y Análisis de Resultados.

Nota adicional: La diferencia en la granularidad de los datos es un aspecto crucial. Mientras que en *"Relatedness in the era of machine learning"* [4] se utilizan datos a nivel global –por país–, lo que facilita la detección de patrones de *relatedness* en un conjunto amplio y diverso, en nuestro estudio los datos a nivel cantonal generan un desbalance aún mayor. Esto refuerza la necesidad de aplicar técnicas de balanceo, como SMOTE o métodos cost-sensitive, para lograr una adecuada representación de la clase minoritaria.

EXPERIMENTOS Y ANÁLISIS DE RESULTADOS

Configuración Experimental

Para evaluar el desempeño de los modelos predictivos se utilizó un esquema de **validación cruzada** tradicional, específicamente una validación cruzada con 3 particiones (K-Fold con $k=3$), implementada mediante **RandomizedSearchCV**. Esta técnica permitió estimar la capacidad de generalización de los modelos y ajustar sus hiperparámetros de manera robusta, evitando sesgos asociados al conjunto de prueba.

Las métricas principales utilizadas fueron el **F1-score, la precisión (Precision) y el mAP@20,** que permiten analizar tanto el rendimiento global como el desempeño en la

predicción de nuevos productos (clase 1) frente a la permanencia de productos existentes (clase 0).

Se realizaron múltiples experimentos para cada modelo, y en el caso de **XGBoost** se analizaron las curvas de aprendizaje (por ejemplo, las curvas de Log Loss) para detectar indicios de *sobreajuste* y evaluar la estabilidad durante el entrenamiento. Adicionalmente, se incluyó **Isolation Forest** como modelo de **detección de anomalías**, entrenado exclusivamente con observaciones correspondientes a productos ya existentes (clase 0).

Aunque Isolation Forest es un algoritmo no supervisado, en este estudio se evaluó su rendimiento en un contexto supervisado gracias a la **disponibilidad de etiquetas reales**. Para ello, se mapearon sus salidas predeterminadas (1 para normales, -1 para anómalos) para interpretar correctamente los resultados frente al objetivo de predecir la aparición de nuevos productos (clase 1).

Esta adaptación permitió calcular métricas supervisadas como el F1-score y comparar de manera coherente su desempeño con el resto de los modelos considerados.

Análisis de Modelos Basados en Árboles de Decisión

1. **XGBoost con SMOTE (Modelo Base)**

Este modelo fue diseñado combinando el algoritmo XGBoost con la técnica de oversampling SMOTE, con el objetivo de equilibrar la alta desproporción entre la clase 0 y la clase 1.

Resultados:

- **Clase 1:** Se obtuvo un F1-score de 1.00, lo que indica que el modelo detecta de manera casi perfecta los nuevos productos.
- **Clase 0:** El F1-score alcanzó 0.99, demostrando que el modelo es capaz de reconocer la ausencia de nuevos productos sin incurrir en un alto número de falsos positivos.

Técnica de aceptación: La robustez se comprobó a través de validación cruzada, y las curvas de pérdida mostraron una convergencia estable sin signos de sobreajuste.

2. **XGBoost sin Reducción de Dimensionalidad**

En este experimento se utilizó XGBoost sin aplicar técnicas de reducción de dimensionalidad, de modo que el modelo trabajó directamente sobre el conjunto de datos original.

Resultados:

- **Clase 1:** Se obtuvo un F1-score promedio de 0.95, lo que indica un rendimiento aceptable en la detección de nuevos productos.
- **Clase 0:** El F1-score descendió a aproximadamente 0.48, lo que evidencia un sesgo considerable hacia la clase 1 y una inadecuada detección de la clase 0.

Técnica de descarte: La comparación de las métricas entre ambas clases y la evidencia de un alto sesgo llevaron a descartar este enfoque en favor de modelos que logran un balance adecuado.

3. **XGBoost con Datos Preprocesados (UMAP)**

Este modelo aplicó UMAP para la reducción no lineal de la dimensionalidad antes de entrenar XGBoost.

Resultados:

1. **Clase 1:** Se obtuvo un F1-score de 0.95, similar al modelo sin reducción.
2. **Clase 0:** El F1-score fue de 0.00, lo que indica que el modelo prácticamente no clasifica correctamente la clase 0, mostrando un comportamiento completamente sesgado hacia la clase 1.

Técnica de descarte: Dado que el modelo no logra identificar la clase 0, se concluyó que la reducción mediante UMAP, en este caso, no aporta beneficios y empeora la capacidad de generalización.

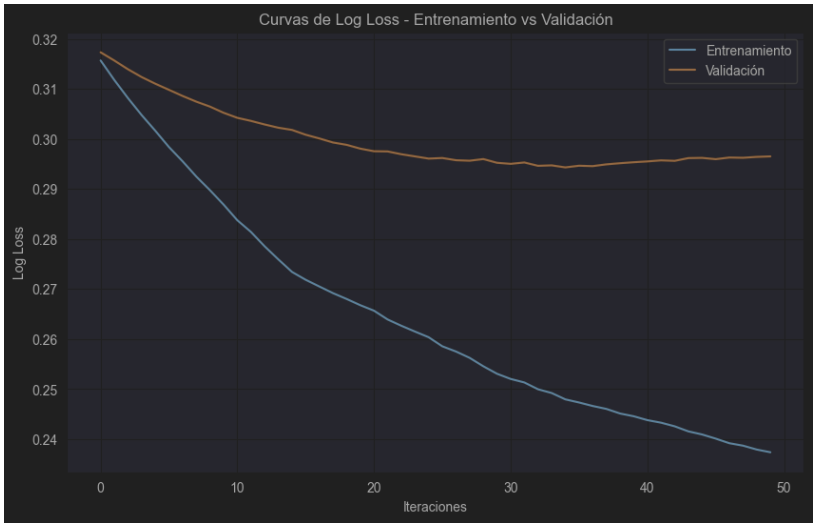

Ilustración 9: Entrenamiento vs Validación de XGBoost con UMAP

4. **Balanced Random Forest (Modelo CPS)**

Se implementó un clasificador Random Forest cost-sensitive, que internamente balancea la distribución de clases mediante submuestreo de la clase mayoritaria.

Resultados:

- **Clase 1:** Se registró un F1-score de 1.00, lo que evidencia una excelente detección de nuevos productos.
- **Clase 0:** Se obtuvo un F1-score de 0.98, demostrando que el modelo clasifica adecuadamente también la ausencia de nuevos productos.

Técnica de aceptación: La consistencia de las métricas en ambos grupos y la estabilidad en la validación cruzada permitieron aceptar este modelo como una alternativa robusta al modelo base.

Análisis de Modelos Basados en Métodos Lineales y de Gradiente

1. **ElasticNet Logistic Regression**

Este modelo combina penalizaciones L1 y L2 para gestionar la alta dimensionalidad y la colinealidad.

Resultados:

- **Clase 1:** Se observó un F1-score de 1.00, indicando una detección casi perfecta de nuevos productos.
- **Clase 0:** Sin embargo, el F1-score fue de 0.00, lo que demuestra un sobreajuste extremo que impide la detección de la clase 0.

Técnica de descarte: La extrema disparidad entre el rendimiento en la clase 1 y la falta de detección en la clase 0 llevó a rechazar este enfoque, pues no cumple con el criterio de balance en ambas clases.

2. **HistGradientBoostingClassifier**

Se aplicó el HistGradientBoostingClassifier para aprovechar la potencia del boosting en grandes conjuntos de datos.

Resultados:

- **Clase 1:** Se alcanzó un F1-score de 1.00, similar a otros modelos de alto rendimiento en la detección de nuevos productos.
- **Clase 0:** No obstante, se obtuvo un F1-score de 0.00, lo que indica que este modelo también sufre de un sobreajuste severo al favorecer la clase mayoritaria.

Técnica de descarte: La incapacidad para clasificar correctamente la clase 0 y el indicio de sobreajuste general llevaron a descartar este modelo en favor de enfoques más equilibrados.

Análisis de Modelo en base a Detección de Anomalías

1. **Isolation Forest**

Incorporó como un modelo de detección de anomalías, diseñado para identificar outliers a través del aislamiento de observaciones inusuales mediante árboles aleatorios. Dado su enfoque no supervisado, se utilizó para evaluar si la exclusión de valores atípicos mejoraba la capacidad predictiva general del sistema.

Resultados:

- **Clase 1:** Se alcanzó un F1 score de **0.95**, con una precisión del **90%**, lo cual indica una detección adecuada de nuevos productos dentro de los casos no considerados anómalos.
- **Clase 0:** Sin embargo, se registró un F1 score de **0.00**, mostrando una completa incapacidad para detectar adecuadamente los casos de no aparición de nuevos productos.

Técnica de descarte: A pesar del buen rendimiento en la clase 1, el modelo falló totalmente en representar la clase 0, lo que refleja un sesgo severo derivado posiblemente del proceso de aislamiento. Esta debilidad en el balance entre clases llevó a descartar su uso como modelo principal, conservando en cambio los enfoques con **XGBoost + SMOTE** y **Balanced Random Forest**, los cuales demostraron resultados robustos y equilibrados.

Tabla 1: Resultados de F1-score y precisión para cada modelo probado

Modelo	**F1-Score (Clase 0)**	**F1-Score (Clase 1)**	**Precisión (Clase 0)**	**Precisión (Clase 1)**
XGBoost + SMOTE (Modelo Base)	0.99	1.00	0.98	1.00
Balanced Random Forest (Modelo CPS)	0.98	1.00	0.98	1.00
Detección de anomalías (Isolation Forest)	0.00	0.95	0.00	0.90
XGBoost sin reducción de dimensionalidad (Modelo 1)	0.48	0.95	0.30	0.91
XGBoost con datos preprocesados (Modelo 5)	0.00	0.95	0.00	0.91
Random Forest (Modelo 2)	0.48	0.95	0.30	0.91
ElasticNet Logistic Regression (Modelo 3)	0.00	1.00	0.00	0.99
HistGradientBoostingClassifier (Modelo 4)	0.00	1.00	0.00	0.99

Criterios de Selección y Validación

Para la evaluación y selección de los modelos predictivos, se establecieron los siguientes criterios, fundamentados en prácticas reconocidas en el ámbito del aprendizaje automático y la estadística:

- **Equilibrio entre Clases:** Dado el desbalance en la distribución de clases del conjunto de datos, se priorizó la evaluación del rendimiento del modelo en ambas clases, utilizando el F1 score para cada una. El F1 score es especialmente útil en contextos de clases desbalanceadas, ya que combina la precisión y la recuperación en una única métrica, proporcionando una visión equilibrada del rendimiento del modelo [10].
- **Generalización y Robustez:** Para asegurar la capacidad de generalización del modelo y evitar el sobreajuste, se empleó la validación cruzada con 3 particiones (K-Fold con $k=3$), implementada mediante **RandomizedSearchCV**. Este método permite evaluar la estabilidad del modelo al entrenarlo y probarlo en diferentes subconjuntos del conjunto de datos, lo que es crucial para obtener estimaciones fiables del rendimiento del modelo en datos no vistos [11].
- **Métricas Complementarias:** Además del F1 score, se utilizaron métricas adicionales como la precisión y el mAP@20. La precisión mide la proporción de verdaderos positivos entre todas las predicciones positivas, lo que es fundamental para entender la exactitud del modelo en la identificación de casos positivos [12]. El mAP@20, por su parte, es una métrica ampliamente utilizada en sistemas de recomendación y recuperación de información, ya que evalúa la precisión promedio de las predicciones en las primeras 20 posiciones, reflejando la relevancia y el orden de los resultados proporcionados por el modelo [13].
- **Interpretabilidad y Complejidad:** La interpretabilidad del modelo es esencial, especialmente en aplicaciones relacionadas con políticas económicas, donde es

importante comprender cómo las variables influyen en las predicciones. Modelos más interpretables permiten una mejor comprensión y confianza en las decisiones basadas en sus resultados.

Resumen del Análisis de Resultados

Modelos Aceptados:

- ***XGBoost con SMOTE (Modelo Base)*** **y** ***Balanced Random Forest (Modelo CPS)***, son los enfoques que muestran resultados equilibrados, con F1-scores cercanos a 1.00 en la clase 1 y alrededor de 0.98 en la clase 0, lo que confirma su robustez y capacidad de generalización.

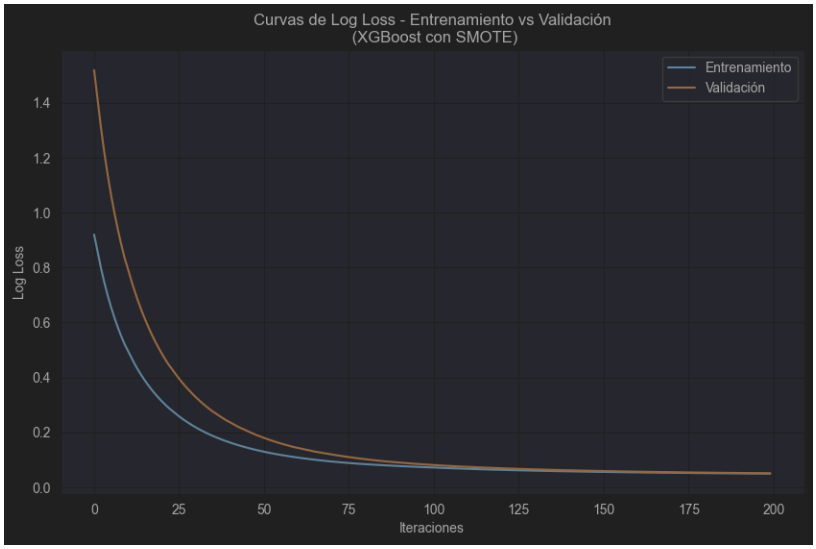

Ilustración 10: Entrenamiento vs validación XGBoost + SMOTE

Modelos Descartados:

- ***XGBoost sin reducción de dimensionalidad*** **y** ***XGBoost con UMAP,*** presentan un rendimiento deficiente en la clase 0 (F1-scores de 0.48 y 0.00, respectivamente), evidenciando un sesgo hacia la clase 1.

- ***ElasticNet Logistic Regression* y *HistGradientBoostingClassifier,*** muestran sobreajuste extremo, ya que logran un F1-score perfecto en la clase 1 pero ignoran completamente la clase 0.
- ***Isolation Forest,*** implementado como modelo no supervisado para la detección de nuevos productos considerados anómalos, fue descartado debido a su bajo desempeño en la identificación correcta de la clase 0. Aunque el modelo logró un F1-score elevado en la clase 1 con un valor de 0.95, su desempeño sobre la clase 0 fue completamente deficiente (F1-score = 0.00), clasificando erróneamente muchas observaciones legítimas como anomalías. A pesar de los ajustes realizados para interpretar sus salidas dentro de un esquema supervisado de clasificación binaria desbalanceada, Isolation Forest no logró un equilibrio adecuado por lo que fue considerado inviable frente a los otros modelos supervisados.

CONCLUSIONES Y TRABAJO FUTURO

Este trabajo demuestra que es posible predecir la aparición de nuevas actividades productivas empresariales en Ecuador mediante un enfoque basado en la métrica de *Relatedness Density* y técnicas de *Machine Learning*. La integración de registros administrativos del SRI con modelos como XGBoost + SMOTE y Balanced Random Forest (CPS) permitió alcanzar un desempeño robusto, equilibrando adecuadamente la detección de ambas clases en un entorno de datos altamente desbalanceados.

A nivel nacional, esta propuesta representa una herramienta estratégica para apoyar la formulación de políticas públicas orientadas al desarrollo productivo. A nivel internacional, se alinea con estudios recientes sobre complejidad económica, pero se diferencia en su

enfoque a escala microeconómica (nivel de empresa y cantón), frente a enfoques tradicionales más agregados (por país).

Durante el desarrollo del proyecto se evidenció la importancia de adaptar los modelos al contexto de los datos, superando retos como el desbalance de clases y el sobreajuste. Asimismo, se observó que técnicas de detección de anomalías como Isolation Forest, aunque prometedoras en otros contextos, no son adecuadas para tareas de clasificación binaria en este tipo de datos, lo que refuerza la necesidad de validación rigurosa.

Entre las principales dificultades enfrentadas se destacan la obtención y depuración de datos administrativos, así como la selección de técnicas adecuadas para lograr interpretabilidad y generalización simultáneamente. Como aprendizaje, se consolidaron competencias en análisis de datos económicos, validación de modelos y adaptación de técnicas complejas al contexto local.

Como líneas de trabajo futuro, se propone:

- **Incorporar** datos más recientes y variables adicionales (como redes empresariales o localización geográfica fina).
- **Explorar** enfoques causales o generativos que complementen los modelos predictivos.
- **Implementar** visualizaciones interactivas que faciliten la interpretación de recomendaciones a nivel institucional.
- **Analizar** la evolución temporal de los patrones de diversificación para detectar cambios estructurales.

Este estudio abre la puerta a futuras investigaciones aplicadas en planificación territorial, políticas de incentivos y monitoreo inteligente del desarrollo productivo en economías emergentes como la ecuatoriana.

REFERENCIAS BIBLIOGRÁFICAS

[1] C. A. Hidalgo, "Economic complexity theory and applications," *Nature Reviews Physics*, vol. 3, no. 2, pp. 92-113, Jan. 2021.

[2] I. Afolabi, T. C. Ifunaya, F. G. Ojo, and C. Moses, "A model for business success prediction using machine learning algorithms," *Journal of Physics: Conference Series*, vol. 1299, no. 1, pp. 1-10, 2019. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1299/1/012034

[3] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51-59, Mar. 2013.

[4] A. Tacchella, A. Zaccaria, M. Miccheli, and L. Pietronero, "Relatedness in the era of machine learning," *Chaos, Solitons and Fractals*, vol. 176, 2023.

[5] V. Stojkoski, C. Hidalgo, "Optimizing Economic Complexity," *Center for Collective Learning, Corvinus University of Budapest*, 2025. [Disponible en: https://www.tse-fr.eu/publications/optimizing-economic-complexity]

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[7] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

[8] A. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[9] C. O'Sullivan, "Guía del Bosque del Aislamiento: Explicación e implementación en Python," Sep. 25, 2024. https://www.datacamp.com/es/tutorial/isolation-forest?utm_source=chatgpt.com

[10] P. Kashyap, "Understanding Precision, Recall, and F1 Score Metrics," *Medium*, 2023. [En línea]. Disponible en: https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

[11] DataCamp, "Guía completa para la validación cruzada K-Fold," *DataCamp*, 2023. [En línea]. Disponible en: https://www.datacamp.com/es/tutorial/k-fold-cross-validation

[12] Google Developers, "Clasificación: Exactitud, recuperación, precisión y métricas relacionadas," *Google*, 2023. [En línea]. Disponible en: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419

[13] Innovatiana, "mAP o 'mean Average Precision', una métrica clave en IA," *Innovatiana*, 2024. [En línea]. Disponible en: https://es.innovatiana.com/post/mean-average-precision