Modelos Predictivos de Defaults Soberanos Utilizando Herramientas de Machine Learning: Una Comparación Basada en Métricas de Desempeño

Universidad de San Andrés

Departamento de Economía

Licenciatura en Economía

Autoras:

Valentina Lucini - Legajo: 30144

María Mercedes García Fagalde - Legajo: 30087

Mentores: Juan Cruz López Del Valle y Belén Michel Torino

Victoria, Provincia de Buenos Aires, Diciembre de 2022

Resumen

Las crisis de deuda soberana se han vuelto cada vez más preeminentes en la opinión pública, principalmente en países en desarrollo. Frente a la situación mundial post pandemia y de la guerra en curso, un número cada vez mayor de economías emergentes están experimentando dificultades financieras, lo que aumenta el riesgo de una ola de defaults soberanos y por lo que cada vez cobra mayor relevancia el hecho de poder realizar una predicción sobre el momento en el cual las naciones entrarán en tales crisis de deuda. El presente trabajo propone una actividad de predicción de defaults de una selección de cinco países quienes resultaron ser aquellos con mayor cantidad de incumplimientos de deuda soberana desde 1970 hasta la actualidad. En este sentido, el trabajo contribuye a la literatura por un lado, proporcionando un dataset para el conjunto de países en cuestión. Para ello, se seleccionaron variables macroeconómicas y de deuda para los países seleccionados; la elección de dichas variables fue resultado de una extensa revisión de la literatura existente sobre las variables más adecuadas y eficaces a la hora de predecir defaults. Por otro lado, el trabajo brinda un exhaustivo análisis de predicción de dichos defaults utilizando modelos de Machine Learning. La propuesta consta de una comparación o "carrera de caballos", del entrenamiento de diversos modelos, entre ellos regresión logit, KNN y árboles de regresión, y a través del cálculo de diversas métricas de desempeño se seleccionó al modelo de Random Forest como aquel que presentó las mejores métricas de exactitud evaluadas.

Índice

1. Introducción

Existen importantes razones que justifican la profunda búsqueda de la existencia de modelos predictivos sobre deuda soberana. Siguiendo a Catao et al (2002), a pesar de que en las últimas décadas se ha podido observar una significativa mejora de los fundamentos macroeconómicos en las economías emergentes, los defaults soberanos han aumentado notablemente. Dado que el impacto sistémico de estos acontecimientos puede ser considerable en un mundo globalizado, es natural que las propuestas para abordar el problema ocupen un lugar destacado en la agenda política multilateral. Sin embargo, mientras se siguen debatiendo las recientes propuestas para rediseñar los contratos de deuda soberana y establecer nuevos mecanismos para promover una reestructuración ordenada de la deuda, la anticipación de la crisis y la prevención temprana siguen siendo cruciales.

Sturzenegger y Zettelmeyer (2006) señalan que los episodios de impago tienden a producirse en clusters y, por lo general, siguen a un boom de préstamos. En la década de 1970, se produjo dicho auge de préstamos hacia países en desarrollo, lo que consecuentemente causó una cadena de impagos. Fueron 15 los episodios observados de impagos de préstamos y a partir de la "crisis de la deuda" con la suspensión de pagos de México en agosto de 1982, siguieron más de 70 defaults (34 episodios de países africanos y 29 de países latinoamericanos). Esto produjo la suficiente evidencia como para crear la necesidad de implementar un sistema de predicción con el poder de prevenir y alertar a las economías nacionales de estos eventos de crisis generales. Este estudio tiene como principal objetivo mejorar las técnicas de predicción de defaults ya existentes, aportando la tecnología de Machine Learning y la producción de un dataset compuesto por variables relevantes para el output en cuestión. Se realizó el análisis para aquellos países que mayor cantidad de ocasiones incurrieron en defaults desde 1970 (datos disponibles) a la actualidad, donde además de Argentina, se encuentran Ecuador con cinco defaults, Paraguay con dos, Bolivia con tres y por último Nigeria con cinco.

La motivación principal del análisis surge como consecuencia de habitar un país como Argentina, donde la palabra “default” es un tema recurrente pero no menor. La secuencia comienza en 1827 pero se exacerbó en los últimos cuarenta años, siendo hoy la novena vez en la que el país se encuentra frente al incumplimiento de su deuda. El último préstamo solicitado al Fondo Monetario Internacional fue en 2018, durante la gestión del Presidente Mauricio Macri donde se acordó un monto de 57 mil millones de dólares, el más alto de los préstamos acordados entre el FMI y el Estado argentino.

Según S&P un emisor es calificado como en incumplimiento de pagos o default si no ha cumplido con alguna o algunas de sus obligaciones financieras y lo considera como incumplimiento general, cuando el deudor no pagará específicamente ninguna o casi ninguna de sus obligaciones dentro de los plazos establecidos en las condiciones del préstamo. Muy a menudo existen las negociaciones

internacionales donde se cancela parcialmente o se realiza una reestructuración de la deuda. Este tipo de acuerdo asegura el reembolso parcial cuando el acreedor acepta una renuncia de una gran parte de la deuda, como es el caso argentino donde el gobierno logró reestructurar su deuda con el FMI por 45.000 millones de dólares y con el Club de París por una deuda de 2.400 millones de dólares.

Diversas son las consecuencias que recaen en un país al dejar en deuda a sus acreedores, siendo la más inmediata el impedimento del acceso a los mercados internacionales de crédito. La literatura teórica asume que los impagos y las reestructuraciones conducen a la exclusión de dichos países de los mercados internacionales de capitales, así como a un aumento de los costos de los préstamos posteriores (Ams. et al., 2019). Los períodos de reestructuración de la deuda suelen ser largos, y mientras no se llega a un acuerdo entre el gobierno y sus acreedores, el país pierde el acceso a los mercados internacionales; como es el ejemplo de Argentina, que luego de su default en 2002 estuvo catorce años sin acceso al financiamiento externo. No solo el gobierno se ve perjudicado sino también los municipios y empresas, que no son ajenos a las consecuencias que supone el no acceso al crédito; si los acreedores no prestan al gobierno, tampoco lo hará a las empresas que operan en el país. Ante esta situación el gobierno se ve obligado a elegir una de las siguientes alternativas: recortar el gasto público, aumentar los impuestos o financiarse a través de la emisión de moneda, todas con fuertes consecuencias en la economía. Por otro lado, la confianza de los agentes en una economía es vital para que esta funcione adecuadamente y luego de una cesación de pagos el aumento de la incertidumbre en la economía es un problema que cobra relevancia. El ambiente de incertidumbre que se genera luego de que un país entre en default, hace que los consumidores e inversores restrinjan sus consumos e inversiones, afectando la actividad económica y consecuentemente generando presión en el mercado cambiario. Borensztein y Panizza (2008), señalan que retrasar los impagos puede ser costoso por tres razones:

1. Las políticas fiscales restrictivas no creíbles son ineficaces para evitar el impago y provocan contracciones de la producción;
2. Puede prolongar el clima de incertidumbre y los altos tipos de interés y, por tanto, tener un efecto negativo en la inversión y los balances de los bancos;
3. Puede tener efectos perjudiciales directos en el sector financiero.

Antes del siglo XIX los defaults eran resultado de eventos extraordinarios pero a partir de entonces se han vinculado en mayor medida a desmanejos financieros (Cabrera, 2014). Las décadas del 70 y 80 fueron protagonistas de tres grandes crisis; crisis mundial del petróleo, crisis del dólar y la de deuda externa latinoamericana. Las dos primeras crearon una elevada inflación en los países en desarrollo mientras que la última dejó a los países latinoamericanos en una posición difícil para hacer frente a los retos que impondría la economía mundial a fines del siglo 20. Agosto de 1982 marca convencionalmente el inicio de la denominada “década perdida” para el desarrollo latinoamericano,

pero esta crisis reconoce sus raíces en sus desequilibrios macroeconómicos y choques externos que se desarrollaron en la década previa, en casi todos los casos el gasto total se expandió por encima del producto. Puntualmente, para los países estudiados, tanto Argentina como Paraguay, Ecuador y Bolivia se enfrentaron a gobiernos cívico militares en la década del 70, acompañado con alza de precios, déficit en sus balanzas de pagos y consecuentemente creciente deuda externa para financiarlos. Por su parte, Nigeria no es la excepción de las consecuencias de la crisis mundial del petróleo y del dólar de esos tiempos. El país comenzó a tener problemas de endeudamiento desde principios de la década de 1980, cuando los ingresos en divisas se desplomaron como consecuencia del colapso de los precios en el mercado internacional del petróleo y los préstamos externos comenzaron a adquirirse indiscriminadamente. Para el período en cuestión, dejó de pagar sus deudas en cinco ocasiones; 1982, 1986, 1992, 2001 y 2004.

Como se mencionó anteriormente, no son pocas las consecuencias que recaen sobre la economía luego de un default. En este sentido, este trabajo contribuye a la literatura por un lado, proporcionando un dataset para el conjunto de países que han atravesado crisis de deuda la mayor cantidad de veces. Para obtener este conjunto de datos se seleccionaron variables de deuda y macroeconómicas para los cinco países en cuestión; la elección de dichas variables fueron resultado de un extensa revisión de la literatura existente sobre cuales son las variables más adecuadas para predecir defaults. Es importante mencionar la relevancia del dataset dado que al ser países en desarrollo es de gran dificultad encontrar variables que no presenten missing values para el conjunto de años en cuestión y para los cinco países analizados. A su vez, al considerar una definición más extensa de "default", como se denomina posteriormente, el dataset proporciona datos sobre préstamos y desembolsos incurridos de parte del FMI, para así poder considerar los cuasi defaults, a diferencia de la literatura previa. Por otro lado, el trabajo brinda un exhaustivo análisis de predicción de dichos defaults utilizando modelos de Machine Learning. La explosión de nuevas técnicas de aprendizaje automático (ML) cumplen un papel importante y pueden utilizarse para gestionar complejos modelos económicos equipados con enorme cantidad de datos. Por otro lado, ML representa una técnica clave que ajusta formas funcionales complejas y muy flexibles a los datos sin sobre ajustar, es decir, encuentra funciones que predicen bien fuera de la muestra. En 2003, Manasse et al. intentaron predecir las crisis de la deuda soberana para un conjunto de 47 países. Utilizaron regresión logística y árboles de decisión, y fueron capaces de predecir impagos soberanos con una precisión de alrededor del 90% con árboles de decisión, utilizando una base de datos propia. Además, Huang y Sethi (2017) realizaron el mismo ejercicio, con una muestra de 1.334 datos que abarcaban 43 países y 30 años, y entrenaron cuatro modelos en su conjunto de entrenamiento: regresión logística, SVM, una red neuronal y Random Forest. Concluyen que incluso con la eliminación de una variable que creían estaba altamente correlacionada con el resultado, Random Forest era el más indicado para predecir.

Al enfrentarse ante un problema de predicción y sobre todo cuando se dispone de un gran conjunto de datos, hay mejores opciones que la elección de Regresión Logística o Lineal. Los modelos de Machine Learning son más sofisticados que los utilizados en la estadística convencional y proporcionan seguridad en cuanto a la precisión del modelo incluso ante un mayor número de regresores potenciales en comparación con la presencia de pequeños volúmenes de observación. De esta forma, en este trabajo se propone analizar una suerte de carrera de caballos entre modelos predictivos con técnicas de ML, entre ellos Regresión Logística, Análisis Discriminante Lineal, Vecinos Cercanos, Árboles de Decisión, CART, Random Forest, Support Vector Machines, Bagging y Boosting. A partir de estos, se intentará evaluar cuál de ellos ha tenido un mejor desempeño respecto a métricas como la Precisión, el Área bajo la Curva, el Error Cuadrático Medio, entre otras. Para ello se propone analizar cuales son los determinantes de defaults soberanos según la literatura previa, y más aún cuales son significativos para predecir. En cuánto los modelos hayan sido evaluados, comparándolos e identificado distintos parámetros relevantes en cada uno, se llegará a la conclusión de cuál es el modelo que mejor se adecúa a los eventos ya acontecidos. Durante todo el proyecto, se utilizó SKLearn para implementar, entrenar y probar los modelos. Para el ajuste de los hiperparámetros y la validación, se utilizó k-fold validación cruzada (k = 5) y luego la elección del modelo fue a partir de aquel que presentase menor Error Cuadrático Medio. Esta métrica es una de las más utilizadas la cual permite evaluar hasta qué punto las predicciones se ajustan realmente a los datos observados. En otras palabras, permite cuantificar hasta qué punto el valor de respuesta predicho para una observación dada se acerca al verdadero valor de la respuesta para esa observación.

Se observó que Randon Forest es el modelo que mejor predice los defaults, con métricas superadoras contra los demás modelos. En primer lugar, es el que presenta menor ECM, minimizando así los errores de predicción y con mayor valor de precisión o accuracy. Sin embargo, respecto a las demás métricas utilizadas no se distingue de los demás modelos, dado que presenta el mismo AUC que regresión logística, KNN y SVM. Hay dos razones principales que pueden estar ocasionando estos resultados. En primer lugar, la faltante de datos que existe para determinados países en los años estudiados, es decir, las variables utilizadas son las mismas para cada país pero, dada la frecuencia, hay ciertas variables con faltantes para algunos años en estudio y dificulta la precisión en la predicción. En segundo lugar, el hecho de predecir un fenómeno que ocurre en el 6,8% de los años estudiados, resulta en un desbalance ya mencionado de la base.

Es importante estudiar este tipo de cuestiones y más aún en países en cuestión, donde el sector económico está continuamente analizando, evaluando y actuando en base a ello. A su vez, con esta investigación se propone entender las razones por las que un país no paga sus deudas, y ver qué factores son los que más influyen en dicha cesación de pagos. Por estos motivos se decidió tomar los mencionados países como casos de estudio, siendo los que más veces entraron en cesación de pagos desde 1970 hasta la actualidad.

El trabajo se organiza de la siguiente manera. En la siguiente sección se realiza una revisión de la literatura existente sobre distintas aplicaciones de aprendizaje estadístico y de Machine Learning para predecir la situación de default. La sección 3 explica los datos a utilizar, mientras que la sección 4 detalla la metodología general a seguir en cada modelo. En la sección 5 se exponen los principales resultados obtenidos y, por último, en el apartado 6 se presentan las reflexiones finales.

2. Literatura relacionada

La previsión de impagos soberanos como campo de investigación surgió en la década de 1970, cuando los niveles de deuda externa de los países en desarrollo aumentaron significativamente, lo que llevó a un creciente volumen de reestructuraciones soberanas en la década de 1980. A raíz de la caída del comunismo, se produjo un importante volumen de inversiones extranjeras en Europa del Este, Asia, América Latina y África, y los inversores internacionales comenzaron a alarmarse de los riesgos que podría implicar la globalización del comercio mundial y de los mercados financieros. De esta forma, se volvió cada vez más relevante investigar este tipo de acontecimientos para que las consecuencias que conlleva sean por lo menos más leves.

La literatura teórica destaca una variedad de factores que pueden desencadenar el impago de la deuda soberana y crisis de deuda. En un capítulo del documento del Banco Mundial, Primo Braga y Vincelette (2011) utilizan una base de datos que abarca 25 años para 46 países emergentes utilizando técnicas de promedios de modelos Bayesianos para determinar el conjunto de determinantes de defaults. Un primer examen de los datos indica que para toda la muestra, la probabilidad de impago se asocia sólo con el nivel de endeudamiento. Las variables que representan los costes de la deuda y el riesgo de refinanciación no parecen robustas como predictores del impago de la deuda, por lo que este resultado confirma la opinión de la literatura de que sólo unas pocas variables macroeconómicas y de calidad institucional son necesarias para predecir los defaults soberanos (Kraay y Nehru 2006). Estos autores examinan empíricamente los determinantes del "debt distress". Utilizando regresiones probit, encuentran que los defaults soberanos dependen de un pequeño conjunto de factores: la carga de la deuda, la calidad de las políticas e instituciones, y las perturbaciones. Se demuestra que estos resultados son robustos y muestran que sus especificaciones básicas tienen un importante poder predictivo fuera de la muestra. Además, la evidencia empírica sugiere que la probabilidad de una crisis de deuda está negativamente correlacionada con el crecimiento del PBI (Sturzenegger, 2004), y el nivel de reservas internacionales (Dooley, 2000) y positivamente relacionada con el nivel de deuda total (McFadden et al., 1985) y deuda total a corto plazo (Detragiache y Spilimbergo, 2001). En cambio, Manasse et al. (2003) señalan que este tipo de crisis responde fuertemente a las condiciones políticas, mientras que Reinhart et al. (2003) sostienen que el historial previo de deuda está fuertemente correlacionado.

Por otro lado, en los últimos años las técnicas de Machine Learning fueron poco a poco predominando en los estudios empíricos por su mayor precisión al momento de predecir en comparación a los enfoques econométricos tradicionales. Por un lado, Savona y Vezzoli (2013) utilizan un nuevo enfoque basado en un árbol de regresión para alcanzar el mejor compromiso entre la bondad de ajuste dentro de la muestra y la predictibilidad fuera de la muestra de los impagos soberanos. Utilizan datos de países emergentes como Grecia, Irlanda, Portugal y España (GIPS) durante el período 1975-2010. Los resultados muestran que la iliquidez y el historial de impagos junto con el crecimiento del PIB real y los tipos de interés de EE.UU, son los principales determinantes de la reciente crisis de la deuda soberana europea. Por otro lado, Manasse y Roubini (2009) desarrollaron un análisis utilizando el modelo CART para examinar factores macroeconómicos, financieros y políticos que explican las crisis de la deuda soberana. Las 50 variables iniciales se redujeron a 10 utilizando árboles de decisión en los que se desarrollaron reglas para reconocer las características de los países en situación de impago. Se llegó a la conclusión de que no todas las crisis eran similares y podrían diferenciarse en términos de solvencia, liquidez y riesgos macroeconómicos. Alaminos et al. (2019) aplicaron la metodología de árboles de decisión difusos para predecir las crisis de la deuda soberana utilizando datos entre 1970 y 2017, y aplicando 30 variables y una validación cruzada de cinco veces. El área bajo la curva ROC (AUROC) del modelo global fue del 94%, lo que indica un gran poder de predicción. A su vez, como se mencionó anteriormente, Huang y Sethi (2017) desarrollaron modelos de Support Vector Machine, Random Forest y Logit utilizando una base de datos del FMI que contiene 1200 observaciones. Las variables se redujeron mediante el análisis de componentes principales, y los resultados se sometieron a pruebas retrospectivas mediante un método de validación cruzada con $k = 5$. Random Forest resultó ser el mejor modelo de predicción, con una precisión de clasificación del 91%, seguido de los métodos SVM (89%), KNN (88%) y logit (87%).

El presente trabajo busca hacer una contribución a esta literatura desde diferentes perspectivas. Si bien la metodología utilizada de Machine Learning es similar a la de otros estudios anteriores, hasta el momento, no se habían registrado investigaciones acerca de la predicción de defaults en países emergentes con algoritmos de aprendizaje supervisado a través de técnicas de tipo clasificatorias. La literatura previa indica una fuerte discusión en base a las metodologías econométricas tradicionales, y es por eso que, a través de esta investigación se busca contribuir a lo ya estudiado aprovechando las nuevas tecnologías que Machine Learning brinda. Esta investigación se lleva a cabo incluyendo variables económicas, políticas y sociales que se consideran lo suficientemente abarcativas como para poder construir una decisión en base a métricas de desempeño, de un modelo óptimo que resulte en la reducción de la incertidumbre económica y social que sufre la sociedad actual. A su vez, se ha decidido expandir la mera definición de defaults para poder capturar no solo las crisis de deuda clasificadas por el índice de Standard & Poor´s, sino que también los cuasi defaults que se previnieron con programas de ajustes por parte del FMI, factor que no había sido considerado hasta el momento

en investigaciones de este tipo y que ha sido necesario incluir luego de analizar lo que los datos demuestran.

3. Set de Datos

Para poder llevar a cabo una investigación con alta precisión, es importante establecer *a priori*, cuáles son los determinantes de un default de deuda soberana, e identificarlo para cada uno de los cinco países que se abordan en este trabajo. En conjunto, la literatura existente contribuye a nuestra comprensión de los posibles predictores de las crisis de deuda, que a su vez, pueden clasificarse como:

1. Riesgo de insolvencia: incluye variables de la cuenta de capital y de la cuenta corriente de capital (reservas internacionales, inversión directa extranjera, tipo de cambio real, balanza por cuenta corriente) y variables de deuda (deuda externa pública, deuda externa total, deuda externa a corto plazo);
2. Riesgo de iliquidez: aproximado por variables de liquidez (deuda a corto plazo en relación con las reservas, servicio de la deuda en relación con las reservas y/o las exportaciones, M2 en relación con las reservas);
3. Riesgo macroeconómico: medido por variables macroeconómicas (crecimiento del PIB real, tasa de inflación, tipo de cambio, tipos de interés internacionales);
4. Riesgo político, medido por factores institucionales/estructurales (apertura del mercado internacional de capitales, liberalización financiera, grado de inestabilidad política y derechos políticos, e historial de impagos);
5. Riesgo sistémico, la variable de contagio, que se suele aproximar al número/proporción de otras crisis de deuda.

Los datos fueron extraídos de las estadísticas de deuda internacional del Banco Mundial y de la fuente de datos del Fondo Monetario Internacional, mientras que los datos macroeconómicos para cada país, de los indicadores internacionales del Banco Mundial. Además, para la elección de los países seleccionados para el estudio, se distinguieron aquellos en los que se pueden notar importantes y numerosos períodos de default de deuda soberana en los últimos 50 años. Es por esto que el período en consideración se extiende entre 1970 a 2020.

Gran parte de la investigación previa a la predicción es explorar y entender en profundidad los datos que se disponen para poder usarlos de la manera más eficiente posible. Es por esto que, en primer lugar, se define que un país se encuentra en crisis de deuda si está clasificado como en default por el índice de Standard & Poor´s o si recibe un importante préstamo de parte del Fondo Monetario Internacional excediendo el 100 por ciento de la cuota asignada disponible. Investigaciones de la calificadora evidencian que las mayores vulnerabilidades del perfil crediticio de los países con historial de incumplimiento de pagos de deuda provienen del entorno financiero que se deteriora

rápidamente, la falta de confianza en los mercados financieros acerca de las iniciativas políticas bajo próximas administraciones, y la incapacidad del Tesoro de financiarse a corto plazo con el sector privado.

Como se mencionó anteriormente, según S&P un emisor es calificado como en incumplimiento de pagos o default si no ha cumplido con alguna o algunas de sus obligaciones financieras y lo considera como incumplimiento general, cuando el deudor no pagará específicamente ninguna o casi ninguna de sus obligaciones dentro de los plazos establecidos en las condiciones del préstamo. Como se menciona en el paper de Manasse et al. (2003), existe un potencial problema con esta definición ya que no captura los cuasi defaults que fueron prevenidos con un programa de ajuste y un importante paquete financiero proveniente del FMI. Es por esto mismo, que se decide añadir nuevas condiciones al significado de default de S&P para poder así abarcar estos potenciales problemas, en este caso, considerando los datos sobre préstamos y desembolsos incurridos a los 5 países considerados de parte del FMI. La institución del FMI se conforma de cuotas que son los componentes principales de la estructura financiera y de gobierno del FMI. La cuota de cada país miembro del fondo refleja su disposición económica relativa en la economía mundial. Estas se definen como Derechos especiales de giro (DEG), y funciona como unidad de cuenta del FMI. Basado en esta última definición, se clasifica a un país en crisis de deuda, si recibe un préstamo significativo que excede el porcentaje de cuota asignado y que un desembolso es realizado sobre dicho préstamo durante el primer año.

Es relevante notar que la base de datos utilizada en el presente trabajo está desbalanceada. Para esto, se deben encontrar los algoritmos de aprendizaje más adecuados, ya que la gran mayoría asumen una distribución relativamente equilibrada. Como se explica en la investigación de Miravet (2021), la presencia de clases desbalanceadas en bases de datos supone, en muchos casos, un problema en los modelos predictivos ya que estos tienden a centrar su atención sobre los casos de la clase mayoritaria. Es decir, en la mayoría de los años en estudio de cada país no hubo default, entonces es una fracción menor la de los años en default. En la muestra aproximadamente el 6,8% de los años hubo default. En consecuencia, se obtienen resultados que aparentan ser buenos, pero en definitiva, la predicción se realiza sólo en base a la clase mayoritaria. Es por esto que, el análisis de datos desbalanceados presenta una serie de características particulares que necesitan ser tratados de manera diferente tanto a la hora de entrenar como a la hora de evaluar en comparación a los datos con distribución equilibrada. Un algoritmo que clasifique a todos los años como no default calificaría correctamente casi al 93,2% de las observaciones. Es por esto que mientras mayor sea el desbalance en la muestra, aumenta de manera artificial la precisión de las estimaciones. Según Somasundaram y Reddy (2016), existen dos formas principales de lidiar con este fenómeno. Por un lado existe la posibilidad de modificar los modelos para agregar una ponderación mayor a las observaciones en la categoría minoritaria, en el presente trabajo la categoría minoritaria es la de los años en que hubo default en los países en estudio. Por otro lado, existen los métodos de remuestreo aleatorio que su función es equilibrar la proporción

de observaciones de las categorías de la variable de respuesta. Esto incluye el sobremuestreo que incrementa aleatoriamente el número de las observaciones minoritarias de manera que ambas categorías queden igualmente representadas, y el submuestreo que como el nombre lo indica, subrepresenta a la clase mayoritaria de no default, con el mismo fin.

4. Los modelos

Siguiendo la literatura de Athey e. Imbens (2019), el procedimiento tradicional en econometría, como se ha explicado en los conocidos textos de Angrist y Pischke (2008) y Wooldrige (2010), es especificar un objetivo que es funcional de una distribución conjunta de datos. Este objetivo es generalmente un parámetro de un modelo estadístico que describe la distribución de un conjunto de variables, condicionales a otras variables, en términos de un set de parámetros, que puede ser tanto finito como infinito. Dada una muestra aleatoria de la población de interés, los parámetros son estimados buscando los valores que mejor se ajusten a los datos, y esto se lleva a cabo, usando una función objetivo como la suma de los errores al cuadrado o con una función de probabilidad. Dicho esto, se puede identificar que en econometría tradicional, el foco se encuentra en la calidad de los estimadores del objetivo, que por lo general se mide en base a la eficiencia de muestras grandes. Los autores también mencionan el interés por construir intervalos de confianza, y que típicamente, se reportan errores estándares y estimaciones puntuales. En otras palabras, lo que importa es conocer la forma de la función de donde salen los datos. Este enfoque se concentra en la estimación de un modelo, típicamente representado por $y_i = x_i'\beta + \mu_i$, exógeno, en donde la relación entre la variable y_i dependiente y el vector de regresores x_i' está determinada por una teoría o una "estructura" (Sosa Escudero, 2018). Esto significa que lo que se busca es estimar de la mejor manera posible a los coeficientes β en donde la calidad que posee el estimador suele asociarse con ciertas propiedades deseables. Existen preferencias lexicográficas por la insesgadez en la mayoría de los casos, dejando a la eficiencia del estimador relegado en un segundo plano.

Sin embargo, en contraste con la literatura de la econometría tradicional, en la literatura de Machine Learning, los autores argumentan que el foco principal se encuentra en el desarrollo de algoritmos. El objetivo de estos algoritmos es generalmente, hacer predicciones sobre algunas variables dadas otras variables, o clasificar unidades sobre la base de información limitada. La diferencia con técnicas anteriores está en su capacidad para adaptarse a los cambios en los datos a medida que van entrando en el sistema y aprender de las propias acciones del modelo. Es por esto que, el aspecto iterativo es importante porque a medida que los modelos son expuestos a nuevos conjuntos de datos, estos pueden adaptarse independientemente. La forma del modelo, se aprende en base a los datos, y no se realiza inferencia, sino una predicción puntual, ya que se requiere obtener un poder predictivo fuera de la muestra. En otras palabras, el aprendizaje automático busca predecir y en base a x, donde el modelo

en sí, no tiene un rol destacado. Particularmente, lo que se quiere es predecir correctamente fuera de la muestra, esto es, evaluar la capacidad de predicción en observaciones que no se utilizaron para la construcción del modelo. Existe una preferencia por tolerar métodos más sesgados a cambio de una significativa disminución en la varianza, en comparación a los métodos de econometría frecuentista (Sosa Escudero, 2018). Los modelos con mayor grado de complejidad tienden a ser menos sesgados, pero a tener una mayor varianza y ser más erráticos, mientras que un menor grado de complejidad en el modelo permite disminuir la varianza a costas de un sesgo mayor. Para controlar dicha complejidad existen los hiperparámetros que maximizan la precisión de la predicción en base a una función de pérdida. Estas estrategias de regularización incorporan una penalización en el ajuste por mínimos cuadrados ordinarios para poder evitar un sobreajuste y por ende reducir la varianza, atenuando el efecto de la correlación entre predictores y el impacto que un modelo puede generar sobre los predictores de menor importancia.

Otra gran ventaja del enfoque de Machine Learning es a la hora de entender o encontrar patrones generales en los datos. Dentro del aprendizaje automático existen diferentes algoritmos o tipos de modelo que difieren en el tipo de datos ya sea de entrada como de salida, la estructura y la complejidad computacional. Existen dos tipos de algoritmos de aprendizaje estadístico que son comúnmente empleados en este tipo de estudio. Como señala Plukikova (2016), los algoritmos de aprendizaje no supervisado buscan elaborar e identificar alguna estructura entre los datos sin distinguirla de antemano. Funcionan a partir de un espacio de características sin la necesidad de una variable de respuesta. Además, distinguen un patrón inicial en los datos y por lo general, funcionan con variables predictoras y con variables de respuesta. Estos se utilizan para tareas predictivas como lo es el problema de clasificación binaria de este trabajo.

En este caso fueron entrenados 8 modelos en el set de entrenamiento: regresión logística, análisis discriminante lineal, KNN, árbol de decisión, support vector machine (SVM), bagging, random forest y boosting. Para poder llevarlo a cabo, se eligió usar SKLearn para implementar, entrenar y testear los modelos (Pedregosa et al., 2011). Para el ajuste y la validación de hiperparámetros se utilizó k-fold validación cruzada, con un $k = 5$, ya que el desbalance muestral deja en evidencia la necesidad de una disminución en el número de particiones de la muestra. Para ello, se siguen los siguientes pasos generales para cada algoritmo utilizado. En primer lugar, luego de la selección de los modelos a evaluar en dicha carrera de caballos, se los entrena con datos pertenecientes al período 1970-2020 y se realiza la elección del valor óptimo de sus respectivos parámetros a partir de la técnica de validación cruzada. Con esto se consigue que la complejidad del modelo quede optimizada para evitar problemas de sobreajuste. En segundo lugar, una vez entrenados se utilizan los datos no entrenados como base de prueba, y se evalúa en qué medida estos modelos logran clasificar correctamente a este conjunto de observaciones que no fueron utilizados para dicho entrenamiento. Finalmente, se computan las

métricas de desempeño tradicionales para describir la capacidad predictiva de cada método, y así poder identificar el mejor modelo.

A continuación abordaremos, con detalle, la explicación del funcionamiento de cada uno de los modelos propuestos. Es importante remarcar que la selección de los modelos no es exhaustiva a todos los que se podrían tener en cuenta en dicha carrera de caballos, pero fueron incluídos aquellos que se identificaron como los más relevantes para el caso dado. Esta selección fue basada en dos principales bases de información. Por un lado, se consideraron los modelos estudiados de manera teórica en la literatura existente relacionada al tema en cuestión y por otro lado, se implementaron técnicas de Machine Learning y características de modelos predictivos que fueron enseñados durante la carrera a través de diferentes trabajos prácticos y teóricos.

Regresión Logística

La Regresión Logística es utilizada para estimar la probabilidad de que una variable pertenezca a una clase particular, y es conocida como la técnica más tradicional. Esta regresión consiste en un modelo de probabilidad condicional de ocurrencia de Y dado un conjunto de predictores X:

$$p = P(Y = 1 \mid X)$$

Donde el modelo, que no es lineal, para p se identifica como,

$$p = F(x'\beta) = \frac{e^{x'\beta}}{1+e^{x'\beta}}$$

Luego de obtener cada predicción de probabilidades individuales, es decir para cada año y país en particular, se crea un umbral de decisión σ para determinar si una observación pertenece a una clase o a la otra. Es así que, si para un año y país en particular la probabilidad predicha es mayor a σ, entonces se clasificará como *default*. En caso contrario, si es menor que σ, se clasificará como año y país de *no default*.

Support Vector Machine

Este modelo de SVM, por sus siglas en inglés, o Máquina de Soporte Vectorial, forma parte de los métodos de aprendizaje supervisado para la regresión, clasificación o detección de extremos y fue inicialmente implementado por Cortes y Vapnik (1995). En este modelo se identifican funciones base que están centradas en los puntos de datos de entrenamiento de la muestra y, seguidamente, se selecciona un subconjunto de dichos puntos los cuales se los nombra vectores de soporte. Este tipo de modelos predictivos son empleados para la clasificación de conjuntos de datos complejos, pero donde el tamaño es mediano o pequeño, y además pueden realizar tareas de clasificación lineales y no lineales (Burges, 1998). Este consiste en establecer el límite de clasificación no lineal mediante la

construcción de límites lineales de decisión, específicamente, hiperplanos separadores (James et al. (2013)), en un espacio transformado del conjunto de los predictores originales. A la hora de elegir el hiperplano maximizador (aquel que separa perfectamente a las dos clases de observaciones), aparece el problema de que probablemente estos sean infinitos. Por definición, en un espacio de p dimensiones, el hiperplano sería un subespacio plano de $p - 1$ dimensiones. Dicho esto, un ejemplo es que en un espacio de dos dimensiones, el hiperplano se identifica como una línea, y en un espacio de tres, el hiperplano es un plano (James et al. (2013)). Es por esto que, en un espacio de p dimensiones, el hiperplano es definido por la ecuación,

$$\beta_o + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ + \beta_p x_p = \beta_0 + \ \beta' x = 0$$

Dada la situación, se podría elegir aquel hiperplano separador que maximice la distancia a los puntos más cercanos de cada clase, los llamados support vectors. De la distancia perpendicular entre los support vectors y el hiperplano, es decir del margen, se desprende el hiperplano separador óptimo. No obstante, dicho hiperplano de óptima separación suele ser sensible a los datos y de poca robustez, lo que podría provocar un sobreajuste y por consiguiente, una mala predicción por fuera de la muestra. Además, no es recurrente que las observaciones de la muestra de entrenamiento sean perfectamente separables de modo lineal.

Árboles de Decisión

La principal elección de Árboles de Decisión fue basada en la evidencia de usos previos y en el hecho de que se clasifica como uno de los mejores modelos a la hora de predecir con una base de datos desbalanceada. Fue introducido por primera vez por Breiman (2001), con el principal objetivo de poder capturar y modelar automáticamente las relaciones no lineales entre predictores. Para una mejor explicación de este algoritmo, es necesario introducir los árboles de clasificación y regresión, ya que son las raíces de este algoritmo complejizado.

En primer lugar, propondremos utilizar la metodología de CART, que realiza un análisis en árbol de regresión a través de una secuencia de reglas para predecir un outcome binario, descrito en la literatura de Manasse et al. (2003). CART realiza una regresión no paramétrica, es decir que no se estiman parámetros, sino que directamente los datos nos indican la distribución. Por este motivo, puede detectar relaciones complejas entre las variables explicativas o independientes (X) y las variables dependientes (Y). En esencia, lo que hace este método es partir los espacios de atributos (de las X), y ajustar un modelo simple para Y dentro de cada región, donde propone como predicción la media muestral de Y. Este procedimiento es una partición recursiva binaria, y se debe encontrar el mejor ajuste global, es decir, la variable y el punto de partición óptimos. Este método funciona muy bien para estructuras no-lineales. El procedimiento específico comienza con la división del espacio de

los X_p predictores en j regiones, R_1, R_2, … , R_j. Seguidamente, a cada observación que cae en la región R_j se la clasifica en la categoría k que resulta más frecuente en esa región. Es así como la variable y el punto de partición se eligen de modo que se maximice la homogeneidad, respecto de la variable de respuesta. Dicho esto, dada la variable j y el punto de partición s, es posible definir los siguientes semiplanos,

$$R_1(j,\ s)\ = \{X|X_j \leq s\},\ R_2(j,s) = \{X|X_j > s\}$$

Sin embargo, es un método poco robusto a los datos, ya que una pequeña alteración en estos puede dar como resultado un árbol estimado completamente diferente. Es un método que presenta una varianza muy alta; esta inestabilidad se debe a las características jerárquicas que poseen naturalmente los árboles de decisión.

Bagging

Una solución al problema mencionado anteriormente es mediante Bagging, introducido también por Breiman (2001). Este consiste en primero, tomar *B* muestras aleatorias de tamaño n con reemplazo de la muestra original. Segundo, se debe estimar un árbol para cada una de esas muestras y se guardan las predicciones individuales. Finalmente, la clasificación se realiza a través del voto por mayoría; se toma como predicción el promedio de las predicciones de cada árbol. La estimación de Bagging resulta en un vector $\widehat{f}_{bag(x)}$ de tamaño $K\ [p_1(x),\ p_2(x),\ \dots\ ,\ p_K(x)]$ en donde $p_K(x)$ representa la proporción del total de los árboles que predicen la clase k para una observación determinada. Es por esto que, este algoritmo clasifica a las observaciones a partir de la clase con mayor cantidad de votos de los *B* árboles,

$$\widehat{G}_{bag}(x)\ =\ argmax_K \widehat{f}_{bag}(x)$$

La intuición detrás se basa en la idea de que la varianza del promedio es menor que la varianza de un árbol único, es por esto que, la predicción de todos los árboles conjunta, es más robusta que la individual.

La Regresión Logística se distingue fácilmente de los modelos de Random Forest y SVM ya que el primero es lineal en el sentido de que no incorpora términos polinómicos, términos de interacción, o alguna forma más compleja de los predictores. Esto difiere de los otros dos modelos porque tienen la capacidad de captar todos esos tipos de relaciones no lineales más complejas de manera automática.

Random Forest

Un modelo Random Forest está formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping. Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por nodos generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo, es decir, es la media de las predicciones de todos los árboles que lo forman.

La diferencia entre Bagging y Random Forest es el número de predictores que se usan por iteración, en Bagging se utilizan todos los predictores por cada iteración, mientras que en Random Forest se utiliza una muestra de predictores de tamaño m, lo cual tiene una ventaja. Si se supone un predictor fuerte en el conjunto de datos, junto con otros predictores moderadamente fuertes entonces, en Bagging, la mayoría o todos los árboles utilizarán este predictor. En consecuencia, todos los árboles bagging serán muy similares entre sí y por ende están altamente correlacionadas. Promediar cantidades altamente correlacionadas no conduce a una reducción significativa de la varianza como promediar cantidades no correlacionadas. Random Forest supera este problema al obligar a cada división a considerar sólo un subconjunto de los predictores. En el caso de este modelo, el trade-off entre sesgo y varianza es regulado por la cantidad de m predictores aleatorios utilizados para realizar dichas particiones binarias. Es por esto que, cuanto más grande es m el modelo va a tener un menor sesgo y mayor varianza, y cuando es más chico, la varianza disminuye a cambio de un sesgo mayor. Entonces de los p predictores originales, este algoritmo sólo utilizará $m < p$ elegidos al azar, donde por definición, $m = \sqrt{p}$. Según Friedman et al. (2009), primero este modelo entrena *B* árboles a partir de muestras obtenidas por bootstrap de la muestra original. Para esto se requiere repetir recursivamente los pasos a continuación,

1. m variables de los p predictores originales son seleccionados aleatoriamente, donde por lo general, $m = \sqrt{p}$
2. A partir de dichos m predictores seleccionados anteriormente, se realiza la partición y el punto que optimicen el resultado

Luego de este proceso, se clasifica a las observaciones a partir del voto mayoritario de los árboles que fueron entrenados. Si $\widehat{C}_b(x)$ es la clase predicha por el árbol b, entonces

$$\widehat{C}_{rf}^{\,B}(x) = votomayoritario\left\{\widehat{C}_b(x)\right\}_1^B$$

Vecinos Cercanos

K - Vecinos cercanos o K - Nearest Neighbours (KNN), por otro lado, es un modelo relativamente simple, pero que podría brindarnos mucha información acerca de las características de los países que entran en default, y cuáles de ellas son relevantes para poder realizar una predicción correcta. Aunque no creemos que termine siendo el modelo que mejor prediga, se puede explotar la facilidad con la que este modelo se extiende a múltiples categorías para poder tener en cuenta distintas características cualitativas o cuantitativas de los países (cómo distancia (en días) respecto al último default, nivel de deuda respecto a gasto público, entre otros) y conseguir un modelo, por falta de mejor palabra, de país propenso a caer en defaults.

$N_k(x_o)$: conjunto de las k observaciones más cercanas a x_o,

$$\widehat{g}_{knn}(x_o) = \frac{1}{k}\sum_{i=1}^{n} 1[x_i \epsilon N_k(x_o)|Y_i$$

5. Métricas de Desempeño Predictivo

A cada modelo entrenado en este trabajo, se lo califica y evalúa según una serie de métricas de desempeño. Como la variable de interés *default* es binaria, son cuatro los escenarios posibles que existen para la clasificación de error de tipo 1 y 2, que conforman la matriz de confusión: que un modelo prediga que un país está en default cuando no lo está (falso positivo), que está en default cuando lo está (verdadero positivo), que prediga que no está en default cuando está en default (falso negativo), que no está en default cuando no lo está (verdadero negativo).

Según la literatura relacionada, las métricas comúnmente más utilizadas en este espacio son las siguientes.

- *Tasa de aciertos*: donde se mide la proporción del total de las observaciones que fueron predichas correctamente. El cálculo es el siguiente:

$$\frac{VP + VN}{VP + FP + VN + FN}$$

- *Tasa de verdaderos positivos o sensibilidad*: indica la proporción de años en default clasificados correctamente en relación al total de años que pertenecen a dicha categoría:

$$\frac{VP}{VP + FN}$$

- *Tasa de verdaderos negativos o especificidad*: indica la proporción de los años sin default clasificados correctamente en relación al total de observaciones que pertenecen a dicha categoría:

$$\frac{VN}{VN + FP}$$

- *Accuracy o precisión:* de todas las observaciones clasificadas en default, mide cuál es la proporción que fue clasificada correctamente:

$$\frac{VP}{VP + FP}$$

- *Area under Curve (AUC), o área bajo la curva ROC (Receiver Operating Characteristic)*: representa gráficamente la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes umbrales de clasificación. El área debajo la curva con un rango entre 0 y 1, indica el grado de separabilidad. La intuición detrás es que la curva mide la capacidad que tiene un modelo para identificar correctamente los casos acertados de falsas alarmas. Dicho esto, cuanto mayor sea el área de la región debajo la curva, mejor clasifica el modelo.
- *Error Cuadrático Medio (ECM):* es una de las métricas más utilizadas en la literatura la cual mide el error cuadrado promedio de las predicciones que se computan para cada modelo. Para esto, se calcula para cada punto, la diferencia cuadrada entre la predicción del modelo real y el objetivo de predicción, y luego se realiza un promedio de estos valores. Al momento de comparar entre modelos predictivos, se espera que el mejor de ellos sea el que contiene el mínimo ECM. Cabe aclarar que esta métrica de desempeño destaca grandes errores entre los pequeños, y además al ser diferenciable, se compromete a encontrar valores mínimos y máximos a través de la utilización de métodos matemáticos de forma más efectiva.

$$ECM(\widehat{\beta}) = Sesgo^{2}(\widehat{\beta}) + Var(\widehat{\beta})$$

6. Resultados

Con base en la metodología expuesta, el objetivo del análisis de los modelos cuyos resultados se mostrarán en este apartado, es establecer cuál es el más adecuado para la estimación de los defaults soberanos. Dicho esto, en esta sección se abordarán los principales resultados obtenidos evaluando la habilidad predictiva de los algoritmos propuestos. La tabla 1 resume para cada modelo entrenado los valores de las medidas de desempeño predictivo mencionadas en el apartado anterior.

Tabla 1: Resultados

Modelo	Penalty	Hiperparámetro	VP	FP	VN	FN	Accuracy	AUC	ECM
Regresión Logística	L2	0.00001	81.0	0.0	0.0	8.0	0.910112	0.500000	0.089888
Análisis discriminante lineal	-	-	74.0	7.0	0.0	8.0	0.831461	0.456790	0.168539
KNN	-	5.0	81.0	0.0	0.0	8.0	0.910112	0.500000	0.089888
Árbol de decisión	-	1.0	75.0	6.0	0.0	8.0	0.842697	0.462963	0.157303
Suppport Vector Machine (SVM)	-	10.0	81.0	0.0	0.0	8.0	0.910112	0.500000	0.089888
Bagging	-	50.0	78.0	3.0	0.0	8.0	0.876404	0.481481	0.123596
Random Forest	-	20.0	81.0	0.0	0.0	8.0	0.915001	0.500112	0.089001
Boosting	-	1.0	74.0	7.0	0.0	8.0	0.831461	0.456790	0.168539

[1]

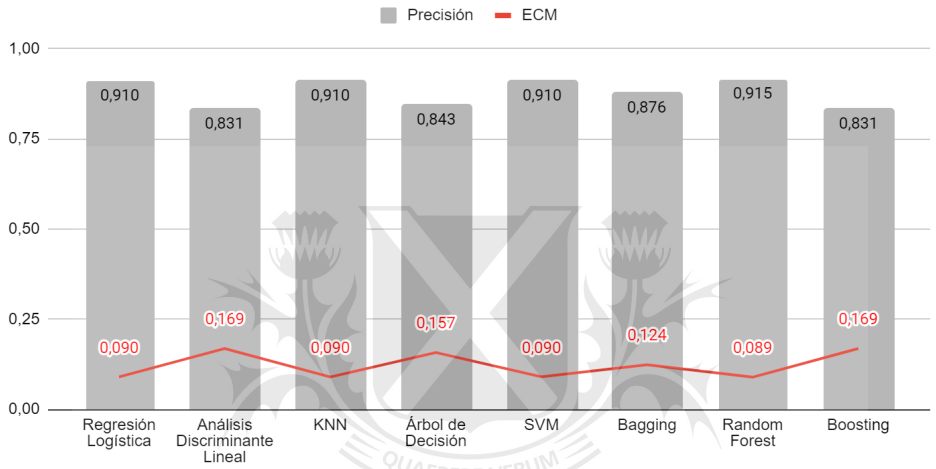
Como fue mencionado a lo largo de este trabajo, el foco se encuentra en poder elegir uno de los modelos utilizados para predecir default, en base a las medidas de desempeño que determinan la capacidad de cada modelo para el caso dado. En esta "carrera de caballos" entre modelos, el ganador es aquel que, luego de realizada la comparación, tiene las mejores habilidades para predecir la probabilidad de default en un año y país determinado.

En primer lugar, se puede observar que los resultados obtenidos en términos de accuracy, AUC y ECM para 3 de los modelos estudiados son iguales entre sí, lo que puede deberse entre otras cosas, al principal problema de desbalance de datos en la base. Siguiendo las definiciones establecidas anteriormente para las métricas de desempeño, y haciendo un análisis comparativo entre modelos, se puede identificar como modelo ganador al de Random Forests siendo el que presenta métricas superadoras. Se utiliza el hiperparámetro de n-estimators, el cual refleja que se han incluido 20 árboles en el modelo. El valor de precisión es el mayor y el valor del ECM el menor entre los modelos, minimizando así los errores de predicción. A su vez, este presenta una alta tasa de verdaderos positivos o sensibilidad, aunque iguales a otros modelos, lo mismo para el AUC que es igual a otros modelos como regresión logística, KNN y SVM pero menor que el resto, por lo que se requiere encontrar el valor más alto debajo de la curva para poder identificar al modelo que mejor clasifica. En base a esta evidencia el modelo elegido que mejor predice la situación de default en un año y país determinado es el de Random Forests, el cual tiene la capacidad de predecir correctamente el 92% del total de las observaciones. Junto a los modelos de regresión logística, KNN y SVM tienen la mayor cantidad de verdaderos positivos (81) del total de las observaciones, esto indica la

[1] VP hace referencia a la tasa de Verdaderos Positivos, FP a Falsos Positivos, VN a Verdaderos Negativos, FN a Falsos Negativos, AUC al área bajo la curva ROC y ECM al Error Cuadrático Medio. La tabla presenta 15.379 observaciones

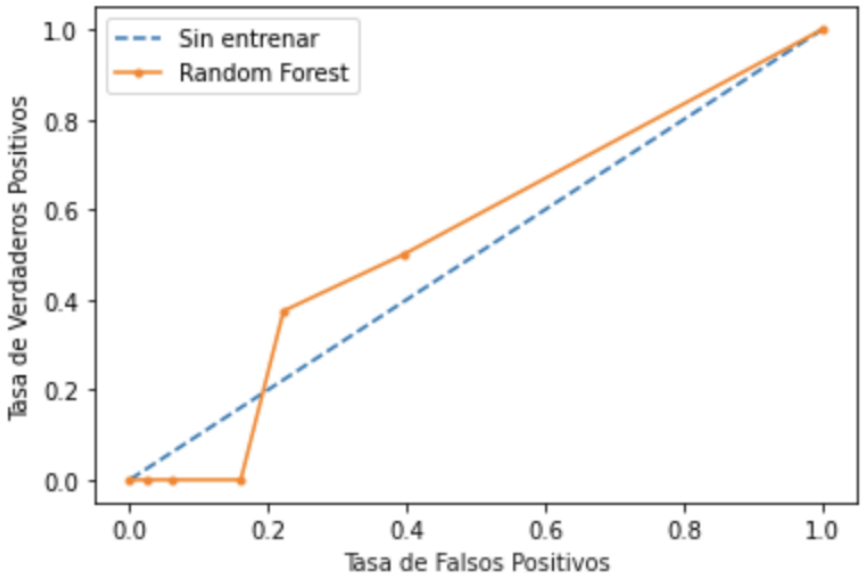
proporción de años en default clasificados correctamente en relación al total de años que pertenecen a dicha categoría. A su vez junto a los mismos 3 modelos mencionados anteriormente, la tasa de falsos positivos (0.0) es la menor, por lo que no se han clasificado años y países en default incorrectamente, y el área debajo de la curva (AUC) es la mayor, indicando que el modelo puede identificar correctamente en un 50% los casos acertados de falsas alarmas. En el gráfico 1 se detalla el Accuracy y ECM para todos los modelos y Random Forest se destaca como aquel con menor ECM y mayor Accuracy entre ellos.

Gráfico 1: Resultados de Accuracy y ECM (Error Cuadrático Medio)



Se puede identificar en los valores de la tabla de los resultados obtenidos que existe espacio para una posible mejora en la base de datos utilizada. Esto puede realizarse a través de la obtención de una base de datos más completa, que incluya las mismas variables y observaciones para todos los países y para cada año en estudio. También, agregando años en estudio y países para poder obtener una mayor cantidad de observaciones, aumentando el porcentaje del grupo minoritario *default* y, las predicciones puedan ser más precisas. Se observan valores repetidos para el ECM o Accuracy, y valores similares entre los modelos elegidos. Hay dos razones principales que pueden estar ocasionando estos resultados. En primer lugar, la faltante de datos que existe para determinados países en los años estudiados, es decir, si bien las variables utilizadas son las mismas para cada país, hay ciertas variables con faltantes para algunos años en estudio y esto hace que se cree un problema a la hora de evaluar la precisión en la predicción. En segundo lugar, el hecho de predecir un fenómeno que ocurre en el 6% de las veces resulta en un desbalance ya mencionado de la base. El hecho de que las observaciones que indican la ocurrencia de default pertenezcan a un grupo minoritario, hace que los datos no se ajusten correctamente a los algoritmos de predicción y el análisis comparativo entre los modelos no sea el deseado.

Gráfico 2: Curva ROC en Random Forest y sin entrenamiento



Como se puede observar en el gráfico 2, la curva ROC ilustra la sensibilidad y especificidad de cada uno de los posibles puntos de corte de un test diagnóstico cuya escala de medición es continua. La sensibilidad representa el ratio de verdaderos positivos, es decir, $VP = VP / (VP + FN)$ mientras que (1 - especificidad) está representada por la tasa de falsos positivos, $VN = VP / (FP + VN)$. En el gráfico se observa la forma en la que actúa el modelo elegido de Random Forest. Lo que se demuestra es que, a pesar de las limitaciones de la base de datos, aún así el modelo elegido presenta una mejora al momento de predecir, en comparación de una clasificación azarosa (modelo sin entrenamiento). Si bien es pequeña la mejora del modelo Random Forest, es notable la diferencia entre este modelo y un modelo que no presente ningún tipo de entrenamiento en sus datos.

7. Conclusiones

La necesidad de modelos estándares que puedan generar una predicción concreta acerca de la situación económica futura de un país, es una de las principales preocupaciones y motivo de continua investigación en los centros económicos de los países estudiados. La incertidumbre, inestabilidad, desconfianza y continua desilusión por parte de las economías emergentes son otras de las tantas razones por las cuales este es un problema recurrente, en continuo desarrollo y aún no solucionado. Es por esto que, mediante este trabajo se busca agregar valor a la literatura ya existente y poder alcanzar, de alguna manera, un paso más hacia una solución futura. Las nuevas tecnologías son herramientas primordiales para cualquier tópico de estudio en la actualidad y son las que generan, con evidencia, grandes avances en la sociedad actual desde todos los puntos de vista. Por esto es que, se ha decidido implementar técnicas innovadoras de Machine Learning aprovechando los algoritmos automatizados que esta ofrece a través de la programación de datos.

Es menester mencionar los principales obstáculos y desventajas que hemos encontrado a lo largo de esta investigación y que también son cuestiones a mejorar en el futuro, abriendo así la posibilidad de continuar con una segunda parte de este trabajo. En primer lugar, el principal motivo por el cual los resultados no han sido los deseados fue por la falta de datos de libre acceso provenientes de las entidades públicas nacionales e internacionales, ya que solo proveen información incompleta y con faltantes en algunos años. Este fue un inconveniente que ha estado presente a lo largo de todo el trabajo, originando un desbalance de la base de datos y que todavía es una cuestión para resolver en un futuro. En segundo lugar, otro obstáculo fue la dificultad de cuantificar variables realmente importantes y de alto impacto en el resultado final, como lo son la inestabilidad, la desconfianza y el descontento político de los ciudadanos en un país determinado. Hay muchos aspectos y eventos políticos y sociales que son altamente complejos de cuantificar, por lo que son muchos los factores relevantes que quedarán fuera de consideración a la hora de realizar una predicción como la es la de defaults en determinados países.

La falta de datos en algunos años y países, junto con el hecho de predecir un fenómeno que ocurre en el 6% de las observaciones totales, representando así una clase minoritaria como lo es la de años y países en default, y la dificultad de cuantificar variables relevantes de alto impacto en los resultados, son las razones principales por las cuales consideramos a este trabajo como una mera aproximación a un resultado final, y con un alto potencial para ampliar el estudio en cuestión. Esperamos poder generar en un futuro, una investigación con una base de datos evolucionada, más precisa y completa que la aquí implementada.

Referencias

Alaminos, D., Fernández, S. M., Neves, P. M., & Santos, J. A. C. (2019). "Predicting sovereign debt crises with fuzzy decision trees".

Ams, J., Baqir, R., Gelpern, A., & Trebesch, C. (2019). "Sovereign default. Sovereign Debt: A Guide for Economists and Practitioners", 275-327

Angrist JD, Pischke JS. (2008). "Mostly Harmless Econometrics: An Empiricist's Companion", Princeton, NJ: Princeton Univ. Press.

Athey, Susan and W.Imbens, Guido (2019). "Machine Learning Methods That Economists Should Know About", Annual Review of Economics.

Arellano, C. (2008)**.** "Default risk and income fluctuations in emerging economies". American economic review, 98(3), 690-712.

Breiman, L. (2001). "Random forests. Machine learning", 45 (1), 5–32

Burges C. J. C. , (1998). "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121–167.

Borensztein, E., & Panizza, U. (2009). *"The costs of sovereign default".* IMF Staff Papers, 56(4), 683-741.

Cabrera, A. A. (2014). "Historia económica mundial 1950–1990". Economía Informa, 385, 70-83.

Catao, L., Sutton, B. (2002). 'Sovereign Defaults: The Role of Volatility', IMF Working Paper, 02/149.

Cavallo, E. A. and Frankel, J. A. (2008). 'Does Openness to Trade Make Countries More Vulnerable to Sudden Stops, or Less? Using Gravity to Establish Causality', Journal of International Money and Finance, Vol. 27, pp. 1430-1452.

Cortes, C., y Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20 (3), 273–297.

Detragiache, E., Spilimbergo, A,. (2001). "Crises and Liquidity: Evidence and Interpretation". IMF Working Paper No. 01/2.

Dooley, M. (2000). 'A Model of Crises in Emerging Markets', The Economic Journal, Vol. 110, pp. 256-272.

Eichengreen, B., Rose, A. and Wyplosz, C. (1996). "Contagious Currency Crises: First Tests", Scandinavian Journal of Economics, Vol. 98, pp. 463-84.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hébert, B., Schreger, J. 2017. "The costs of sovereign default: Evidence from Argentina". American Economic Review, 107(10), 3119-45

Huang, Andrew, and Taresh Sethi. (2017). "Predicting sovereign default". In Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR 70

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Krayenbuehl, Thomas E. 1985. "Country Risk: Assessment and Monitoring". Cambridge: Woodhead-Faulkner. [CrossRef]

Kraay, A., and V. Nehru. 2006. "When Is External Debt Sustainable?" World Bank Economic Review 20 (3): 341–65.

Manasse, P. and Roubini, N. (2009). 'Rules of Thumb for Sovereign Debt Crises', Journal of International Economics, Vol. 78, pp. 192-205.

Manasse, P., Roubini, N. and Schimmelpfennig, A. (2003). 'Predicting Sovereign Debt Crises', IMF Working Paper, 03/221.

McFadden, D., Gershon, F., Vassilis, H., O'Connell, S,. (1985). "Is There Life After Debt? An Econometric Analysis of the Creditworthiness of Developing Countries", in Gordon Smith and John Cuddington, eds., International Debt and the Developing Countries. World Bank: Washington, DC.

Miravet, Blanca Abella (2021). "Mejora de las predicciones en muestras desbalanceadas", Universidad Autónoma de Madrid de Escuela Politécnica Superior.

Muñoz Jaramillo, Victor Daniel (2021). "Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín", Universidad Nacional de Colombia.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Plulikova, N. (2016). Poverty analysis using machine learning methods. Bachelor's in Mathematics Thesis, Comenius University in Bratislava.

Reinhart, C., Rogoff, K. and Savastano, M. (2003). 'Debt Intolerance', Brookings Papers on Economic Activity, Vol. 1, pp. 1-74

Savona, R., & Vezzoli, M. 2015. "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals". Oxford Bulletin of Economics and Statistics, 77(1), 66–92.

Somasundaram, A., y Reddy, U. S. (2016). Data imbalance: effects and solutions for classification of large and highly imbalanced data. En International conference on research in engineering, computers and technology (2016) (pp. 1–16)

Sosa Escudero, W. (2018). Big data y aprendizaje automático: Ideas y desafíos para economistas, en una nueva econometría: Automatización, big data, econometría espacial y estructural. Universidad Nacional del Sur.

Sturzenegger, F. (2004). 'Toolkit for the Analysis of Debt Problems', Journal of Restructuring Finance, Vol. 1, pp. 201-203

Sturzenegger, F., & Zettelmeyer, J. (2007). "Debt defaults and lessons from a decade of crises". MIT press.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28 (2), 3–28.

Wooldridge JM. (2010). "Econometric Analysis of Cross Section and Panel Data", Cambridge, MA: MIT Press.