Universidad de
SanAndrés

Big Data - Propuesta de Investigación

Unveiling Innovation: A Machine Learning Approach to Predicting Innovation in Firms

Alumnos: Juan Diego Barnes, Fransisco Legaspe & Rodrigo Martin

Profesora: Noelia Romero

Tutora: Victoria Oubiña

1. Introducción

De acuerdo con la Agenda 2030 para el Desarrollo Sostenible de las Naciones Unidas, la inversión y la innovación en el ámbito empresarial privado son reconocidas como los motores primordiales de la productividad, el crecimiento económico integral y la generación de empleo. Además, la Conferencia de las Naciones Unidas sobre Comercio y Desarrollo enfatizó la importancia de que los *policymakers* adopten estrategias de innovación, en especial en países en vía de desarrollo (Sirimanne et al., (2018); UNCTAD, (2021); Naciones Unidas, (2022)).

Por su naturaleza, la definición de «innovación» no es trivial. En un esfuerzo por llegar a un consenso al respecto, el Manual de Oslo - Primera Edición (1992) propuso un sistema de indicadores de innovación, que se convirtió en un canon internacional y ha sido usado en las diversas encuestas europeas sobre innovación, así como por la OCDE. El manual se ha ido actualizando en el tiempo. Actualmente se cuenta con el Manual de Oslo - Cuarta Edición (2018).

En él se hace la importante distinción entre una firma innovadora y una innovativa. Una firma *innovativa* es aquella que realizó alguna actividad de innovación. Que sea *innovadora* se refiere a una firma que desarrolló o implementó nuevos productos o procesos. En otras palabras, innovativa apunta a sí la empresa está realizando esfuerzos de innovación, mientras que innovadora hace referencia a si la empresa incluyó nuevos productos o procesos (independientemente de si hizo o no un esfuerzo activo). Una firma puede o no ser innovadora e innovativa, generando así cuatro grupos: Innovativa e Innovadora, Sólo Innovadora, Sólo Innovativa, No innovativa y No innovadora (Manual de Oslo, (2008)).

Pese a la relevancia y trabajo dedicado al tema, no hay consenso acerca de como realizar la medición de innovación e innovatividad de una firma. Esto es especialmente difícil en países en desarrollo, donde la innovación no suele ser por crear nuevas tecnologías, sino por adoptar las creadas en los países más desarrollados. Dificultando la observabilidad y por lo tanto la medición (Verhoogen, (2021)).

El enfoque más común es el de tomar cambios en la productividad total de factores (TFP) como proxy. Esto requiere asumir que la productividad subyacente de las firmas y la inversión tienen una relación monotónica. Esto se rompe en casos de heterogeneidades de restricciones de crédito (Olley & Pakes, (1996)) o si hay costos de ajuste a la inversión (Grilichese & Mairesse, (1998)).

Ambos suelen estar presentes en países en desarrollo, por lo que los supuestos bajo los cuales se usa TFP para medir innovación son muy fuertes y es de esperar que no se cumplan.

Este trabajo busca responder la siguiente pregunta de investigación: *¿ Cómo predecir si una firma es innovadora y/o innovativa en base a su características observables?*. Para esto explotamos una base con información de más de 4.000 firmas Argentinas, relevada por el Ministerio de Ciencia, Tecnología e Innovación Productiva de la Presidencia de la Nación.

Esto expande la frontera del conocimiento en dos aspectos principales. Por un lado, ilumina la discusión acerca de la medición de innovación de países en desarrollo, brindando un primer modelo predictivo basado en métodos de Machine Learning, que circunvala los supuestos del método de TFP. Por otro lado, permite identificar cuales son las principales variables que explican el carácter innovador e innovativo de una firma en particular, basandose en los datos (en base a un análisis mediante CART).

Además, buscaremos la misma predicción pero limitándonos al set de variables a las que tiene un policy maker tiene acceso directo (i.e., sin necesidad de relevar por una encuesta). Predecir el carácter innovativo e innovador de cada firma brinda la posibilidad de aumentar la eficiencia e impacto de políticas publicas al respecto. Hasta donde conocemos no hay ningún trabajo que proporcione herramientas para este fin.

2. Revisión de la literatura

En base a esto, nuestro trabajo se encuentra en la literatura de innovación en economías emergentes. La misma se centra en entender cuales son los mecanismos que conducen a que una empresa realice o no una actividad de innovación, en el contexto de países en vías de desarrollo.

Para esto debemos plantear una definición consensuada para qué es una innovación. El manual de Oslo (2018), define *innovación* como:

> *"Producto o proceso (o una combinación de ambos) nuevo o mejorado que difiere significativamente de los productos o procesos previos de la unidad y que se ha hecho disponible para potenciales usuarios (producto) o puesto en uso por la unidad (proceso)"*
>
> — Manual de Oslo (2018) - Resumen Ejecutivo.

En este trabajo, nos centraremos en una caracterización de las actividades que realizan las firmas que en el manual de Oslo se define como *innovation activities*. Se define *innovation activities* como:

> *"Todas las actividades de desarrollo, financieras y comerciales tomadas por una firma con la intención de resultar en una innovación para la firma".*
>
> — Manual de Oslo (2018) - Resumen Ejecutivo.

Bajo esta definición una empresa que realiza una actividad de innovación, es innovativa pero no necesariamente innovadora. Esto es debido a que para que una firma sea innovadora requiere adopción del producto o proceso, lo cual en los datos se observa como la existencia de un resultado de la innovación. Un aspecto fundamental de esta definición es que estas actividades de innovación pueden resultar en innovaciones, pueden estar en curso, ser pospuestas o ser canceladas.

Si bien la definición del manual de Oslo es la más utilizada en esta literatura no es la única y puede ser una definición problemática en casos particulares. Verhoogen (2021) comenta que la falta de una definición universalmente aceptada ha dado lugar a un desorden terminológico en el campo ya que algunos investigadores utilizan el término innovación para referirse únicamente a innovaciones nuevas para el mundo mientras que algunos se refieren a innovaciones no necesariamente nuevas para el mundo.

Alternativamente, Verhoogen (2021) se enfoca en innovaciones de menor grado de novedad, así que usa un término distinto: *upgrading*. El concepto de upgrading hace referencia a la adopción tecnologías y productos *avanzados*, mientras que el concepto de innovación pone el foco en que los productos o procesos sean *nuevos o mejorados*, sin importar su nivel de avance relativo en el sector de la empresa.

En este trabajo nos centraremos con el objetivo de conseguir una medida comparable y con consenso entre los policy makers decidimos que una mejor clasificación es la expuesta en el manual de Oslo (2018) teniendo en cuenta sus ventajas y limitaciones.

Estas definiciones son de gran importancia a nuestros fines, debido a que nuestra base de datos nos permite diferenciar en los tipos de innovación que realizo la empresa y aproximar un potencial determinante de esta innovación. A continuación, explicaremos con más detalle la estructura de nuestra novedosa base de datos. Hasta donde sabemos no hay papers que predigan el carácter

innovador e innovativo de una firma. Lo más cercano es el paper de Rojas-Cordova et al. (2020), que busca usar métodos de Machine Learning para identificar barreras a la innovación.

3. Los Datos

Con el fin de lograr explorar el problema de *upgrading* e innovacion en firmas en países en desarrollo, explotamos los datos de Encuesta Nacional de Dinámica del Empleo e Innovación (2014-2016), en su segundo operativo (ENDEI II – MINCyT y MTEySS), realizada en 2017. Esta encuesta es llevada acabo por el Ministerio de Ciencia, Tecnología e Innovación Productiva de la Presidencia de la Nación.[1] Los datos de la ENDEI nos permiten conocer si una firma innovo, que tipo de innovación realizo, cual fue su gasto en innovación absoluto y relativo. Además nos brinda medidas de beneficios y de valor agregado que nos permiten evaluar si existen retornos de la invocación. Una descripción de la principales variables incluidas en la encuesta y que son de nuestro interés, se puede encontrar en la Tabla 1 en el Apéndice.

Estos datos son especialmente útiles para identificar innovación o *upgrading*, al recolectares y presentarse siguiendo todas las pautas y recomendaciones del Manual de Oslo (2006). Esto nos permitiéndonos contar con diferentes medidas de innovación, actividades de innovación y posibles determinantes de la innovación. Pudiendo distinguir las actividades de innovación como innovadoras e innovativas. De modo que las actividades pueden clasificarse en:

- **Innovativa e Innovadora:** firmas que han realizado alguna actividad de innovación y han obtenido algún tipo de resultado durante el período 2014-2016.
- **Sólo Innovadora:** firmas que no han realizado actividades de innovación pero han obtenido algún tipo de resultado durante el período 2014-2016.
- **Sólo Innovativa:** firmas que han realizado alguna actividad de innovación pero no han obtenido resultados durante el período 2014-2016.
- **No innovativa y No innovadora:** firmas que no han realizado actividades de innovación ni tampoco han obtenido resultados durante el período 2014-2016

[1] El primer operativo (ENDEI I) se realizo en 2013 recolectando información de 2010, 2011 y 2012. El tercer y ultimo operativo se realizo en 2022, englobando la información de 2019 a 2021 (ENDEI III).

El Manual de Oslo define que una firma innovadora, como aquella que desarrolla e implemento nuevos productos o procesos, en los datos no se pude identificar la implementación, por esto en la clasificación basada en los datos se toman los resultados de la actividad de innovación como proxy de implementación.

A partir de estos datos nos planteamos dos objetivos primero poder determinar cuales parecen ser los principales determinantes, o al menos describir como son aquellas firmas que son innovadoras en los datos (data-driven).[2] Luego, predecir si una empresa es si una firma es innovadora, una dummy que toma valor 1 si la firma es innovadora y 0 en caso contrario. Luego quisiéramos predecir las cuatro clasificaciones de innovativa, innovadora, ambas o ninguna. Definimos el outcome como las 4 categorías descriptas anteriormente, y tomamos como variables todas aquellas variables que describen a las firmas, y no son medidas directas de innovación. Considerando que realizar encuestas es costoso, también planteamos limitar la variables incluidas en el modelo a aquellas que puedan ser obtenidas de registros administrativos (Por ejemplo: empleo, exportaciones, etc.) de modo de generar un modelo predictivo util para el desarrollo de politicas.[3]

4. Metodología

4.1. Reducción de dimensionalidad

Tras limpiar eliminar las preguntas de la encuesta que son directamente sobre innovación, pasamos de tener 685 features de la encuesta, a quedarnos con 486 variables. A pesar que de que el numero de dimensiones es menor tras esta limpieza parcial de los datos, seguimos enfrentando un problema de alta dimensionalidad el cual quisiéramos solucionar.

Para afrontar este problema, adoptamos Análisis de Componentes Principales (PCA). PCA nos permite resumir múltiples variables en factores (combinaciones lineales de las variables), maximizando la varianza pero reduciendo la cantidad de variables. El método buscar construir factores que reproduzcan o resuman de mejor manera posible la variabilidad de múltiples variables. A

[2] Predecir si una firma es innovadora, independientemente de si es innovativa es relevante debido a que puede haber problemas de timming y la madurez de la innovación, la empresa pudo haber realizado actividades de innovación, es decir ser innovativa, en los periodos previos.

[3] Adicionalmente a esto, todo lo descripto en la seccion, limpieza y decisiones sobre el manejo de missing values es crucial para la predicción.

priori sin contar con las estimaciones nos interesaría poder reducir el numero de dimensiones a un numero razonable, podemos pensar arbitrariamente en no mas de 20 componentes principales, parece razonable.

Adicionalmente también nos interesa estimar los modelos seleccionando las variables *ad hoc*, manteniendo aquellas que consideramos que pueden ser obtenidas de bases administrativas para lograr obtener un modelo que pueda predecir la innovación sin la necesidad de realizar encuestas. Incluyendo variables como

4.2. Problema de Predicción

Ahora nuestro objetivo es poder predecir, en primer lugar si una firma es innovadora o no. Y en segundo lugar queremos predecir las cuatro categorías: Innovativa e Innovadora, Sólo Innovadora, Sólo Innovativa, No innovativa y No innovadora. Ambos problemas de clasificación son símiles, por los plantearemos la metodología en forma general para ambos casos.[4]

Comenzaremos utilizando el método de Classification and Regression Trees (CART). Este método tiene una gran ventaja es que el proceso de decisión basado en arboles es fácilmente interpretable, presenta altas coincidencias con proceso *naive* de toma de decisión. Dado que el problema de innovación es de alta relevancia para los *policiy maker* estas ventajas comunicacionales son altamente deseables. Adicionalmente, con el algoritmo de *Weakest Link Pruning* (poda del eslabón más débil) podemos consiste en eliminar las ramas menos importantes o débiles del árbol para evitar el *overfitting*. Al quitar estas ramas menos significativas, el modelo simplificado resultante tiende a tener un rendimiento más sólido con nuevos conjuntos de datos, ya que elimina parte del ruido o la complejidad innecesaria. Esta técnica ayuda a mejorar la capacidad del modelo para generalizar y hacer predicciones precisas en datos no vistos.

La estimación de CART nos brinda una lógica de criterios de decisión y determinantes, basada los datos en los datos, que son fácilmente comunicables. CART explota los efectos no lineales entre las variables, dentro de cada nodo del proceso decisorio, pero si la estructura es altamente lineal, no solo dentro de cada nodo de decisión, CART puede no ser un buen en la predicción.

Debido a estar limitaciones, buscaremos otros métodos mas robustos. Para intentar mejorar

[4]La parametrización de todos los modelos es determinada mediantes el método de *K-fold Cross-Validation*.

la predicción aplicaremos los métodos de Bagging, Random Forest, Boosting. Con esto perdemos la alta interpretabilidad que nos brindaba CART, pero ganamos precisión. Con el fin de poder predecir si una firma es innovativa, innovadora o ambas.-

Bagging es otra técnica de conjunto en aprendizaje automático que se basa en entrenar múltiples modelos predictivos independientes utilizando subconjuntos aleatorios de datos de entrenamiento. Cada modelo se entrena con una muestra aleatoria con reemplazo de los datos originales. Luego, para hacer predicciones, se promedian las predicciones de todos los modelos (en el caso de regresión) o se toma la votación (en el caso de clasificación).

Random Forest es una extensión de la técnica de Bagging que se enfoca principalmente en árboles de decisión. En lugar de entrenar múltiples árboles de decisión con el conjunto de datos original, Random Forest introduce aleatoriedad adicional al seleccionar un subconjunto aleatorio de características en cada división de árbol. Luego, realiza la predicción combinando los resultados de múltiples árboles de decisión, reduciendo la tendencia al sobre ajuste y mejorando la precisión predictiva.

Boosting mejorar el rendimiento de un modelo combinando múltiples modelos más débiles en un único modelo fuerte. A diferencia de Bagging, en Boosting los modelos se entrenan secuencialmente, y cada nuevo modelo se enfoca en corregir los errores de predicción hechos por los modelos anteriores. En cada iteración, los pesos se ajustan para dar más importancia a los ejemplos que fueron mal clasificados anteriormente, lo que lleva a un modelo final más robusto y preciso.

En resumen, Bagging se basa en entrenar múltiples modelos independientes y promediar sus predicciones, Random Forest es una variante de Bagging que se enfoca en árboles de decisión con selección aleatoria de características, mientras que Boosting construye un modelo fuerte a partir de modelos débiles, dando más énfasis a los casos difíciles en cada iteración. Estas técnicas brindan ventajas en la predicción por sobre CART.

Por ultimo, a pesar de que los modelos son entrenados y testeados mediantes Cross Validation, es interesante ver si su capacidad predicativa alcanzada se mantiene para otros años. La existencia de la ENDEI III (tercer operativo), que abarca los años 2019 a 2021, nos brinda una posibilidad de poner a prueba la capacidad de nuestro modelo.

5. Conclusiones

Este trabajo busca explotar las técnicas de aprendizaje automático para explorar el problema de innovación en firmas en países en desarrollo, basándose en el caso argentino y explotando la riqueza de los datos de la ENDEI. Tras atenuar el problema de dimensionalidad en los datos mediante PCA, estimamos y validamos diferentes modelos predictivos para clasificar en primer lugar a una firma como innovadora, y luego caracterizar a las firmas en las 4 categorías que sugiere el Manual de Oslo: Innovativa e Innovadora, Sólo Innovadora, Sólo Innovativa, No innovativa y No innovadora.

Con este fin comenzamos proponiendo una metodológica basada en arboles, adoptando la metodología de CART (*Classification and Regression Trees*) la cual brinda grandes ventajas en su interpretabilidad y comunicación de los resultados. Esto hace de CART una metodología valiosa en un contexto de policy donde poder transmitir nuestros resultados es de gran importancia, además nos brinda una noción básicas de que observables parecieran determinar la innovación en las firmas, basándose en los datos.

En una segunda etapa buscamos conseguir un modelo que nos genere una mayor capacidad predicativa. Para esto implementamos diferentes métodos como: Bagging, Random Forest, Boosting. Esto métodos tienen ventajas en la predicción por sobre CART pero a coste de una perdida en la interpretablilidad. Todos estos modelos son entrenados y validados mediantes Cross-Validation con el fin de encontrar un modelo que sea el mejor prediciendo fuera de la muestra.

Esto genera aportes relevantes tanto para el ámbito académico como de policy. Por un lado, la incapacidad de determinar el carácter innovativo de una firma es una de las barrearas principales en la literatura (Verhoogen, (2021)). Esta barrera se intensifica en países en desarrollo, ya que la innovación suele ser inobservable y los supuestos necesarios para usar la TFP cómo proxy suelen no cumplirse. Nuestro trabajo brinda un primer acercamiento a la resolución del problema con un enfoque de Machine Learning. Permitiendo predecir el carácter innovativo de una empresa y además entender cuales son las principales características de las firmas que innovan. Esto puede informar modelos teóricos y motivar futuros trabajos empíricos relacionados.

Además, desde una perspectiva de policy-making es muy relevante tener un modelo que permita determinar el carácter innovativo de una firma en base a variables observables y/o relativamente

fáciles de conseguir. Siendo la innovación uno de los principales drivers del desarrollo económico, tener información a nivel firma sobre la cual accionar es de gran utilidad.

Referencias

ENDEI – MINCyT y MTEySS, disponbile en:

https://www.argentina.gob.ar/ciencia/indicadorescti/documentos-de-trabajo/innovacion/endei-i

Griliches, Z., & Mairesse, J. (1997), Production Functions: The Search for Identification, No 97-30, Working Papers, Center for Research in Economics and Statistics.

OECD/Eurostat/European Union (1997), Proposed Guidelines for Collecting and Interpreting Technological Innovation Data: Oslo Manual, The Measurement of Scientific and Technological Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264192263-en.

OECD/Eurostat (2007), Oslo Manual: Guía para la recogida e interpretación de datos sobre innovación, 3ª edición, Tragsa, Madrid, https://doi.org/10.1787/9789264065659-es.

OECD/Eurostat (2018), Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg, https://doi.org/10.1787/9789264304604-en.

Olley, G. S., & Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. Econometrica, 64(6), 1263–1297. https://doi.org/10.2307/2171831

Shamika, S., Bell, B., Fajarnés, P., Gonzalez, A., Lim, M., Ok, T., Solomon, A. & Ting, B. (2018). Technology and Innovation Report 2018. Harnessing Frontier Technologies. United Nations Conference on Trade and Development—UNCTAD.

UNCTAD — United Nations Conference on Trade and Development. (2021). Technology and Innovation Report 2021. Catching Technological Waves. Innovation with Equity. United Nations.

United Nations — Department of Economic and Social Affairs. (2022). Transforming Our World: The 2030 Agenda for Sustainable Development. United Nations.

Verhoogen, E., (2021). "Firm-Level Upgrading in Developing Countries,ÏZA Discussion Papers 14858, Institute of Labor Economics (IZA).

Descripción de los Datos

Tabla 1: Variables Relevantes

Variable	***Descripción***
ide_endei_II	*Código identificador de la empresa*
Rama_act	*Rama de actividad de la empresa*
Tam_mue	*Tamaño de empresa*
p.1.10.a	*Primer cliente/comprador de su producción*
Factor2	*Factor de expansión*
Outlier	*Outlier en Gastos de AI*
Perfil_inn	*Perfil de innovación*
p.1.7	*Año inicio de la actividad*
p.1.8	*Origen del capital*
p.1.9	*Tipo de empresa*
p.1.11.1	*Proporción de las ventas de su establecimiento que es absorbida por su cliente más importante*
p.1.12	*Alcance geográfico CLIENTES: Localidad, alrededores y Regiones*
p.1.14	*Exportaciones en las ventas totales de # (2014-2016)*
p.1.15.a	*Primer proveedor para su producción*
p.1.15.b	*Segundo proveedor para su producción*
p.1.16	*Alcance geográfico PROVEEDORES: Localidad, alrededores y Regiones*
p.2.#.#	*Conjunto de preguntas sobre practicas de gestión*
p.3.#.#	*Conjunto de preguntas sobre tipos de innovación*
p.4.#.#	*Conjunto de preguntas sobre resultados de innovación*
p.5.#.#	*Conjunto de preguntas sobre descentralización de la innovación*
p.6.#.#	*Conjunto de preguntas sobre protección (Ej. Patentes)*
p.7.#.#	*Conjunto de preguntas sobre Fuentes, motivaciones y obstáculos*
p.8.#.#	*Conjunto de preguntas Conocimiento de fuentes de financiamiento*
p.9.#.#	*Conjunto de preguntas vinculaciones con otras empresas*
p.10.#.#	*Conjunto de preguntas sobre administración de personal*
p.11.#.#	*Conjunto de preguntas sobre codificación de actividades*
p.12.#.#	*Conjunto de preguntas sobre medio ambiente*
Ingr_Total_#	*Total ingresos año # (2014-2016)*
Prop_Ingr_nocorr_#	*Total ingresos año # (2014-2016)*
Prop_Egr_sueldo_#	*Proporción de sueldos sobre egresos totales # (2014-2016)*
Prop_Otr_egr_corr_#	*Proporción de otros egresos corrientes sobre egresos totales # (2014-2016)*
Prop_egr_nocorr_#	*Proporción de otros egresos no corrientes sobre egresos totales # (2014-2016)*
Cant_NivTotal_#	*Dotación total 2014*
Inno_Idint	*Investigación y desarrollo interna*
Inno_Idext	*Subcontratación de I+D*
Inno_Dindus	*Diseño industrial e ingeniería*
Inno_Maq	*Adquisición de maquinaria y equipos*
Inno_HwSw	*Adquisición de hardware y software para innovación*
Inno_Tec	*Transferencia tecnológica*
Inno_Capac	*Capacitación para la introducción de innovaciones*
Inno_Consul	*Consultorías*

Fuente: ENDEI – MINCyT y MTEySS.

Notas: También se cuenta con información de personal, especializado y no especializado, remuneraciones. Monto y proporción de presupuesto o gasto dedicado a las diferentes formas de innovación. Y fuentes de financiamiento adoptadas. Preguntas sobre acceso y uso de internet en el proceso (Ej. si realiza e-commerce). Medidas de Valor Agregado de las firmas y VA por trabajador.