**Un modelo predictivo sobre el retiro voluntario**

|  |  |
| --- | --- |
|  |  |
| Ing. Plácido E. Maidana  *(placidomaidana@gmail.com)* | |
| Universidad Tecnológica Nacional (FRRe) | |

Resumen

El propósito del siguiente trabajo es encontrar un modelo de inteligencia artificial para aplicarse al Instituto de Seguridad Social, Seguros y Préstamos (InSSSeP). El mismo debe permitir una predicción certera del número de jubilaciones anticipadas para el próximo año.

Este documento, realiza un estudio exploratorio, referente a los temas de jubilación y “Random Forests”. Random forests es un algoritmo de inteligencia artificial que puede utilizarse tanto en problemas de regresión, o de clasificación como es el presentado en este artículo. Utilizaremos la librería scikit-learn para el desarrollo de un modelo que permita hacer posible la solución.

Palabras clave:Jubilación anticipada, Random forests, InSSSeP, Algoritmos de clasificación.

Abstract

The purpose of the following work is to find an artificial intelligence model to be applied to the Institute of Social Security, Insurance and Loans (InSSSeP). It should allow an accurate prediction of the number of early retirements for next year.

This paper conducts an exploratory study on the topics of retirement and "Random Forests". Random forests is an artificial intelligence algorithm that can be used in regression or classification problems such as the one presented in this paper. We will use the scikit-learn library to develop a model to make the solution possible.

Keywords: Early retirement, Random forests, InSSSeP, Classification algorithms.

# Introducción

El Instituto de Seguridad Social, Seguros y Préstamos (InSSSeP), tiene a su cargo, la caja de previsión social de la provincia del Chaco. Dicho organismo está regulado por Ley 4044, la cual es una Normativa provincial sancionada por la cámara de diputados de la provincia. La ley establece en el artículo cinco con título Obligaciones de la entidad, el funcionamiento de la Caja Complementaria Financiera, tendiente a asegurar el equilibrio financiero del organismo y concurrir con los montos necesarios para liquidar y abonar los beneficios jubilatorios establecidos.

En el Art 34: habla de la caja complementaria financiera la cual funcionará precisamente como departamento responsable de invertir las sumas sobrantes del organismo, a efecto de mantener el valor constante debidamente actualizado y cumplir con los objetivos de la presente Ley(LEY 4.044,1994).

La norma regula el accionar de la institución exigiéndole el cumplimiento de sus obligaciones e instruyendo la creación de un fondo para realizar movimientos financieros a fin de que sus fondos no pierdan valor.

De sus funciones se desprende que la caja de previsión social necesita conocer el presupuesto requerido para cumplir con las prestaciones. Poder determinar el momento en que sus afiliados empezaran a usufructuar sus beneficios jubilatorios permitirá al InSSSeP la confección de dicho presupuesto de una manera precisa.

Hay una tendencia de las personas a permanecer en la fuerza laboral por más tiempo para usufructuar un mejor salario demorando así su retiro. Esta predicción es compleja y tiene todas las características para que sea un problema a abordar con herramientas de aprendizaje automático supervisado.

Las técnicas de inteligencia artificial se han vuelto muy populares en las organizaciones públicas y privadas debido a que permiten un proceso de toma de decisiones más preciso. Las compañías de seguros privadas se han aventurado en este campo mediante la implementación de algoritmos que permiten una mejor comprensión de los datos disponibles. El conocimiento de las decisiones de jubilación permite a las compañías de seguros detectar el retiro temprano en un momento dado para tener una provisión presupuestaria adecuada. (Salazar & Boado-Penas., 2019)

En este trabajo, se utilizara “Random forests” una técnica mejorada de los árboles de decisión que tienen una ventaja significativa como lo explica (Navlani, 2018) ”Los árboles de decisión profundos pueden sufrir un sobreajuste, pero los bosques aleatorios previenen el sobreajuste al crear árboles en subconjuntos aleatorios”.

# Marco Teórico

Dos razones fundamentales explican que se haya creado la jubilación como tal. La primera es que los trabajadores en algún punto llegarían al límite de sus capacidades debido a su edad. En segundo lugar, difícil que un trabajador de edad avanzada consiga trabajo, ya que los empleadores prefieren contratar personas más jóvenes.

Hoy los sistemas de jubilación se encuentran amenazados debido a cambios demográficos. La tasa de dependencia total de la población no activa con respecto a la población activa tiende a seguir creciendo. El envejecimiento demográfico progresivo se mide por los siguientes factores: la tasa de dependencia demográfica de las personas mayores y la tasa de dependencia económica (Mendryk, 2017). Las empresas que se ocupan de las previsiones sociales se encuentran amenazadas no solo por la demografía sino también por las jubilaciones anticipadas.

Son muchas las variables, que pueden intervenir en las decisiones de jubilación anticipada como puede leerse en (Salazar & Boado-Penas., 2019) hay variables que son del entorno como lo es la economía, o como lo son eventos fortuitos como el COVID-19, o propios del entorno laboral. Existen otros indicadores que pudieran ser de una particular utilidad para nuestro análisis.

Pasaremos a describir estas características que poseen una relación bastante estudiada como lo podemos ver en (Salazar & Boado-Penas., 2019) (Mendryk, 2017)

Podemos ver cuáles son las características que intervienen en la decisión de jubilarse de forma anticipada.

Los años de aporte: Si los trabajadores cumplen con los requisitos legales para solicitar el beneficio tendrán menos motivación por seguir en la fuerza laboral.

La edad: A mayor edad mayor el deterioro y cansancio físico para realizar las labores ordinarias.

Salario: Con un salario mayor, se cuentan con más reservas para compensar la disminución del salario por el retiro.

Estado civil: Con más compromisos los trabajadores procuraran estar más tiempo en el grupo laboral activo.

Dependientes: Se trata de los hijos, se sigue aplicando la lógica anterior a mayor número de obligaciones, mayor la motivación a continuar con la actividad laboral.

Sexo: El sexo es una razón poderosa para continuar en la actividad laboral. En la región el patrón cultural indica que los hombres son quienes deben ser el soporte de la familia.

Enfermedad: La enfermedad hace que la experiencia laboral se haga extenuante sino imposible.

# Metodología:

# Preparación de datos

Para el siguiente proyecto se utilizan datos provistos por la Obra Social de la provincia, los mismos fueron discretizados para evitar vulnerar la privacidad de los afiliados.

Para la elección de las características se realizó un estudio sobre las características Demográficas Socioeconómicas y Previsionales que afectan al fenómeno en estudio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Caracte-rísticas** | **Descripción** | **Valor Original** | **Valor Discretizado** | **Obser-vacio-nes** |
| Edad | edad | En años. | Se discretizó en 3 valores según tabla 1 del anexo |  |
| AAporte | Años de aporte | En años | En años |  |
| Sexo | sexo | Hombre- Mujer | 0= Hombre  1= mujer |  |
| Enfermedad | Esta variable es 1 si la persona tiene una condición relacionada con las 10 enfermedades mortales más comunes en Argentina. | Cantidad en años desde que se detectó la enfermedad | En años |  |
| NivEd | Nivel de educación |  | Se discretizó en 2 valores según tabla 1 del anexo |  |
| Estado civil | Estado civil |  | 1 si la persona está casada, 0 en caso contrario. |  |
| Salario | El último salario obtenido en el momento de la jubilación. | Valor numérico decimal |  |  |
| Dependientes | El número de personas que dependen económicamente del | Valor numérico | Se discretizó en 5 valores según tabla 3 del anexo |  |

**Cuadro 1: Codificación del Conjunto de Datos**

Se muestra a continuación la descripción de la Etiqueta “Jubilado”

|  |  |  |
| --- | --- | --- |
| **ETIQUETA** | **Valor Original** | **Valor Discretizado** |
| **Internación** | S: Retirado antes de los 65  N:Retirado después de los 65 | 0: N  1: S |

**Cuadro 2: Etiqueta y su valor**

Valor discretizado obtenido de la base de datos del sistema de la Obra Social de la provincia

.

# Modelo de predicción

Como se muestra en el cuadro 2 la etiqueta Jubilación puede asumir dos valores, por lo tanto se trata de una clasificación binaria.

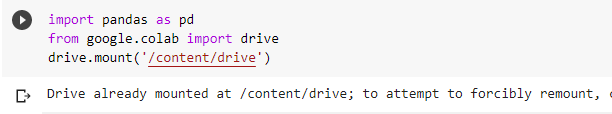
En este trabajo se entrenará un modelo de Random forests classifier. Random forests puede usarse de dos manera tanto para variables continuas, regresión, o para una clasificación. Para el problema que nos proponemos será usado como un clasificador.

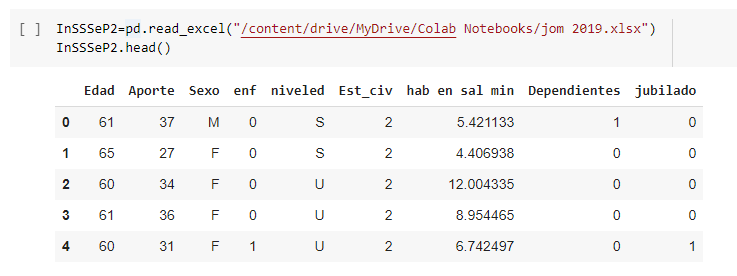
Para el entrenamiento del modelo y la obtención de las métricas de error se utilizará la librería Scikit-learn de Python.

# Aplicación

Se utilizara Google Colabaratory para la implementación del algoritmo de clasificación.

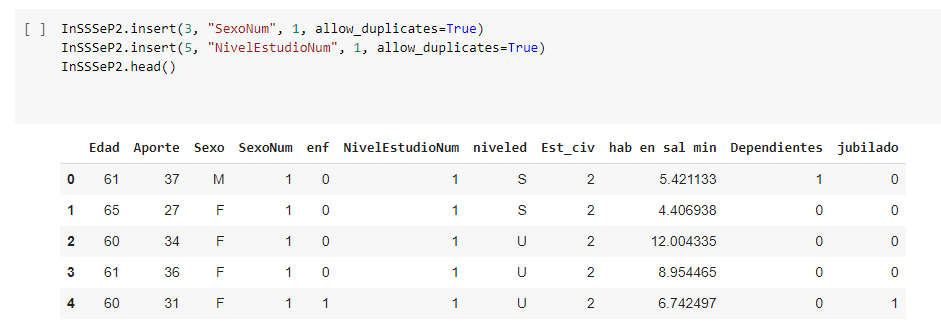
Lo primero será la carga del dataset a la plataforma.

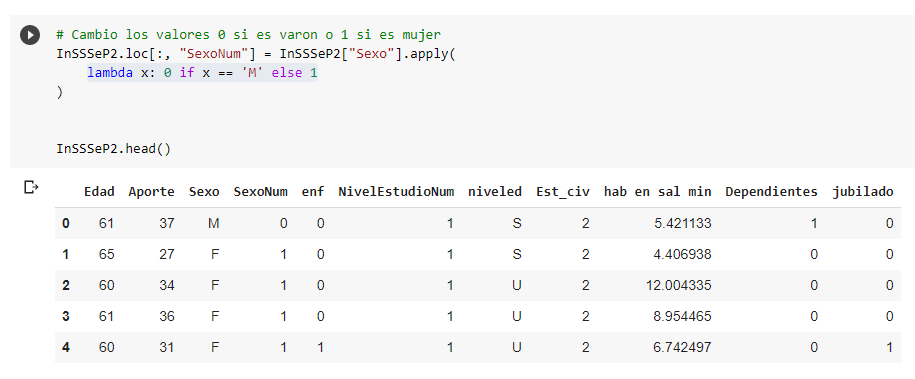




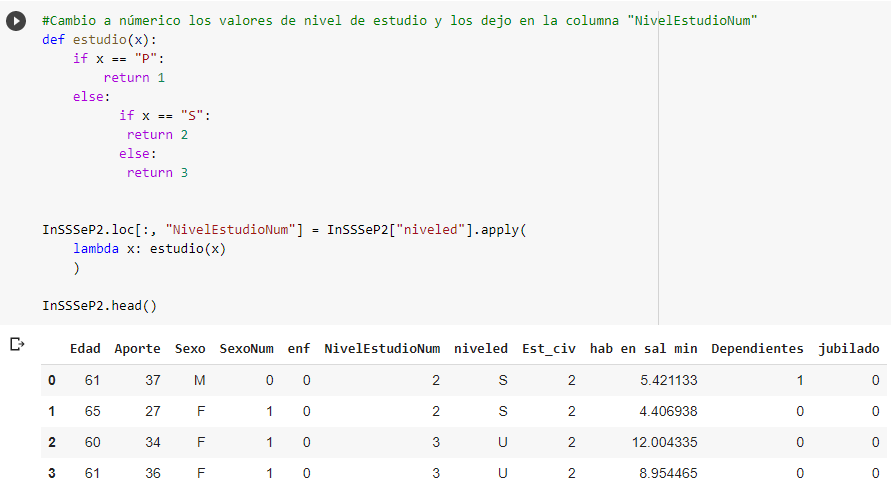
Podemos ver que en los datos encontramos algunas características que no son del tipo numéricos. Por lo que los modificaremos para tener toda la información en formato numérico.

Se creara dos campos nuevos en el dataset “sexoNum” y “NivelEstudioNum”. Luego se completaran el campo SexoNum con 0 si se trata de un hombre, o 1 si se trata de una mujer.

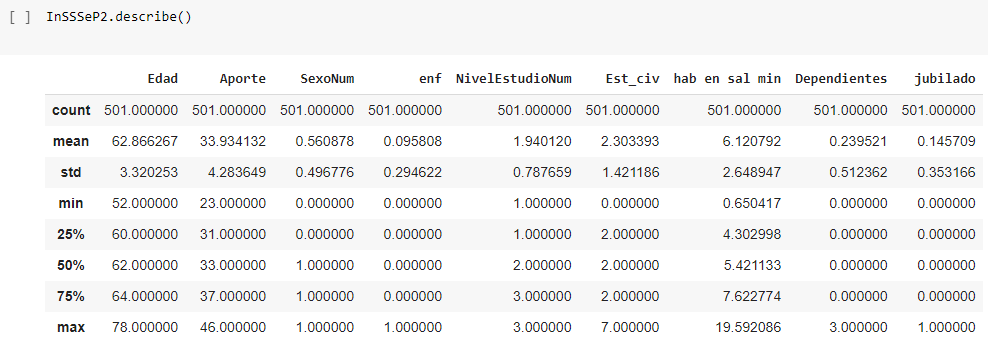




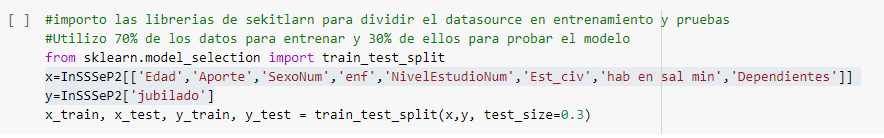
Llenaremos el campo “NivelEstudioNum” con números. Pondremos uno cuando el nivel de estudio es primario o secundario, 2 cuando el nivel de estudio es superior, 3 cuando el nivel de estudio es universitario postuniversitario.



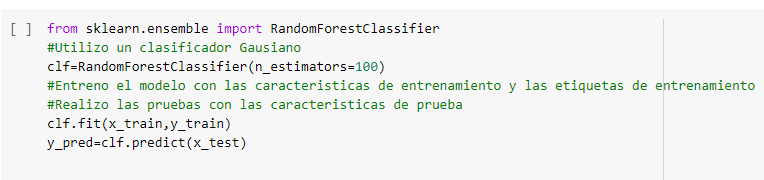
Depurados los datos hacemos una exploración a sus características generales.



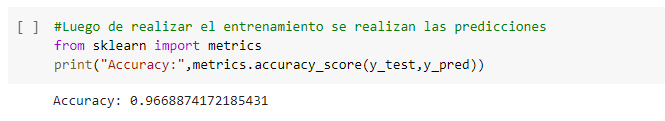
Se tomara como estrategia, tomar el 70% de los datos para realizar el entrenamiento del modelo y el 30% para comprobaciones. En el caso de random forests los riesgos de sobreajuste son menores que las aplicaciones sobre árboles por lo que podemos suponer una buena predicción con estos porcentajes.



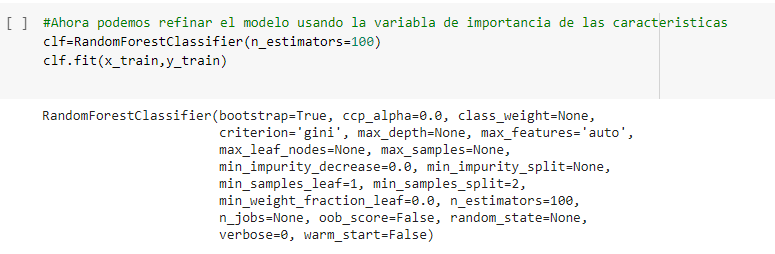
Realizamos la clasificación y las pruebas con los datos de prueba.

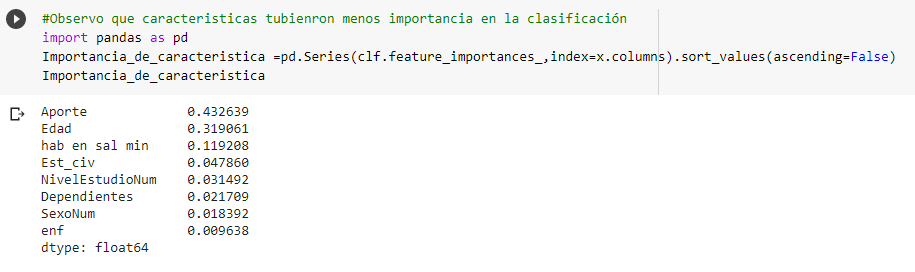


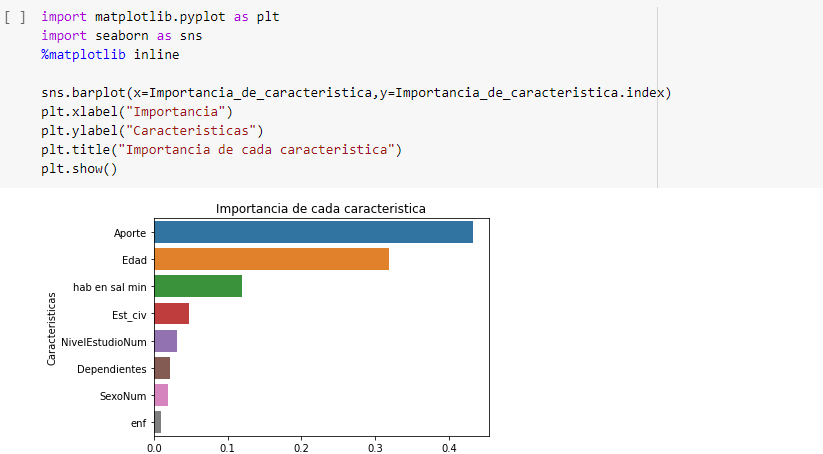
Medimos la precisión del modelo.



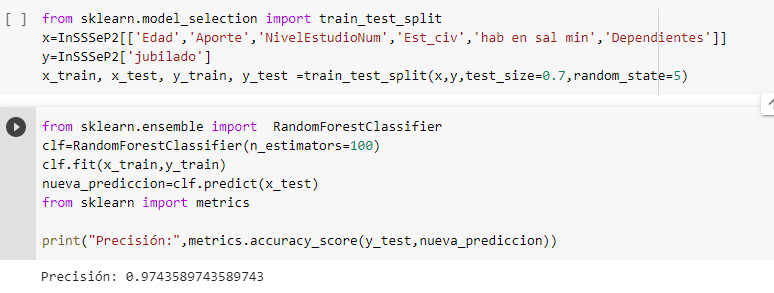
La precisión del modelo es bastante buena pero aun así podemos ir más lejos. Usaremos la librería “scikit-learn” para que nos indique que características son más importantes en la clasificación y de esa manera realizar ajustes al modelo de aprendizaje.







Podemos observar que existen características que aportan poco al modelo como ser “SexoNum” y “enf” las quitaremos y volveremos a correr el modelo.



Como podemos ver la precisión a aumentado y la estrategia usada haciendo uso de la plataforma Google Colaboratory fue sencilla y exitosa.

Conclusiones

La aplicación de Random Forest haciendo uso de la librería “scikit-learn” fue exitosa y promete buena precisión. Con el modelo presentado el InSSSeP podrá determinar el número de jubilaciones anticipadas previstas para el próximo año. Con esta información podrá conseguir mayor precisión en el presupuesto del año próximo.

Puede verse que la aplicación de inteligencia artificial está al alcance de cualquier institución. Solo vasta ser creativos, y analiticos a la hora de encontrar las variables dependientes de los problemas que esperamos resolver.

# Referencias

Bagnato, J. I. (03 de 03 de 2020). *Sets de Entrenamiento, Test y Validación*. Obtenido de Aprende Machine Learning: https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/

De Los Santos, P. R. (28 de 04 de 2020). *Datos de entrenamiento vs datos de test*. Obtenido de Think Big Empresas: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/

LEY 4.044,1994,Instituto de seguridad social, seguros y prestamos (InSSSeP)

Navlani, A. (16 de 05 de 2018). *Clasificadores de bosques aleatorios en Python*. Obtenido de DataCamp: https://www.datacamp.com/community/tutorials/random-forests-classifier-python

Salazar, J. d., & Boado-Penas., M. d. (2019). *SCORING AND PREDICTION OF EARLY RETIREMENT USING MACHINE LEARNING TECHNIQUES: APPLICATION TO PRIVATE PENSION PLANS.*

Segal, M. (2004). *Machine Learning Benchmarks and Random Forest Regression.*

Ying, X. (2019). *An Overview of Overfitting and its Solutions.* J. Phys.: Conf. Ser. 1168 022022.

**Anexo**

**Tabla 1: Estado civil**

|  |  |
| --- | --- |
| Codigo | Descripcion |
| 0 | Sin Datos |
| 1 | Soltero |
| 2 | Casado |
| 3 | Separado de Hecho |
| 4 | Unido de Hecho |
| 5 | Separado Legalmente |
| 6 | Viudo |
| 7 | Divorciado |

**Tabla 2: Discretización del Nivel de Estudio**

|  |  |
| --- | --- |
| **Nivel de Estudio** | **valor** |
| Significa que la persona tiene educación primaria o secundaria. “P” | **1** |
| Superior “S” | **2** |
| Estudios de grado. “U” | **3** |
| Estudios de posgrado. “POST” | **4** |

**Tabla 3: Discretización del número de Dependientes**

|  |  |
| --- | --- |
| **Dependientes** | **Valor** |
| Ningun dependiente | 0 |
| Un dependiente | 1 |
| Dos dependientes | 2 |
| Tres dependientes | 3 |
| Cuatro o más de cuatro dependientes | 4 |