UNIVERSIDAD DE LOS ANDES
Departamento de Ingeniería Industrial - Segundo Semestre de 2023
Statistical Learning for Data Analytics - IIND 4123
Profesor: Carlos Valencia

**PROYECTO FINAL**

- ∼ El informe del proyecto debe ser entregado antes del jueves 7 de diciembre a las 5pm
- ∼ La solución puede ser elaborada en grupos de máximo 3 personas
- ∼ Se recomienda no hacer el proyecto de manera indiviudal por la carga de trabajo que requiere
- ∼ Utilice procedimientos explícitos y análisis detallados

El proyecto consiste en analizar dos problemas basados en datos reales a partir de técnicas y modelos vistos en clase. En cada uno de estos problemas se realizará un competencia basada en modelos predictivos, uno para clasificación y otro para regresión. Cada grupo deberá resolver los dos problemas.

## I. Descripción del Problema 1: Music Year Prediction.

El proyecto consiste en utilizar datos reales para predecir el año en que fue lanzada una canción a partir de sis características del timbre en la grabación. en total, son 90 atributos predictores. Los 12 primeros correspondel al timbre promedio y los 78 siguientes a la covarianza. Originalmente, eset tipo de datos fue recolectado en un proyecto llamado *Million Song Dataset* de la Universidad de Columbia (`http://millionsongdataset.com/`). Lo datos a trabajar tienen los mismo predictores pero con canciones no consideradas en la base de datos original.

## II. Descripción del Problema 2: Bankruptcy.

La bancarrota (quiebra) es el resultado de una condición crónica que ocurre cuando el valor total de los pasivos de una compañía excede el valor de sus activos totales. Por lo tanto, saber de antemano cuándo una empresa se va a ir a la quiebra es información importante que impacta en muchos problemas de toma de decisiones empresariales. La predicción de la quiebra ha sido un problema importante para banqueros, inversionistas, administradores de activos, auditores y académicos. En particular, para las instituciones financieras, la quiebra tiene un impacto significativo en sus decisiones de préstamo y rentabilidad.

Por lo anterior, este proyecto consiste en la predicción de quiebra empresarial (si una empresa entrará en bancarrota o no al año siguiente), utilizando una serie de métricas financieras como variables predictoras. Lograr esto, como se estableció anteriormente, permite planear de mejor manera los préstamos que ofrecen entidades financieras a diferentes compañías. La descripción de las variables predictoras se presenta a continuación:

Fracaso: 1 si la empresa entró en bancarrota al día siguiente, 0 de lo contrario (variable de respuesta)
B11: cociente entre el flujo de caja y las ventas.
B12: cociente entre el flujo de caja y los activos totales.
B13: cociente entre el flujo de caja y el patrimonio neto.
B14: cociente entre el flujo de caja y la deuda total.
B21: cociente entre los ingresos netos y las ventas.
B22: cociente entre los ingresos netos y los activos totales.
B23: cociente entre los ingresos netos y el patrimonio neto.
B24: cociente entre los ingresos netos y la deuda total.
B31: cociente entre los pasivos corrientes y los activos totales.
B32: cociente entre los pasivos de largo plazo y los activos totales.
B41: cociente entre el efectivo disponible y los activos totales.
B42: cociente entre los activos líquidos (activos corrientes menos inventarios) y los activos totales.
B43: cociente entre los activos corrientes y los activos totales.
B44: cociente entre el Working Capital (activos corrientes menos pasivos corrientes) y los activos totales.
B45: cociente entre el patrimonio neto y los activos totales.
B46: cociente entre el patrimonio neto y los pasivos totales.
B47: cociente entre los pasivos y los activos totales.
B48: cociente entre el Working Capital y el patrimonio total.
B51: cociente entre el efectivo disponible y los pasivos corrientes.
B52: cociente entre los activos líquidos y los pasivos corrientes.
B53: cociente entre los activos corrientes y los pasivos corrientes.
B54: cociente entre el efectivo disponible y los pasivos totales.
B61: cociente entre el efectivo disponible y las ventas.
B62: cociente entre las cuentas por cobrar y las ventas.
B63: cociente entre el inventario y las ventas.
B64: cociente entre las ventas y los activos líquidos.
B65: cociente entre las ventas y los activos corrientes.
B66: cociente entre el Working Capital y las ventas.
B67: cociente entre las ventas y el patrimonio neto.
B68: cociente entre las ventas y los activos totales.
B69: cociente entre el efectivo disponible y los gastos operacionales.
B610: cociente entre los activos corrientes y los gastos operacionales.
B611: cociente entre el Working Capital y los gastos operacionales.
B612: cociente entre la utilidad bruta y las ventas.
B613: cociente entre la utilidad operacional y las ventas.
B81: cociente entre los gastos financieros y los pasivos.
B82: cociente entre los gastos financieros y los pasivos corrientes.
B83: cociente entre los gastos financieros y los activos.
B84: cociente entre los gastos financieros y las ventas.

`B85: cociente entre los gastos financieros y la utilidad`

Los datos a trabajar corresponden a empresas de *retail* en Colombia.

## III. Datos y Competencia.

Para obtener los datos, y la desripción de las variables, puede ir al sitio:
`https://www.kaggle.com/t/f0530f103835434e8c422a531ab4d49f`
para el problema de regresión, y para el problema de clasificación puede ir a:
`https://www.kaggle.com/t/019209cedd9048bc88753f7eb22c2602`
Podrá obtener un archivo con los datos de entrenamiento (`train.csv`) que incluyen tanto las variables como la respuesta. Además encontrará un archivo con datos de prueba (`test.csv`).
Parte del proyecto incluye participar en una competencia entre todos los grupos. Usted debe encontrar un modelo predictivo en cada caso (regresión y clasificación) y usarlo para predecir en los inputs proporcionados en los datos de prueba (`test.csv`). El desempeño de sus modelos será evaluado automáticamente por el sitio online (`kaggle.com`), y se hará el ranking de los grupos. Más instrucciones sobre el formato de la competencia, pueden ser revisados en los `url` suministrados.

## III. Guía para desarrollo de proyecto y elaboración de informe.

En general, cada grupo tiene libertad para desarrollar su proyecto. Tenga en cuenta que el objetivo fundamental es utilizar corrrectamente los conocimientos adquiridos en clase para encontrar un buen modelo predictivo. La evaluación no se basa únicamente en el desempeño de su algoritmo, sino en el buen desarrollo del modelo y la justificación de su uso.
Para encontrar un modelo apropiado, se recomienda que su proyecto contenga las siguientes etapas:

1. **Exploración y Visuaización de datos**. Antes de empezar a estimar funciones, es mejor si aduqiere un conocimiento previo de qué exactamente son los datos, cómo se relacionan entre sí y cómo afectan la variable de respuesta. Algunas referencias que puede usar para hace exploración y visualizaciń de datos las puede encontrar en el archivo anexo (`data-exploration.zip`), o puede revisar el link:
   `http://www.rdatamining.com/docs/data-exploration-and-visualization-with-r`

2. **Selección y Extracción de variables**. Uno de los resultados importantes que salen de la exploración de los datos, determinar qué variables, o qué transformación de ellas se debe usar en cada modelo. Referencias para realizar estos procesos se pueden encontrar en el capítulo 3 del libro *Applied Preditive Modeling* de Kunh y Johnson, el cual recomendamos leer. Tenga en cuenta que en esta etapa se debe usar la intución que se tiene del problema real, para determinar patrones que sean útiles para predecir o explicar la respuesta.

3. **Preparación y limpieza de datos**. Esta parte requiere mirar la validez de algunos datos y decidir que inofrmación no es útil. Los datos suministrados son bastante limpios y no requieren mucha limpieza. Procesos típicos de limpieza incluyen tratamiento de datos perdidos (*missing data*) y datos atípicos (*outliers*).

4. **Evaluación de modelos**. Acá se incluye el ajuste y la evaluación de diferentes modelos predictivos. Recuerde que es muy importante calibrar los modelos correctamente y escoger adecuadamente entre ellos.

5. **Selección de modelo final, análsis y conlcuiones**

## IV. Evaluación.

La nota del proyecto será sobre 100 puntos. El informe, con todos los detalles de la solución tendrá un valor de 90 puntos. El resto de la nota dependerá de su desmpeño en las competencias. Para cada problema (regresión y clasificación), habrá 15 puntos máximos disponibles (para el primer lugar). La bonificación para cada grupo será proporcional al ranking que ocupe. Por ejemplo, si un grupo queda justo en la mitad (mediana!!!) en la competencia de clasificación, bonificará 7.5 puntos. Note que la nota máxima que un grupo puede obtener e de 120/100.