Pronóstico de Graduación Estudiantes Subgraduados Universidad de Puerto Rico Recinto Mayagüez

Ramineh Lopez Yazdani[1] & Jesús D. Hernández Londoño[2]

[1]Universidad de Puerto Rico. Departamento de Matemáticas.
Recinto Mayagüez.

[2]Universidad de Puerto Rico. Departamento de Matemáticas.
Recinto Mayagüez.

9 de septiembre de 2020

Contenido

1. Objetivo
2. Antecedentes
3. Metodología
4. Referencias

Contenido

1. Objetivo
2. Antecedentes
3. Metodología
4. Referencias

Objetivo

Objetivo

Elegir un Modelo de Machine Learning, para la predicción de graduación de estudiantes subgraduados a tiempo o en general (150 %).

Contenido

1 Objetivo

2 Antecedentes

3 Metodología

4 Referencias

Antecedentes

Tasa de Graduación

Es la proporción de estudiantes que completan sus estudios dentro del 150 % (en general) del tiempo del programa o a tiempo, i.e., 8 años para programas de 5 años (Facultad de Ingeniería) y 6 años para los demás programas, excepto grados asociados.

Antecedentes

¿Por qué graduarse a tiempo?

Según "Complete College America Alliance of States"

- Grados Asociados pierden aproximadamente $\$35,000$ en salarios.
- Programas de 4 años $\$45,327$ en salarios.
- Prestigio y Acreditación

Causas

- Consejería académica insuficiente o inadecuada.
- Grupos grandes, ausentismo frecuente.
- Deficiencias en las materias básicas.
- Pocos cursos.

Antecedentes

Estrategias

- Establecer un sistema de alerta temprana.
- Consejería académica compuestos por estudiantes.

Tasa de Graduación UPRM

Según OPIMI en su informe del 2019, el 51,4 % de los estudiantes con cohorte en 2010, habría completado su grado al cabo del 150 % del tiempo.

Antecedentes

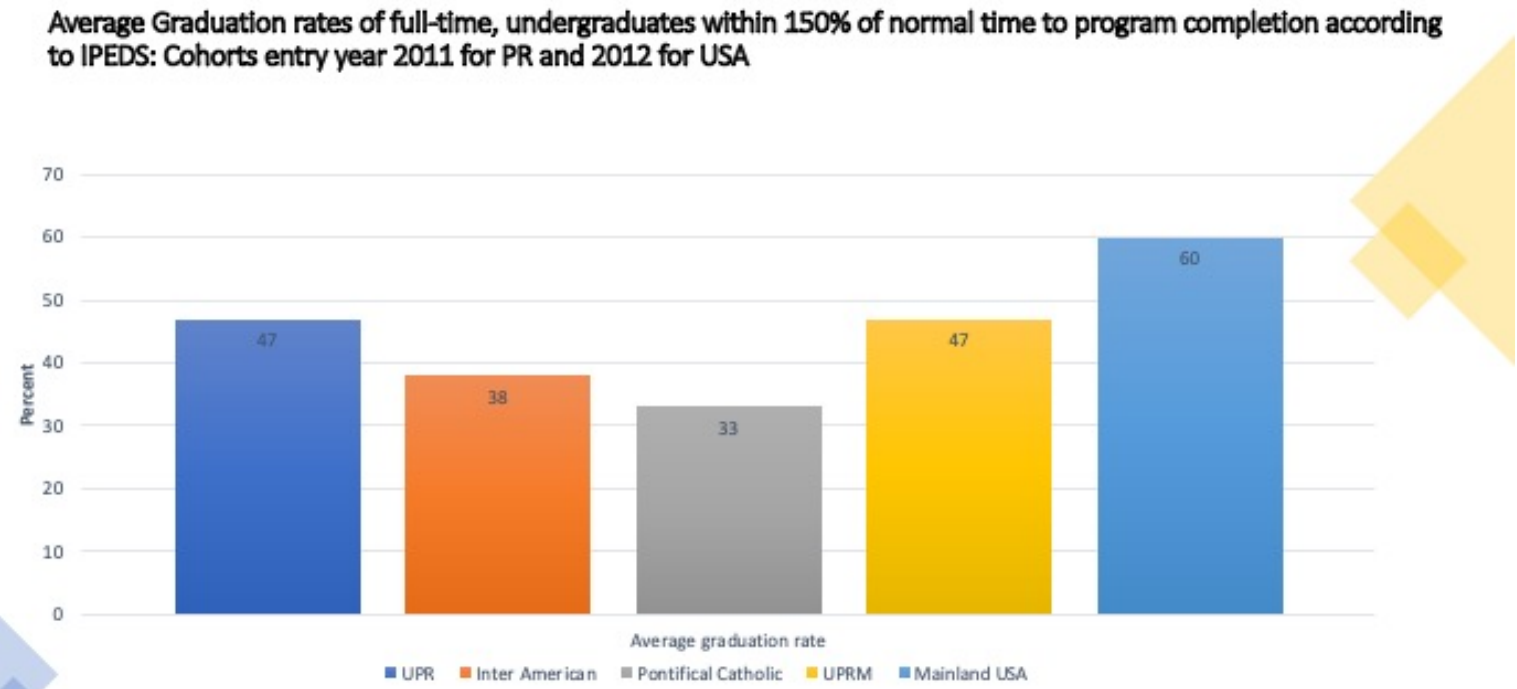


Figura: Autoría propia

Contenido

1 Objetivo

2 Antecedentes

3 Metodología

4 Referencias

Metodología

- Se inicia con un preprocesamiento, extracción y creación de variables, eliminación de datos duplicados, eliminación Na's
- Exploración de los datos.
- Comparar Modelos de Machine learning para validar predicciones. (Logistic Regression, Random Forests, Boosting, Support Vector Machines, Deep Learning).

Metodología

Software

- Rstudio
- tidyverse, dplyr, rpart, partykit, ROCR, gbm, RandomForest, caret, e1071, nnet.

Metodología

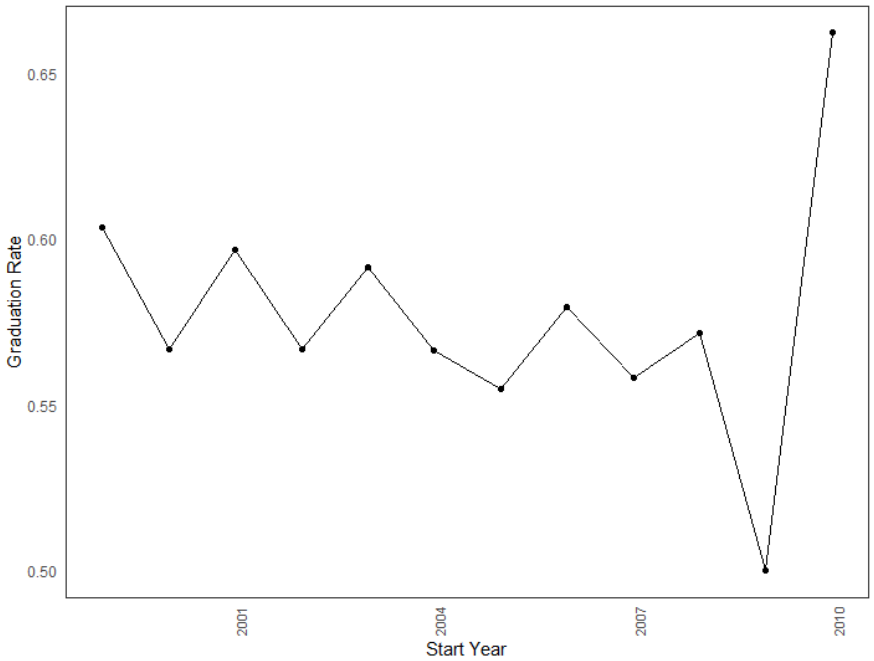
Data

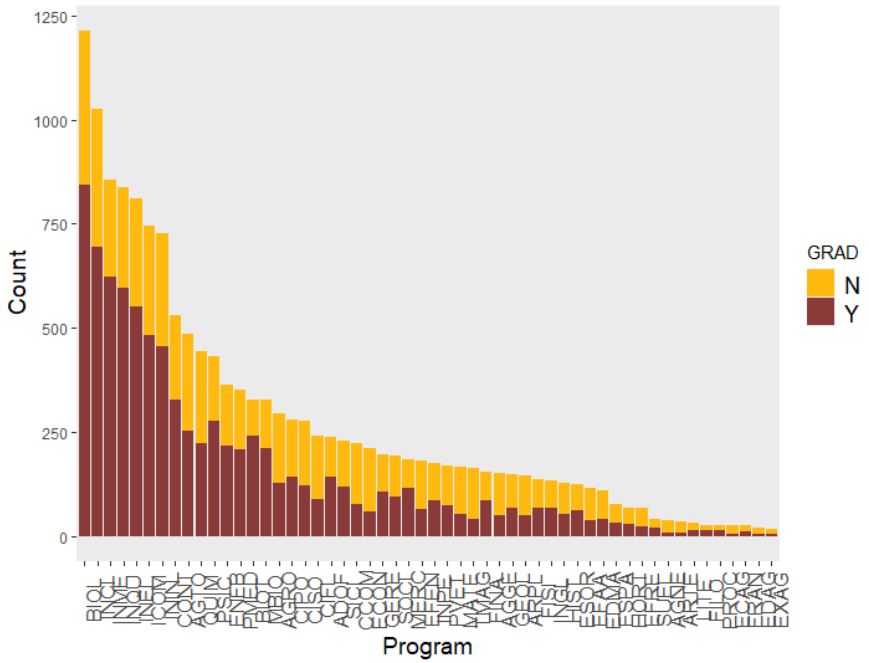
- Oficina de planeación, investigación y mejoramiento institucional (OPIMI), (43,515 ; 22), (43,515 ; 11), (18,955 ; 8), años de ingreso 1999-2018.
- Observaciones 18413, 21 variables.
- Cohorte 2010.

Variables

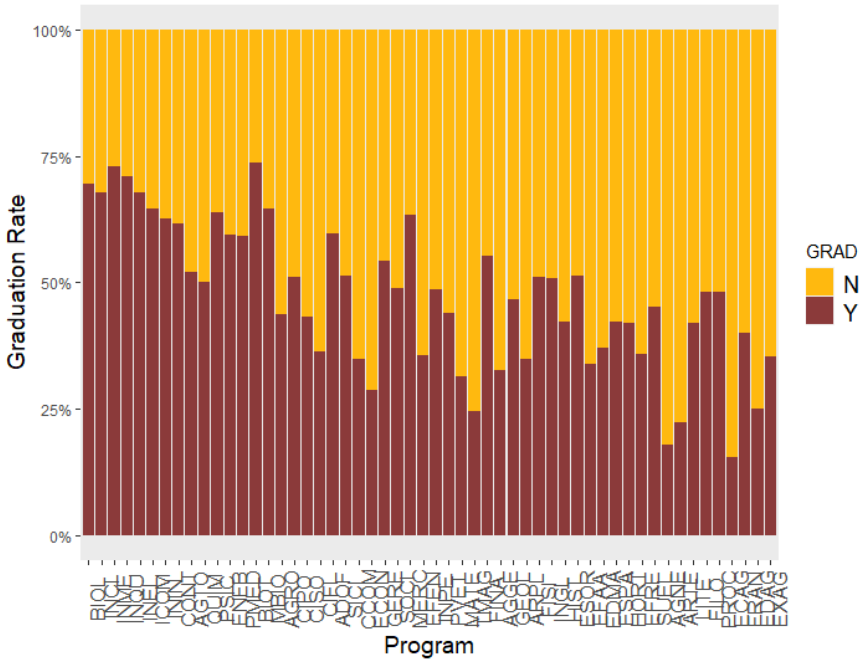
STUDENT RECORD KEY, Year, Faculty, Major, Apt Verbal, Aprov Matem, Apt Matem, Aprov Espanol, Aprov Ingles, Highschool GPA, INGRESO FAMILIAR, EDUC PADRE, EDUC MADRE, Gender, School Type, GPA 1ER ANO, GRAD, Year Grad, Rel Stud GPA, Rel School GPA, grad intime

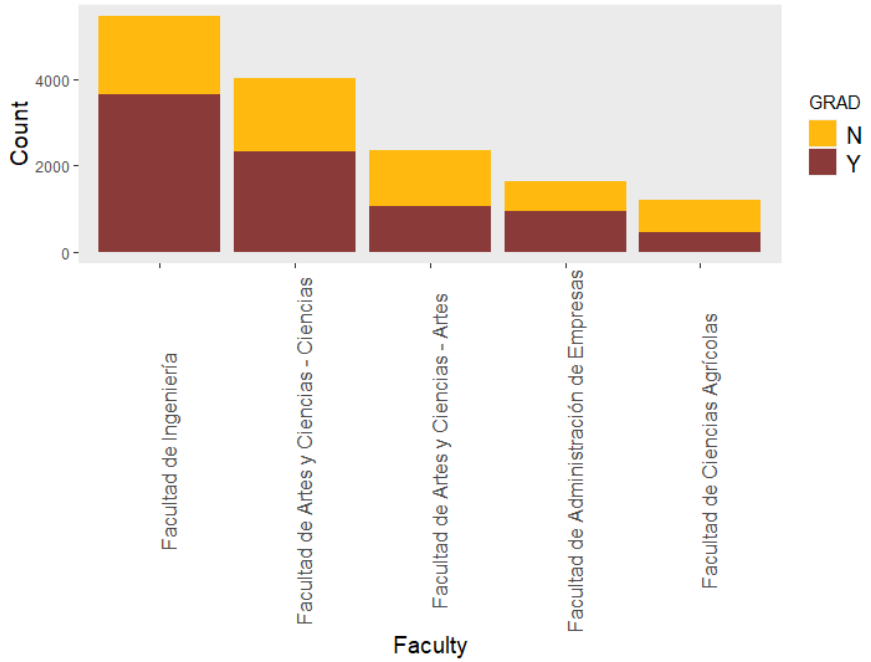
Metodología





Metodología





Metodología

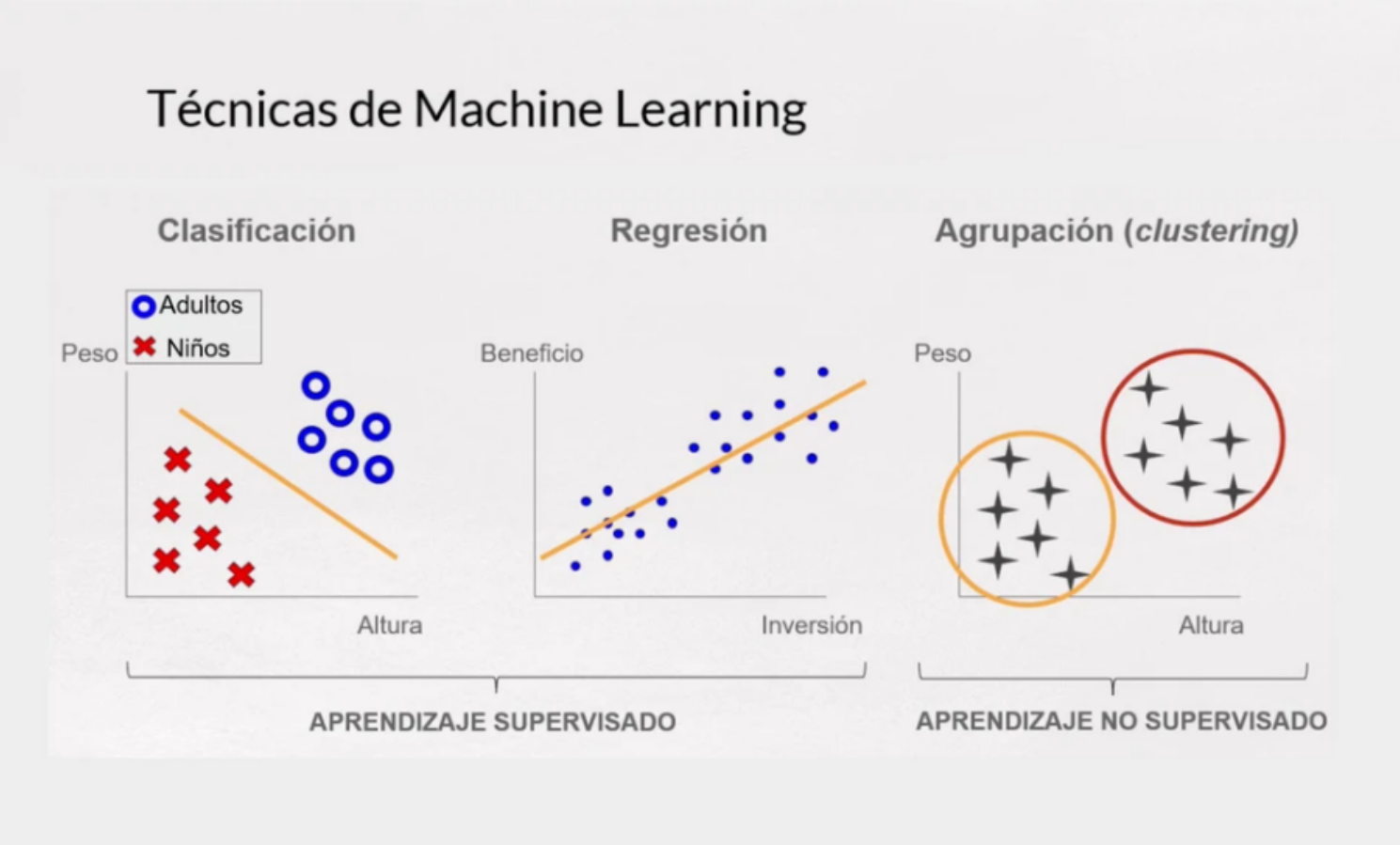


Figura: Extraída de OpenWebinars

Metodología

1. Clasificadores de Bayes
   - Linear Discriminant Analysis.
   - Naive Bayes.
   - k-nearest neighbors.
   - Logistic Regression.
2. Decision Trees.
3. Ensembles.
4. Neural Networks and Deep Learning.
5. Support vector machines.

Metodología

Matriz de confusión

Es una herramienta que permite conocer el desempeño de un algoritmo.

	Actual class=Yes	Actual class=No
Predicted class =Yes	True Positives=TP	False Positives=FP
Predicted class=No	False Negative=FN	True Negatives=TN

Figura: Acuña E.

- Accuracy=(TP+TN)/Total.
- Missclassification rate=(FP+FN)/Total = 1- Accuracy.
- True Positives Rate=Sensitivity=Recall=TP/(TP+FN).
- False Positives Rate =FP(/FP+TN).
- Specifity=TN/(FP+TN)=1-False Positive Rate

ROC

Es el punto de corte en el que se alcanza la sensibilidad y especificidad más alta. Evalúa la capacidad del modelo para clasificar correctamente. Un valor de 1 significa que el modelo es perfecto, un valor de 0.5 indica que el modelo no es útil.

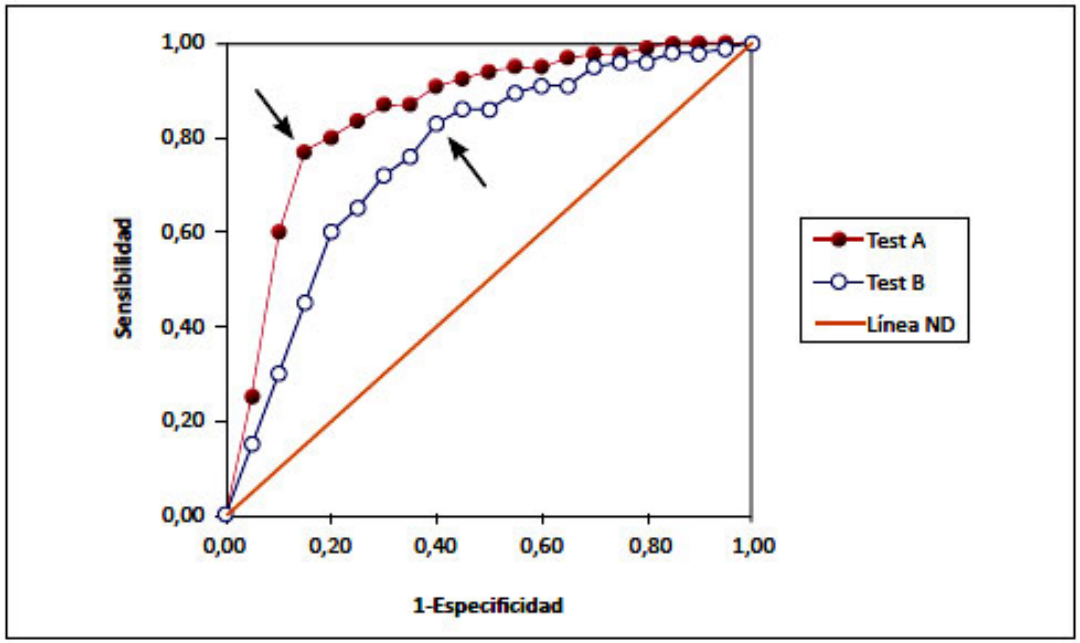


Figura: Jaime Cerda y Lorena Cifuentes.

Metodología

AUC

Es el área bajo la curva, Esta refleja que tan bueno es la prueba o modelo para discriminar pacientes con y sin la enfermedad.

Contenido

1. Objetivo
2. Antecedentes
3. Metodología
4. Referencias

Referencias

Tibshirani R., Hastie T., Witten D., James G.
An Introduction to Statistical Learning with applications in R.

Acuña E.
Data Mining and Machine Learning.
Decision Trees

Retención de Estudiantes.
http://ponce.inter.edu/.
Rodríguez A.

Referencias


University of Puerto Rico-Mayaguez.
https://nces.ed.gov/ipeds.
IPEDS (Integrated Postsecondary Education Data System)


Cerda J., Cifuentes L.
Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos.
Rev. chil. infectol. vol.29 no.2 Santiago abr. 2012


Complete College America Alliance of States.
https://completecollege.org/.