UNIVERSIDAD ESPÍRITU SANTO

ESAI BUSINESS SCHOOL

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

PROYECTO INTEGRADOR DE INTELIGENCIA DE NEGOCIOS

INFORME EJECUTIVO FINAL

TITULO: MODELO DE ANÁLISIS PREDICTIVO DE LOS FACTORES SOCIALES QUE INCIDEN EN LA APROBACIÓN DEL EXAMEN DE INGRESO A LA CARRERA DE CPA

AUTOR: CRISTHIAN GEOVANNY GUANOQUIZA CALERO

DOCENTE: ALEXANDRA JACQUELINE ARCINIEGAS CORAL

SAMBORONDÓN, 16 de junio del 2023

Modelo de análisis predictivo de los factores sociales que inciden en la aprobación del examen de ingreso a la carrera de CPA

A. Resumen Ejecutivo

Actualmente en la carrera de Contador Público Autorizado (CPA) de la Universidad “ABC”, para dar la prueba de ingreso se matriculan aproximadamente 350 aspirantes, de la cual pocos aprueban, habiendo un desconociendo de los motivos por el cual no acceden a un cupo universitario; llenando de incertidumbre a los aspirantes, es por ello que el problema abordado se centra en la aprobación del examen de ingreso a la carrera de Contador Público Autorizado (CPA) y cómo los factores sociales inciden en ese resultado.

Aprobar este examen es muy importante para el aspirante, debido a que, aquí comienza su carrera profesional; por ello se trata de identificar qué factores sociales pueden afectar positiva o negativamente el éxito profesional, el objetivo principal de esta investigación es desarrollar un modelo de análisis predictivo utilizando las técnicas de Decision Tree y Random Forest, para identificar y comprender los factores sociales que tienen el mayor impacto en la aprobación del examen de ingreso a la carrera de CPA.

Los resultados de la presente investigación demuestran que los factores sociales inciden en la aprobación del examen de ingreso a la carrera de CPA, esto debido a las técnicas aplicadas, en donde se analizó la influencia de cada una de las variables con el modelo predictivo, dando como mayor énfasis el estado civil de los padres.

El valor de la solución propuesta se basa en la aplicación de las técnicas de Decision Tree y Random Forest, el cual permitió corroborar que, si la mayoría de aspirantes en donde sus padres tengan un estado civil de “Unidos”, la aprobación de ingreso a la carrera de CPA será efectiva.

B. Resumen Ejecutivo Grafico

En la imagen 1, se puede visualizar el diagrama final del árbol podado con 4 nodos y sus porcentajes,

donde la variable que tuvo más énfasis fue “EPadres”, el cual representa el “estado civil de los padres”, en base a la variable de predicción se demuestra que, si el aspirante es de padres “Unidos” y el consumo alcohol es de: algunas veces y casi nunca, el aspirante puede aprobar; caso contrario, si el aspirante viene de un número de familia menor a tres, va a seguir aprobando, y por ultimo si el aspirante tiene tiempo libre en un nivel de: casi siempre y algunas veces, este puede aprobar, cabe resaltar que va a existir una mayor probabilidad de que, el aspirante apruebe el examen siempre y cuando los padres se mantengan en un estado de “Unidos”.

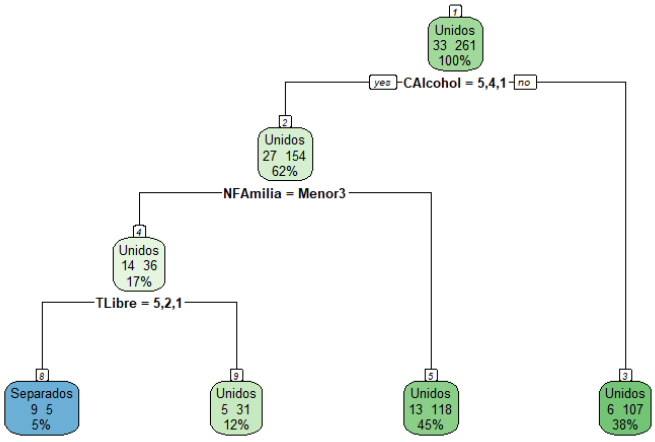


***Imagen 1**. Diagrama de Árbol podado con 4 nodos.*

En la imagen 2 muestra las variables más significadas usando la técnica Random Forest, en el cual se visualiza que la variable “TLibre” tiene más énfasis que las demás variables.

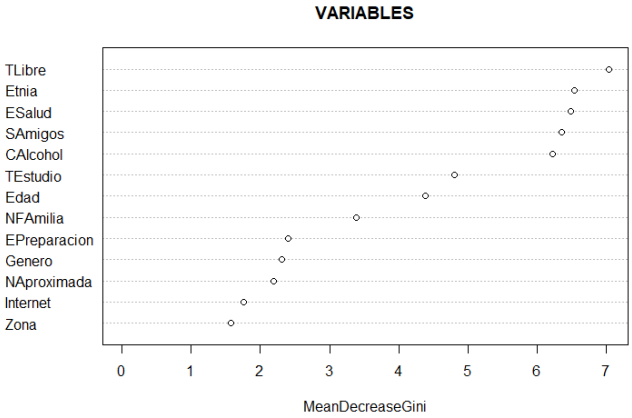


***Imagen 2**. Gráfico de variables significativas según la técnica de Random Forest,*

C. Detalle Técnico

1. Fuentes de Datos

La fuente de datos es todo aquello que nos proporciona información para reconstruir hechos y obtener conocimiento (Maranto & González, 2015), por ello se utilizó datos actuales proporcionados por los estudiantes de la carrera de Sistemas de la Universidad “ABC”. Los datos forman parte de los aspirantes que se presentan a rendir la prueba de admisión para entrar a la carrera de CPA (Contador Público Autorizado).

La base de datos cuenta con 5152 registros, 368 observaciones y 14 variables como se muestra en la imagen 3, de la cual se utilizó toda la población debido a que las observaciones son menores a 400 (López, 2004).

```
"| Observaciones:  368 | Variables:  14 | Registro:  5152 |"
```

***Imagen 3**. Datos del DATASET.*

Las variables que presenta el Dataset se visualizan en la imagen 4, en donde se aprecia la descripción adecuada de las variables utilizadas, además de su valor y niveles que fueron tomados por parte de los estudiantes de Sistemas.

Variables	Descripción	Valor	Niveles	
Genero	Genero	(Masculino, Femenino)	Nivel 1	5 => Excelente
Etnia	Etnia	(Afroecuatoriana, Mestizo, Indígena, Montubia, Otros)		4 => Muy Bueno
				3 => Bueno
Edad	Edad			2 => Malo
Zona	Zona	(Urbana, Rural)		1 => Muy Malo
NFAmilia	Numero de Familia	(Mayor a 3, Menor a 3)	Nivel 2	5 => Siempre
EPadres	Estado Civil de Padres	(Unidos, Separados)		4 => Casi Siempre
EPreparacion	Estudio de Preparación	(Si, No)		3 => Algunas Veces
TEstudio	Tiempo de Estudio	(Nivel 3)		2 => Casi Nunca
Internet	Internet	(Si, No)		1 => Nunca
SAmigos	Salida entre Amigos	(Nivel 2)	Nivel 3	1 => < 2 H
CAlcohol	Consumo de Alcohol	(Nivel 2)		2 => 2 - 5 H
ESalud	Estado de Salud	(Nivel 1)		3 => 5 - 10 H
TLibre	Tiempo Libre	(Nivel 2)		4 => > 10 H
NAproximada	Nota Aproximada	(Aprobado, Reprobado)		

***Imagen 4**. Descripción de variables.*

2. Técnica

La técnica que se utilizo fue Decision Tree, el cual es un método usado en distintas disciplinas para predicción (Mendoza, 2018), esta técnica permite observar que variable incide en la aprobación del examen de ingreso, ya que, en base a un diagrama se pude visualizar el posible resultado y tomar una decisión adecuada, debido a que su modelo en la estructura de un árbol junto con nodos de decisión y nodos hoja (Talari, 2022), por ello es fácil de visualizar, debido a que su modelo es ajustable.

```
90
91  bd_estudiantes$Genero <- as.factor(bd_estudiantes$Genero)                                  # 0.6081
92  bd_estudiantes$Etnia <- as.factor(bd_estudiantes$Etnia)                                    # 0.2703
93  bd_estudiantes$Zona <- as.factor(bd_estudiantes$Zona)                                      # 0.7162
94  bd_estudiantes$NFAmilia <- as.factor(bd_estudiantes$NFAmilia)                              # 0.5405
95  bd_estudiantes$EPadres <- as.factor(bd_estudiantes$EPadres)                                # 0.8784
96  bd_estudiantes$EPreparacion <- as.factor(bd_estudiantes$EPreparacion)                      # 0.6216
97  bd_estudiantes$TEstudio <- factor(bd_estudiantes$TEstudio, levels = nivel1)                # 0.5135
98  bd_estudiantes$Internet <- as.factor(bd_estudiantes$Internet)                              # 0.7703
99  bd_estudiantes$SAmigos <- factor(bd_estudiantes$SAmigos, levels = nivel2)                  # 0.3649
100 bd_estudiantes$CAlcohol <- factor(bd_estudiantes$CAlcohol, levels = nivel2)                 # 0.2973
101 bd_estudiantes$ESalud <- factor(bd_estudiantes$ESalud, levels = nivel2)                     # 0.2703
102 bd_estudiantes$TLibre <- factor(bd_estudiantes$TLibre, levels = nivel2)                     # 0.3649
103 bd_estudiantes$NAproximada <- as.factor (ifelse(bd_estudiantes$NAproximada >= 70 ,"Aprobado", "Reprobado")) # 0.5270
104
```

***Imagen 5**. Variables con su respectivo análisis de Accuracy.*

En la imagen 5, se puede observar el valor de Accuracy de cada una de las variables comprobadas con el método Confusion Matrix, el cual permitió ver la variable con mayor énfasis y así seleccionar la variable predictiva.

Por otro lado para comprobar la viabilidad del árbol de decisión se aplica la técnica Random Forest, el cual es un método que combina una cantidad grande de árboles de decisión independientes probados sobre

conjuntos de datos aleatorios (Espinosa, 2020), esta técnica me permite comprobar que los resultados sean exactos y precisos, con el fin de obtener un modelo único más ajustado en comparación con los resultados del Decision Tree,

Para la utilización de estas técnicas se emplearon las librerías “rpart( )”, “rpart.plot()” y “prune( )” del paquete “rpart” (Therneau & Beth, 2018), las cuales se encargan respectivamente de extender y podar un árbol tanto como se les indique, esto aplicada al modelo Decision Tree y para el modelo Random Forest se aplicó mediante la librería “randomForest( )” del paquete que lleva el mismo nombre (Liaw & Wiener, 2002), para llegar a una mayor precisión de los datos empleados como se muestra en la imagen 5.

Para llegar a utilizar estas 2 técnicas se utilizó datos de entrenamiento y de prueba con un porcentaje de 80 y 20, 80% para entratamiento, conocido también como Training y 20% para prueba denominada Testing, esta partición es la más precisa debido a que su sobreestimación de error de aproximación es el menor (Gholamy, Kreinovich, & Kosheleva, 2018).

D. Resultados

De acuerdo con la investigación realizada, el resultado obtenido en base a las técnicas empleadas arroja que el modelo en Decision Tree tiene una presión de 87,84% con un error de 12,16% y utilizando Random Forest arroja una precisión del 90,54% con un error de 9.46%, el cual demuestra que el modelo aplicado tiene una presión mayor al 85% en la variable “EPadres”, con una diferencia de error entre ambos modelos de 2,7%.

Finalmente se concluye que para que exista más estudiante en la carrera de Contador Público Autorizado (CPA), el estado civil de los padres de los futuros estudiantes tiene que tener un estado civil de “Unidos”, dando como resultado la aprobación del examen de ingreso a la carrera, el cual permitirá una mayor demanda y exista un porcentaje más elevado de graduados para evitar que se cierre la carrera de CPA.

E. Referencias Bibliográficas

Espinosa, Z. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Scielo*.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (Febrero de 2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Scholarworks*. Obtenido de https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*.

López, P. L. (2004). Población, muestra y muestreo. *SCIELO*. Obtenido de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1815-02762004000100012

Maranto, R. M., & González, F. M. (Febrero de 2015). Fuentes de información. *UAEH*. Obtenido de https://repository.uaeh.edu.mx/bitstream/bitstream/handle/123456789/16700/LECT132.pdf

Mendoza, V. J. (23 de Abril de 2018). *RPubs*. Obtenido de https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

Talari, S. (1 de November de 2022). *kdnugget*.

Therneau, T., & Beth, A. (2018). rpart: Recursive Partitioning and Regression Trees. *R package*.