

Índice general

1. Introducción	12
1.1. Motivación	12
1.2. Alcances y Objetivo General	13
1.3. Objetivos Específicos	13
1.4. Estructura de la memoria	13
2. Marco teórico	15
2.1. Definiciones	15
2.2. Marco conceptual	15
2.3. Formulación del problema	16
2.4. Dificultad del problema	17
2.4.1. Clases desbalanceadas	17
2.4.2. Falta de datos reales	18
2.4.3. Dinámica del fraude	18
2.5. Descripción de los orígenes de datos	18
2.5.1. Base propia	19
2.5.2. Base no propia	19
2.5.3. Base de transacciones financieras	19
2.5.4. Bases transacciones fraudulentas	19
3. Descripción de modelos	21

3.1.	Redes neuronales	21
3.1.1.	Motivación biológica	21
3.1.2.	Redes neuronales artificiales	21
3.2.	Support Vector Machines	28
3.2.1.	Introducción	28
3.2.2.	Formulación matemática	28
3.3.	Árboles de decisión	32
3.3.1.	Introducción	32
3.3.2.	C4.5	34
3.3.3.	CART	37
3.4.	Regresión logística	39
3.4.1.	Introducción	39
3.4.2.	Formulación	40
4.	Tratamiento de datos	41
4.1.	Datos de fraude	41
4.2.	Variables relevantes	45
4.2.1.	Base total	47
4.2.2.	Base telefonía	50
4.2.3.	Base transacciones financieras	53
4.2.4.	Base centros de pago	56
4.2.5.	Comparación de variables entre rubros	58
5.	Entrenamiento y test	59
5.1.	Bases de entrenamiento	59
5.2.	Entrenamiento modelos	60
5.2.1.	Entrenamiento	60
5.3.	Test	68

6. Resultados y análisis	70
6.1. Base total	72
6.1.1. Entrenamiento E50	72
6.1.2. Entrenamiento E25	73
6.1.3. Entrenamiento E10	74
6.2. Base recargas telefónicas	75
6.2.1. Entrenamiento E50	75
6.2.2. Entrenamiento E25	76
6.2.3. Entrenamiento E10	77
6.3. Base trx. financieras	78
6.3.1. Entrenamiento E50	78
6.3.2. Entrenamiento E25	79
6.3.3. Entrenamiento E10	80
6.4. Comparación modelos	81
6.5. Evaluación económica	84
6.5.1. Impacto económico en la base total	84
6.5.2. Impacto económico en la base recargas telefónicas	85
6.5.3. Impacto económico en la base trx. financieras	86
6.6. Comparación con la base total	87
6.6.1. Recargas telefónicas	88
6.6.2. Transacciones financieras	88
7. Conclusiones	90
7.1. Trabajos Futuros	92
Bibliografía	93
A. Medidas de rendimiento	I
A.1. Sensibilidad y especificidad	I

A.2. Precisión global	II
A.3. nFP	II
A.4. MCC	II
A.5. KS	II
A.6. AUC	III
B. Detección de fraude a nivel cliente	IV
B.1. Datos a nivel cliente	IV
B.2. Entrenamiento	VII
B.3. Resultados	IX
C. Curvas ROC	XIV

Índice de figuras

2.1. Esquema de los pasos que componen el proceso KDD.	16
3.1. Esquema de una ANN <i>feed-forward</i> con una capa oculta.	22
3.2. Diagrama de un nodo.	23
3.3. Tres diferentes funciones de activación para los nodos.	23
3.4. Red neuronal <i>feed-forward</i> de dos capas, con dos nodos de entrada, dos nodos ocultos y un nodo de salida.	24
3.5. Método general de aprendizaje en una ANN.	25
3.6. Algoritmo back-propagation para actualizar los pesos en una red multicapa.	27
3.7. Ejemplo simple de un árbol de decisión para conceder un crédito.	33
3.8. Árbol construido por el método dividir y conquistar.	34
3.9. Esquema de árbol para poda.	36
B.1. Curvas ROC Mes 9.	XII
B.2. Curvas ROC Mes 9, modelos con 15 variables.	XIII
C.1. Curvas ROC para los modelos entrenados en la base E50, base total.	XIV
C.2. Curvas ROC para los modelos entrenados en la base E25, base total.	XV
C.3. Curvas ROC para los modelos entrenados en la base E10, base total.	XV
C.4. Curvas ROC para los modelos entrenados en la base E50, base recargas.	XVI
C.5. Curvas ROC para los modelos entrenados en la base E25, base recargas.	XVI
C.6. Curvas ROC para los modelos entrenados en la base E10, base recargas.	XVII

C.7. Curvas ROC para los modelos entrenados en la base E50, base trx. financieras.	XVII
C.8. Curvas ROC para los modelos entrenados en la base E25, base trx. financieras.	XVIII
C.9. Curvas ROC para los modelos entrenados en la base E10, base trx. financieras.	XVIII

Índice de tablas

2.1. Matriz de confusión.	17
4.1. Clasificación n° 1 de fraude, datos de un periodo de seis meses.	42
4.2. Clasificación n° 2 de fraude, datos de un periodo de seis meses.	43
4.3. Clasificación de fraude según rubro.	44
4.4. Variables de la base consolidada.	46
4.5. Variables acumuladoras.	46
4.6. KS de variables para la base total.	48
4.7. Correlación de las variables acumuladoras de 90 días.	48
4.8. Variable Monto_Ag para el modelo de la base total.	48
4.9. Variable Cuotas_Ag para el modelo de la base total.	49
4.10. Variable ClassRubro_Ag para el modelo de la base total.	49
4.11. Variable Monto_dia_Ag para el modelo de la base total.	49
4.12. Variable Monto_30d_Ag para el modelo de la base total.	49
4.13. Variable MontoProm_90d_Ag para el modelo de la base total.	50
4.14. Variable RelMontoMax_12M_Ag para el modelo de la base total.	50
4.15. KS de variables para la base de recarga telefónicas.	51
4.16. Variable MarcaMonto para el modelo de la base total.	51
4.17. Variable Cuotas_Ag para el modelo de la base recargas telefónicas.	51
4.18. Variable Comercio para el modelo de la base recargas telefónicas.	52
4.19. Variable Monto_dia_Ag para el modelo de la base recargas telefónicas.	52

4.20. Variable MontoProm_30d_Ag para el modelo de la base recargas telefónicas.	52
4.21. Variable RelMontoProm_90d_Ag para el modelo de la base recargas telefónicas.	52
4.22. Variable RelMontoMax_12M_Ag para el modelo de la base recargas telefónicas.	52
4.23. KS de variables para la base de transacciones financieras.	54
4.24. Variable Monto_Ag para el modelo de la base trx. financieras.	54
4.25. Variable Cuotas_Ag para el modelo de la base trx. financieras.	54
4.26. Variable ClassRubro_Ag para el modelo de la base trx. financieras.	55
4.27. Variable MontoProm_dia_Ag para el modelo de la base trx. financieras.	55
4.28. Variable Monto_30d_Ag para el modelo de la base trx. financieras.	55
4.29. Variable Monto_90d_Ag para el modelo de la base trx. financieras.	55
4.30. Variable RelMontoMax_12M_Ag para el modelo de la base trx. financieras.	55
4.31. KS de variables para la base centros de pago.	57
4.32. Variable Cuotas_Ag para el modelo de la base centros de pago.	57
4.33. Variable Cuotas en el comercio F.	57
5.1. Cantidad de transacciones normales y fraudulentas en las bases de entrenamiento.	60
5.2. Tiempos de entrenamiento en segundos, base total.	62
5.3. Parámetros escogidos para cada modelo según su rendimiento, base total.	63
5.4. Errores de clasificación (%) en el entrenamiento, base total.	63
5.5. Tiempos de entrenamiento en segundos, base recargas.	64
5.6. Parámetros escogidos para cada modelo según su rendimiento, base recargas.	65
5.7. Errores de clasificación (%) en el entrenamiento, base recargas.	65
5.8. Tiempos de entrenamiento en segundos, base trx financieras.	66
5.9. Parámetros escogidos para cada modelo según su rendimiento, base trx financieras.	67
5.10. Errores de clasificación (%) en el entrenamiento, base trx financieras.	67
5.11. Tiempos en segundos de: (i) promedio para el entrenamiento y (ii) $\text{Test}\alpha$, base total.	68
5.12. Tiempos en segundos de: (i) promedio para el entrenamiento y (ii) $\text{Test}\beta$, base recargas telefónicas.	69

5.13. Tiempos en segundos de: (i) promedio para el entrenamiento y (ii) Test γ , base trx. financieras.	69
6.1. (i) Matriz de confusión, (ii) Matriz de confusión porcentual.	70
6.2. Matrices de confusión porcentual para los modelos entrenados en la base E50, base total.	72
6.3. Medidas para los modelos entrenados en la base E50, base total.	72
6.4. Porcentajes de casos fraudulentos para las más altas probabilidades, E50 base total.	72
6.5. Matrices de confusión porcentual para los modelos entrenados en la base E25, base total.	73
6.6. Medidas para los modelos entrenados en la base E25, base total.	73
6.7. Porcentajes de casos fraudulentos para las más altas probabilidades, E25 base total.	74
6.8. Matrices de confusión porcentual para los modelos entrenados en la base E10, base total.	74
6.9. Medidas para los modelos entrenados en la base E10, base total.	75
6.10. Porcentajes de casos fraudulentos para las más altas probabilidades, E10 base total.	75
6.11. Matrices de confusión porcentual para los modelos entrenados en la base E50, base recargas.	75
6.12. Medidas para los modelos entrenados en la base E50, base recargas.	76
6.13. Porcentajes de casos fraudulentos para las más altas probabilidades, E50 base recargas.	76
6.14. Matrices de confusión porcentual para los modelos entrenados en la base E25, base recargas.	76
6.15. Medidas para los modelos entrenados en la base E25, base recargas.	77
6.16. Porcentajes de casos fraudulentos para las más altas probabilidades, E25 base recargas.	77
6.17. Matrices de confusión porcentual para los modelos entrenados en la base E10, base recargas.	77
6.18. Medidas para los modelos entrenados en la base E10, base recargas.	78
6.19. Porcentajes de casos fraudulentos para las más altas probabilidades, E10 base recargas.	78

6.20. Matrices de confusión porcentual para los modelos entrenados en la base E50, base trx. financieras.	78
6.21. Medidas para los modelos entrenados en la base E50, base trx. financieras.	79
6.22. Porcentajes de casos fraudulentos para las más altas probabilidades, E50 base trx. financieras.	79
6.23. Matrices de confusión porcentual para los modelos entrenados en la base E25, base trx. financieras.	79
6.24. Medidas para los modelos entrenados en la base E25, base trx. financieras.	80
6.25. Porcentajes de casos fraudulentos para las más altas probabilidades, E25 base trx. financieras.	80
6.26. Matrices de confusión porcentual para los modelos entrenados en la base E10, base trx. financieras.	80
6.27. Medidas para los modelos entrenados en la base E10, base trx. financieras.	81
6.28. Porcentajes de casos fraudulentos para las más altas probabilidades, E10 base trx. financieras.	81
6.29. Promedio de las medidas de los modelos, en las bases total, recargas y trx. financieras.	82
6.30. Diferencia relativa entre el mejor y peor modelo según cada indicador, base total. . .	83
6.31. Diferencia relativa entre el mejor y peor modelo según cada indicador, base recargas.	83
6.32. Diferencia relativa entre el mejor y peor modelo según cada indicador, base trx. financieras.	83
6.33. IE en millones de pesos para los modelos de la base total.	85
6.34. IE en miles de pesos para los modelos de la base recargas telefónicas.	86
6.35. IE en millones de pesos para los modelos de la base trx. financieras.	87
6.36. Comparación modelos de las bases total aplicado a recargas telefónicas y recargas telefónicas.	88
6.37. Comparación modelos de las bases total aplicado a trx.financieras y trx. financieras.	89
B.1. Estructura base de fraude a nivel cliente.	IV
B.2. Número de clientes con fraude en el periodo considerado. (*): N° cuenta único. . . .	V
B.3. Variables a considerar en los modelos.	VI
B.4. Agrupación de variable “Sop_compras” en tres categorías.	VII

B.5. KS de las variables que entran en los modelos.	VIII
B.6. Precisión global en el entrenamiento de los modelos.	IX
B.7. Precisión global modelos.	X
B.8. KS modelos.	X
B.9. Variables definitivas a considerar en los modelos y n° de categorías.	XI
B.10. Precisión global logísticas.	XI