Trabajo Final Integrador

Especialización en Estadística – FCEN UBA

Cecilia Oliva (ceciliamoliva@hotmail.com)

12/02/2020

Introducción

El presente trabajo tiene como objetivo general analizar la base de datos "Communities and Crime", extraída del repositorio *UCI Machine Learning Repository* (http://archive.ics.uci.edu/ml/datasets/communities+and+crime), y buscar el modelo que mejor prediga la cantidad de crímenes violentos por cada cien mil habitantes en las comunidades de Estados Unidos que cuentan con la mayor información necesaria disponible, dentro del marco de algunos modelos estudiados en la Especialización en Estadística de FCEN, UBA. Dicha base combina datos socioeconómicos de las comunidades de Estados Unidos del censo de 1990 y datos de delitos del programa UCR (Uniform Crime Reporting) del FBI de 1995. Además contiene datos de la aplicación de la ley y estadísticas administrativas de LEMAS (Law Enforcement Management And Administrative Statistics) de 1990. LEMAS recopila datos de miles agencias de aplicación de la ley de propósito general, del condado y locales, incluidos todos aquellos que emplean a 100 o más oficiales a tiempo completo y una muestra representativa a nivel nacional de agencias más pequeñas. Entre los datos que se obtienen se encuentran: las políticas de armamento y armaduras, los requisitos de educación y capacitación, los sistemas de información y computación, los vehículos, las unidades especiales y actividades de vigilancia comunitaria.

Descripción general de los datos

El conjunto de datos incluye 122 covariables que, según la fuente, se seleccionaron debido a la existencia de alguna conexión plausible con el crimen. Las mismas serán detalladas en el apartado siguiente. Por otro lado, la variable objetivo "Crímenes violentos per cápita" es la variable a predecir. Además hay 5 variables nominales que especifican a qué comunidad/observación se hace referencia. En total se contabilizan 1994 observaciones, cada una referente a una comunidad de Estados Unidos, y para cada una de las cuales se registran las 128 variables mencionadas anteriormente.

Algunas variables involucran a la comunidad, como el porcentaje de la población considerada urbana y el ingreso familiar promedio, otras involucran a las autoridades policiales, como el número per cápita de oficiales de policía y el porcentaje de oficiales asignados a unidades de drogas. De acuerdo a lo explicado en la página web que brinda los datos, la variable de delitos violentos per cápita se calculó utilizando la población y la suma de las variables de delitos consideradas delitos violentos en los Estados Unidos: asesinato, violación, robo y asalto. Se destaca que ciertas comunidades del medio oeste de Estados Unidos no fueron incluidas en el conteo de violaciones debido a valores faltantes en la recolección de la información. Por otro lado, una limitación es que la encuesta de LEMAS se realiza en los departamentos de policía con al menos 100 oficiales, más una muestra aleatoria de departamentos más pequeños. Es por este motivo que se omitieron en la base las comunidades que no se encuentran en los conjuntos de datos de censos y delitos. De todos modos a muchas comunidades presentes en el conjunto de datos les faltan datos de LEMAS, por lo que en este trabajo serán descartadas.

Descripción detallada de los datos

En este caso los datos que brinda el repositorio UCI están en función de los valores originales. Todos los datos numéricos se normalizaron en el rango decimal 0.00-1.00 utilizando un método de agrupación. Los atributos conservan su distribución y sesgo (por lo tanto, por ejemplo, el atributo de población tiene un valor medio de 0.06 porque la mayoría de las comunidades son pequeñas). Un atributo descripto como 'personas promedio por hogar' es en realidad la versión normalizada (0-1) de ese valor.

La normalización conserva proporciones aproximadas de valores DENTRO de un atributo (por ejemplo, duplicar un valor de la variable población es duplicar la población dentro de la precisión disponible, excepto los valores extremos (todos los valores superiores a 3 veces el desvío estándar por encima de la media se normalizan a 1,00, y todos los valores a más de 3 veces el desvío estándar por debajo de la media se normaliza a 0,00)).

Sin embargo, la normalización no preserva las relaciones ENTRE los valores de dos atributos (por ejemplo, no sería significativo comparar el valor de whitePerCap con el valor de blackPerCap para una comunidad).

A continuación se exponen los nombres de las variables, su correspondiente significado, el tipo de dato y si contiene datos faltantes. Las primeras cinco variables identifican a las observaciones, no son utilizadas como atributos o covariables de predicción. La última variable es la variable a predecir.

- state: estado de EE. UU. (por número) - no se cuenta como predictivo.
- county: código numérico para el condado - no predictivo, y muchos valores faltantes (numérico)
- community: código numérico de comunidad - no predictivo y muchos valores perdidos (numérico)
- communityname: nombre de comunidad - no predictivo - solo para información (cadena)
- fold: número de fold para la validación cruzada no aleatoria de 10 fold, potencialmente útil para depuración, pruebas pareadas - no predictivo (numérico)
- population: población por comunidad: (numérico - decimal)
- householdsize: personas promedio por hogar (numérico - decimal)
- racepctblack: porcentaje de la población que es afroamericana (numérico - decimal)
- racePctWhite: porcentaje de población que es caucásico (numérico - decimal)
- racePctAsian: porcentaje de población que es de herencia asiática (numérico - decimal)
- racePctHisp: porcentaje de población que es de herencia hispana (numérico - decimal)
- agePct12t21: porcentaje de población que tiene entre 12 y 21 años de edad (numérico - decimal)
- agePct12t29: porcentaje de población que tiene entre 12 y 29 años de edad (numérico - decimal)
- agePct16t24: porcentaje de población que tiene entre 16 y 24 años de edad (numérico - decimal)
- agePct65up: porcentaje de la población que tiene 65 años o más (numérico - decimal)
- NumbUrban: número de personas que viven en áreas clasificadas como urbanas (numérico - decimal)
- pctUrban: porcentaje de personas que viven en áreas clasificadas como urbanas (numérico - decimal)
- medIncome: ingreso medio del hogar (numérico - decimal)
- pctWWage: porcentaje de hogares con ingresos salariales o salariales en 1989 (numérico - decimal)

- pctWFarmSelf: porcentaje de hogares con ingresos agrícolas o por cuenta propia en 1989 (numérico - decimal)
- pctWInvInc: porcentaje de hogares con ingresos por inversiones / rentas en 1989 (numérico - decimal)
- pctWSocSec: porcentaje de hogares con ingresos de seguridad social en 1989 (numérico - decimal)
- pctWPubAsst: porcentaje de hogares con ingresos de asistencia pública en 1989 (numérico - decimal)
- pctWRetire: porcentaje de hogares con ingresos de jubilación en 1989 (numérico - decimal)
- medFamInc: ingreso familiar medio (difiere del ingreso familiar para hogares no familiares) (numérico - decimal)
- perCapInc: ingreso per cápita (numérico - decimal)
- whitePerCap: ingreso per cápita para caucásicos (numérico - decimal)
- blackPerCap: ingreso per cápita para los afroamericanos (numérico - decimal)
- indianPerCap: ingreso per cápita para los nativos americanos (numérico - decimal)
- AsianPerCap: ingreso per cápita para personas con herencia asiática (numérico - decimal)
- OtherPerCap: ingreso per cápita para personas con 'otro' patrimonio (numérico - decimal)
- HispPerCap: ingreso per cápita para personas con herencia hispana (numérico - decimal)
- NumUnderPov: número de personas bajo el nivel de pobreza (numérico - decimal)
- PctPopUnderPov: porcentaje de personas bajo el nivel de pobreza (numérico - decimal)
- PctLess9thGrade: porcentaje de personas de 25 años o más con educación inferior a noveno grado (numérico - decimal)
- PctNotHSGrad: porcentaje de personas de 25 años o más que no son graduados de la escuela secundaria (numérico - decimal)
- PctBSorMore: porcentaje de personas de 25 años o más con una licenciatura o educación superior (numérico - decimal)
- PctUnemployed: porcentaje de personas de 16 años o más, en la fuerza laboral y desempleados (numérico - decimal)
- PctEmploy: porcentaje de personas de 16 años o más que están empleadas (numérico - decimal)
- PctEmplManu: porcentaje de personas de 16 años o más que están empleadas en la fabricación (numérico - decimal)
- PctEmplProfServ: porcentaje de personas de 16 años o más que trabajan en servicios profesionales (numérico - decimal)
- PctOccupManu: porcentaje de personas de 16 años o más que están empleadas en la fabricación (numérico - decimal)
- PctOccupMgmtProf: porcentaje de personas de 16 años o más que están empleadas en ocupaciones administrativas o profesionales (numérico - decimal)
- MalePctDivorce: porcentaje de hombres divorciados (numérico - decimal)
- MalePctNevMarr: porcentaje de hombres que nunca se han casado (numérico - decimal)
- FemalePctDiv: porcentaje de mujeres divorciadas (numérico - decimal)
- TotalPctDiv: porcentaje de la población divorciada (numérico - decimal)
- PersPerFam: número promedio de personas por familia (numérico - decimal)
- PctFam2Par: porcentaje de familias (con niños) que están encabezadas por dos padres (numérico - decimal)
- PctKids2Par: porcentaje de niños en viviendas familiares con dos padres (numérico - decimal)
- PctYoungKids2Par: porcentaje de niños de 4 años y menores en dos hogares de padres (numérico - decimal)
- PctTeen2Par: porcentaje de niños de 12 a 17 años en dos hogares de padres (numérico - decimal)
- PctWorkMomYoungKids: porcentaje de madres de niños menores de 6 años en la fuerza laboral (numérico - decimal)
- PctWorkMom: porcentaje de madres de niños menores de 18 años en la fuerza laboral (numérico - decimal)
- NumIlleg: número de hijos nacidos o nunca casados (numérico - decimal)
- PctIlleg: porcentaje de niños nacidos de padres nunca casados (numérico – decimal)
- NumImmig: número total de personas nacidas en el extranjero (numérico - decimal)
- PctImmigRecent: porcentaje de _inmigrantes_ que inmigran en los últimos 3 años (numérico - decimal)

- PctImmigRec5: porcentaje de _inmigrantes_ que inmigran en los últimos 5 años (numérico - decimal)
- PctImmigRec8: porcentaje de _inmigrantes_ que inmigran en los últimos 8 años (numérico - decimal)
- PctImmigRec10: porcentaje de _inmigrantes_ que inmigran en los últimos 10 años (numérico - decimal)
- PctRecentImmig: porcentaje de _población_ que han inmigrado en los últimos 3 años (numérico - decimal)
- PctRecImmig5: porcentaje de _población_ que han inmigrado en los últimos 5 años (numérico - decimal)
- PctRecImmig8: porcentaje de _población_ que han inmigrado en los últimos 8 años (numérico - decimal)
- PctRecImmig10: porcentaje de _población_ que han emigrado en los últimos 10 años (numérico - decimal)
- PctSpeakEnglOnly: porcentaje de personas que solo hablan inglés (numérico - decimal)
- PctNotSpeakEnglWell: porcentaje de personas que no hablan bien el inglés (numérico - decimal)
- PctLargHouseFam: porcentaje de hogares familiares que son grandes (6 o más) (numérico - decimal)
- PctLargHouseOccup: porcentaje de todos los hogares ocupados que son grandes (6 o más personas) (numérico - decimal)
- PersPerOccupHous: personas promedio por hogar (numérico - decimal)
- PersPerOwnOccHous: personas promedio por hogar ocupado por el propietario (numérico - decimal)
- PersPerRentOccHous: personas promedio por hogar de alquiler (numérico - decimal)
- PctPersOwnOccup: porcentaje de personas en hogares ocupados por sus propietarios (numérico - decimal)
- PctPersDenseHous: porcentaje de personas en viviendas densas (más de 1 persona por habitación) (numérico - decimal)
- PctHousLess3BR: porcentaje de unidades de vivienda con menos de 3 dormitorios (numérico - decimal)
- MedNumBR: número medio de habitaciones (numérico - decimal)
- HousVacant: número de hogares vacantes (numérico - decimal)
- PctHousOccup: porcentaje de viviendas ocupadas (numérico - decimal)
- PctHousOwnOcc: porcentaje de hogares ocupados por el propietario (numérico - decimal)
- PctVacantBoarded: porcentaje de viviendas vacantes que se ha cerrado (numérico - decimal)
- PctVacMore6Mos: porcentaje de viviendas vacantes que han estado vacantes por más de 6 meses (numérico - decimal)
- MedYrHousBuilt: unidades de vivienda de año medio construidas (numérico - decimal)
- PctHousNoPhone: porcentaje de unidades de vivienda ocupadas sin teléfono (en 1990, ¡esto era raro!) (Numérico - decimal)
- PctWOFullPlumb: porcentaje de viviendas sin instalaciones de plomería completas (numérico - decimal)
- OwnOccLowQuart: vivienda ocupada por el propietario - valor de cuartil inferior (numérico - decimal)
- OwnOccMedVal: vivienda ocupada por el propietario - valor medio (numérico - decimal)
- OwnOccHiQuart: vivienda ocupada por el propietario - valor del cuartil superior (numérico - decimal)
- RentLowQ: viviendas de alquiler - alquiler por cuartil inferior (numérico - decimal)
- RentMedian: vivienda de alquiler - renta mediana (variable del Censo H32B del archivo STF1A) (numérico - decimal)
- RentHighQ: viviendas de alquiler - alquiler en el cuartil superior (numérico - decimal)
- MedRent: renta bruta mediana (variable de censo H43A del archivo STF3A - incluye utilidades) (numérico - decimal)
- MedRentPctHousInc: renta bruta mediana como porcentaje del ingreso familiar (numérico - decimal)
- MedOwnCostPctInc: los propietarios medianos cuestan como porcentaje del ingreso familiar - para los propietarios con una hipoteca (numérico - decimal)
- MedOwnCostPctIncNoMtg: el costo promedio de los propietarios como porcentaje del ingreso familiar - para los propietarios sin una hipoteca (numérico - decimal)
- NumInShelters: número de personas en refugios para personas sin hogar (numérico - decimal)
- NumStreet: número de personas sin hogar contadas en la calle (numérico - decimal)
- PctForeignBorn: porcentaje de personas nacidas en el extranjero (numérico - decimal)
- PctBornSameState: porcentaje de personas nacidas en el mismo estado que viven actualmente (numérico - decimal)
- PctSameHouse85: porcentaje de personas que viven en la misma casa que en 1985 (5 años antes) (numérico - decimal)
- PctSameCity85: porcentaje de personas que viven en la misma ciudad que en 1985 (5 años antes) (numérico - decimal)

- PctSameState85: porcentaje de personas que viven en el mismo estado que en 1985 (5 años antes) (numérico - decimal)
- LemasSwornFT: número de oficiales de policía jurados a tiempo completo, y muchos valores faltantes (numérico - decimal)
- LemasSwFTPerPop: oficiales de policía jurados a tiempo completo por población de 100K, y muchos valores faltantes (numérico - decimal)
- LemasSwFTFieldOps: número de oficiales de policía jurados a tiempo completo en operaciones de campo (en la calle en lugar de administrativos, etc.), y muchos valores faltantes (numérico - decimal)
- LemasSwFTFieldPerPop: oficiales de policía a tiempo completo jurados en operaciones de campo (en la calle en lugar de administrativos, etc.) por cada 100 K de población, y muchos valores faltantes (numérico - decimal)
- LemasTotalReq: solicitudes totales de policía, y muchos valores faltantes (numérico - decimal)
- LemasTotReqPerPop: solicitudes totales de policías por cada 100 000 habitantes, y muchos valores faltantes (numérico - decimal)
- PolicReqPerOffic: solicitudes totales de policía por oficial de policía, y muchos valores faltantes (numérico - decimal)
- PolicPerPop: agentes de policía por cada población de 100K, y muchos valores faltantes (numérico - decimal)
- RacialMatchCommPol: una medida de la coincidencia racial entre la comunidad y la fuerza policial. Los valores altos indican que las proporciones en la comunidad y la fuerza policial son similares, y muchos valores faltantes (numérico - decimal)
- PctPolicWhite: porcentaje de policías que son caucásicos, y muchos valores faltantes (numérico - decimal)
- PctPolicBlack: porcentaje de policías que son afroamericanos, y muchos valores faltantes (numérico - decimal)
- PctPolicHisp: porcentaje de policías que son hispanos, y muchos valores faltantes (numérico - decimal)
- PctPolicAsian: porcentaje de policías asiáticos, y muchos valores faltantes (numérico - decimal)
- PctPolicMinor: porcentaje de policías que son minoritarios de cualquier tipo, y muchos valores faltantes (numérico - decimal)
- OfficAssgnDrugUnits: número de oficiales asignados a unidades especiales de drogas, y muchos valores faltantes (numérico - decimal)
- NumKindsDrugsSeiz: número de diferentes tipos de drogas incautadas, y muchos valores faltantes (numérico - decimal)
- PolicAveOTWorked: promedio de horas extra trabajadas por la policía, y muchos valores faltantes (numérico - decimal)
- LandArea: área de tierra en millas cuadradas (numérico - decimal)
- PopDens: densidad de población en personas por milla cuadrada (numérico - decimal)
- PctUsePubTrans: porcentaje de personas que utilizan el transporte público para los desplazamientos (numérico - decimal)
- PolicCars: número de coches de policía, y muchos valores faltantes (numérico - decimal)
- PolicOperBudg: presupuesto operativo de la policía, y muchos valores faltantes (numérico - decimal)
- LemasPctPolicOnPatr: porcentaje de oficiales de policía jurados a tiempo completo en patrulla, y muchos valores faltantes (numérico - decimal)
- LemasGangUnitDeploy: unidad de grupo desplegada, y muchos valores faltantes (numérico - decimal - pero realmente ordinal - 0 significa NO, 1 significa SÍ, 0.5 significa Tiempo parcial)
- LemasPctOfficDrugUn: porcentaje de oficiales asignados a unidades de drogas (numérico - decimal)
- PolicBudgPerPop: presupuesto operativo de la policía por población, y muchos valores faltantes (numérico - decimal)
- ViolentCrimesPerPop: número total de delitos violentos por cada 100K población (numérico - decimal) atributo OBJETIVO (variable a predecir)

Para conocer parcialmente el comportamiento de las variables se presentan abajo las medidas resumen y la cantidad de datos faltantes de cada una.

	Min	Max	Mean	SD	Correl	Median	Mode	Missing
population	0	1	0.06	0.13	0.37	0.02	0.01	0
householdsize	0	1	0.46	0.16	-0.03	0.44	0.41	0
racepctblack	0	1	0.18	0.25	0.63	0.06	0.01	0
racePctWhite	0	1	0.75	0.24	-0.68	0.85	0.98	0
racePctAsian	0	1	0.15	0.21	0.04	0.07	0.02	0
racePctHisp	0	1	0.14	0.23	0.29	0.04	0.01	0
agePct12t21	0	1	0.42	0.16	0.06	0.4	0.38	0
agePct12t29	0	1	0.49	0.14	0.15	0.48	0.49	0
agePct16t24	0	1	0.34	0.17	0.10	0.29	0.29	0
agePct65up	0	1	0.42	0.18	0.07	0.42	0.47	0
numbUrban	0	1	0.06	0.13	0.36	0.03	0	0
pctUrban	0	1	0.70	0.44	0.08	1	1	0
medIncome	0	1	0.36	0.21	-0.42	0.32	0.23	0
pctWWage	0	1	0.56	0.18	-0.31	0.56	0.58	0
pctWFarmSelf	0	1	0.29	0.20	-0.15	0.23	0.16	0
pctWInvInc	0	1	0.50	0.18	-0.58	0.48	0.41	0
pctWSocSec	0	1	0.47	0.17	0.12	0.475	0.56	0
pctWPubAsst	0	1	0.32	0.22	0.57	0.26	0.1	0
pctWRetire	0	1	0.48	0.17	-0.10	0.47	0.44	0
medFamInc	0	1	0.38	0.20	-0.44	0.33	0.25	0
perCapInc	0	1	0.35	0.19	-0.35	0.3	0.23	0
whitePerCap	0	1	0.37	0.19	-0.21	0.32	0.3	0
blackPerCap	0	1	0.29	0.17	-0.28	0.25	0.18	0
indianPerCap	0	1	0.20	0.16	-0.09	0.17	0	0
AsianPerCap	0	1	0.32	0.20	-0.16	0.28	0.18	0
OtherPerCap	0	1	0.28	0.19	-0.13	0.25	0	1
HispPerCap	0	1	0.39	0.18	-0.24	0.345	0.3	0
NumUnderPov	0	1	0.06	0.13	0.45	0.02	0.01	0
PctPopUnderPov	0	1	0.30	0.23	0.52	0.25	0.08	0
PctLess9thGrade	0	1	0.32	0.21	0.41	0.27	0.19	0
PctNotHSGrad	0	1	0.38	0.20	0.48	0.36	0.39	0
PctBSorMore	0	1	0.36	0.21	-0.31	0.31	0.18	0
PctUnemployed	0	1	0.36	0.20	0.50	0.32	0.24	0
PctEmploy	0	1	0.50	0.17	-0.33	0.51	0.56	0
PctEmplManu	0	1	0.40	0.20	-0.04	0.37	0.26	0
PctEmplProfServ	0	1	0.44	0.18	-0.07	0.41	0.36	0
PctOccupManu	0	1	0.39	0.20	0.30	0.37	0.32	0
PctOccupMgmtProf	0	1	0.44	0.19	-0.34	0.4	0.36	0
MalePctDivorce	0	1	0.46	0.18	0.53	0.47	0.56	0
MalePctNevMarr	0	1	0.43	0.18	0.30	0.4	0.38	0
FemalePctDiv	0	1	0.49	0.18	0.56	0.5	0.54	0
TotalPctDiv	0	1	0.49	0.18	0.55	0.5	0.57	0
PersPerFam	0	1	0.49	0.15	0.14	0.47	0.44	0
PctFam2Par	0	1	0.61	0.20	-0.71	0.63	0.7	0
PctKids2Par	0	1	0.62	0.21	-0.74	0.64	0.72	0
PctYoungKids2Par	0	1	0.66	0.22	-0.67	0.7	0.91	0
PctTeen2Par	0	1	0.58	0.19	-0.66	0.61	0.6	0
PctWorkMomYoungKids	0	1	0.50	0.17	-0.02	0.51	0.51	0
PctWorkMom	0	1	0.53	0.18	-0.15	0.54	0.57	0
NumIlleg	0	1	0.04	0.11	0.47	0.01	0	0

```
PctIlleg 0 1 0.25 0.23 0.74 0.17 0.09 0
NumImmig 0 1 0.03 0.09 0.29 0.01 0 0
PctImmigRecent 0 1 0.32 0.22 0.17 0.29 0 0
PctImmigRec5 0 1 0.36 0.21 0.22 0.34 0 0
PctImmigRec8 0 1 0.40 0.20 0.25 0.39 0.26 0
PctImmigRec10 0 1 0.43 0.19 0.29 0.43 0.43 0
PctRecentImmig 0 1 0.18 0.24 0.23 0.09 0.01 0
PctRecImmig5 0 1 0.18 0.24 0.25 0.08 0.02 0
PctRecImmig8 0 1 0.18 0.24 0.25 0.09 0.02 0
PctRecImmig10 0 1 0.18 0.23 0.26 0.09 0.02 0
PctSpeakEnglOnly 0 1 0.79 0.23 -0.24 0.87 0.96 0
PctNotSpeakEnglWell 0 1 0.15 0.22 0.30 0.06 0.03 0
PctLargHouseFam 0 1 0.27 0.20 0.38 0.2 0.17 0
PctLargHouseOccup 0 1 0.25 0.19 0.29 0.19 0.19 0
PersPerOccupHous 0 1 0.46 0.17 -0.04 0.44 0.37 0
PersPerOwnOccHous 0 1 0.49 0.16 -0.12 0.48 0.45 0
PersPerRentOccHous 0 1 0.40 0.19 0.25 0.36 0.32 0
PctPersOwnOccup 0 1 0.56 0.20 -0.53 0.56 0.54 0
PctPersDenseHous 0 1 0.19 0.21 0.45 0.11 0.06 0
PctHousLess3BR 0 1 0.50 0.17 0.47 0.51 0.53 0
MedNumBR 0 1 0.31 0.26 -0.36 0.5 0.5 0
HousVacant 0 1 0.08 0.15 0.42 0.03 0.01 0
PctHousOccup 0 1 0.72 0.19 -0.32 0.77 0.88 0
PctHousOwnOcc 0 1 0.55 0.19 -0.47 0.54 0.52 0
PctVacantBoarded 0 1 0.20 0.22 0.48 0.13 0 0
PctVacMore6Mos 0 1 0.43 0.19 0.02 0.42 0.44 0
MedYrHousBuilt 0 1 0.49 0.23 -0.11 0.52 0 0
PctHousNoPhone 0 1 0.26 0.24 0.49 0.185 0.01 0
PctWOFullPlumb 0 1 0.24 0.21 0.36 0.19 0 0
OwnOccLowQuart 0 1 0.26 0.22 -0.21 0.18 0.09 0
OwnOccMedVal 0 1 0.26 0.23 -0.19 0.17 0.08 0
OwnOccHiQuart 0 1 0.27 0.24 -0.17 0.18 0.08 0
RentLowQ 0 1 0.35 0.22 -0.25 0.31 0.13 0
RentMedian 0 1 0.37 0.21 -0.24 0.33 0.19 0
RentHighQ 0 1 0.42 0.25 -0.23 0.37 1 0
MedRent 0 1 0.38 0.21 -0.24 0.34 0.17 0
MedRentPctHousInc 0 1 0.49 0.17 0.33 0.48 0.4 0
MedOwnCostPctInc 0 1 0.45 0.19 0.06 0.45 0.41 0
MedOwnCostPctIncNoMtg 0 1 0.40 0.19 0.05 0.37 0.24 0
NumInShelters 0 1 0.03 0.10 0.38 0 0 0
NumStreet 0 1 0.02 0.10 0.34 0 0 0
PctForeignBorn 0 1 0.22 0.23 0.19 0.13 0.03 0
PctBornSameState 0 1 0.61 0.20 -0.08 0.63 0.78 0
PctSameHouse85 0 1 0.54 0.18 -0.16 0.54 0.59 0
PctSameCity85 0 1 0.63 0.20 0.08 0.67 0.74 0
PctSameState85 0 1 0.65 0.20 -0.02 0.7 0.79 0
LemasSwornFT 0 1 0.07 0.14 0.34 0.02 0.02 1675
LemasSwFTPerPop 0 1 0.22 0.16 0.15 0.18 0.2 1675
LemasSwFTFieldOps 0 1 0.92 0.13 -0.33 0.97 0.98 1675
LemasSwFTFieldPerPop 0 1 0.25 0.16 0.16 0.21 0.19 1675
LemasTotalReq 0 1 0.10 0.16 0.35 0.04 0.02 1675
LemasTotReqPerPop 0 1 0.22 0.16 0.27 0.17 0.14 1675
PolicReqPerOffic 0 1 0.34 0.20 0.17 0.29 0.23 1675
PolicPerPop 0 1 0.22 0.16 0.15 0.18 0.2 1675
RacialMatchCommPol 0 1 0.69 0.23 -0.46 0.74 0.78 1675
PctPolicWhite 0 1 0.73 0.22 -0.44 0.78 0.72 1675
PctPolicBlack 0 1 0.22 0.24 0.54 0.12 0 1675
```

```
PctPolicHisp 0 1 0.13 0.20 0.12 0.06 0 1675
PctPolicAsian 0 1 0.11 0.23 0.10 0 0 1675
PctPolicMinor 0 1 0.26 0.23 0.49 0.2 0.07 1675
OfficAssgnDrugUnits 0 1 0.08 0.12 0.34 0.04 0.03 1675
NumKindsDrugsSeiz 0 1 0.56 0.20 0.13 0.57 0.57 1675
PolicAveOTWorked 0 1 0.31 0.23 0.03 0.26 0.19 1675
LandArea 0 1 0.07 0.11 0.20 0.04 0.01 0
PopDens 0 1 0.23 0.20 0.28 0.17 0.09 0
PctUsePubTrans 0 1 0.16 0.23 0.15 0.07 0.01 0
PolicCars 0 1 0.16 0.21 0.38 0.08 0.02 1675
PolicOperBudg 0 1 0.08 0.14 0.34 0.03 0.02 1675
LemasPctPolicOnPatr 0 1 0.70 0.21 -0.08 0.75 0.74 1675
LemasGangUnitDeploy 0 1 0.44 0.41 0.12 0.5 0 1675
LemasPctOfficDrugUn 0 1 0.09 0.24 0.35 0 0 0
PolicBudgPerPop 0 1 0.20 0.16 0.10 0.15 0.12 1675
ViolentCrimesPerPop 0 1 0.24 0.23 1.00 0.15 0.03 0
```

Se puede observar que hay 22 variables con alta cantidad de datos ausentes (1675 datos faltantes de un total de 1994 observaciones), las cuales serán omitidas para el posterior análisis estadístico.

Objetivo del trabajo

El principal objetivo de este trabajo es seleccionar el modelo con mayor capacidad de generalización y predicción de los delitos violentos por cada 100000 habitantes, utilizando las covariables sin datos faltantes, dentro del marco de las herramientas estadísticas y modelos estudiados en la Especialización en Estadística de FCEN, UBA. Para esto, en primer lugar se utilizarán métodos de selección de variables y de regularización dada la gran cantidad de covariables para disminuir la alta dimensionalidad del problema. Luego se aplicarán distintos modelos de regresión desde los clásicos modelos lineales, pasando por modelos aditivos generalizados, hasta modelos de redes neuronales.

La idea es realizar el entrenamiento de los modelos con el 75% de los datos (1495 observaciones elegidas aleatoriamente), y dejar el 25% restante (499 observaciones) para testeo. Así se podrán calcular tanto los errores cuadráticos medios de entrenamiento como los de test a partir de los cuales se comparará el comportamiento del ajuste en cada caso a fin de decidir el modelo que mejor ajusta los datos y/o que tiene mayor capacidad de predicción.

Selección de variables

Antes de emplear los métodos de selección de variables omitimos de la base las 22 covariables con 1675 datos faltantes, ya que representa una proporción superior al 84% de la totalidad de observaciones. De esta manera, quedan disponibles entonces 100 covaria-

bles, de todas las cuales se quiere obtener un subconjunto de aquellas relevantes para considerar en los modelos de regresión. En el proceso de selección de variables además se debe tener presente que sólo se utilizará la muestra de entrenamiento.

Aplicando a los datos de entrenamiento el método Forward con la función "regsubsets" del paquete "leaps" de R las variables seleccionadas son:

- racePctWhite,
- PctPopUnderPov
- PctKids2Par,
- PctWorkMom,
- PctPersDenseHous,
- HousVacant,
- PctVacantBoarded,
- NumStreet.

Por otro lado, utilizando el método Backward con el mismo paquete de R, las variables seleccionadas resultan:

- racepctblack,
- pctUrban ,
- PctKids2Par*,
- PctIlleg,
- PctPersDenseHous*,
- PctHousOccup ,
- PctVacantBoarded*,
- NumStreet*.

Se puede notar que ambos modelos seleccionan 8 variables cada uno, donde las únicas que se repiten son PctKids2Par, PctPersDenseHous, PctVacantBoarded y NumStreet, que fueron marcadas con un asterisco en el último listado.

Por último, bajo el método mixto, que es una combinación de los dos métodos anteriores, se obtienen:

- racePctWhite*,
- pctUrban*,
- MalePctDivorce,
- PctKids2Par**,
- PctWorkMom*,
- PctIlleg*,
- PctHousOccup*,
- NumStreet**.

Se observa que las variables racePctWhite, pctUrban, PctWorkMom, PctIlleg y PctHousOccup, marcadas con un asterisco, fueron también elegidas por el método de forward o backward, mientras que las que figuran acompañadas de doble asterisco, PctKids2Par y NumStreet, fueron seleccionadas por los tres métodos.

En total, se contabilizan 13 variables, de las cuales 2 son seleccionadas por los tres métodos (PctKids2Par y NumStreet), 7 son seleccionadas por dos de ellos (racePctWhite, pctUrban, PctWorkMom, PctIlleg, PctHousOccup, PctVacantBoarded, PctPersDenseHous), y el resto son elegidas por un sólo método (HousVacant, PctPopUnderPov, racepctblack, MalePctDivorce).

Regularización

Como no se puede asegurar que no existan problemas de colinealidad conviene analizar los datos de entrenamiento con métodos de regularización. Los métodos conocidos son los de Ridge y Lasso. En este caso las covariables ya se encuentran estandarizadas originalmente por lo que se pueden aplicar directamente ambos métodos.

La principal diferencia práctica entre Lasso y Ridge es que el primero consigue que algunos coeficientes sean exactamente cero, por lo que realiza selección de predictores, mientras que el segundo no llega a excluir ninguno. Esto supone una ventaja notable de Lasso en escenarios como éste donde no todos los predictores son importantes para el modelo y se desea que los menos influyentes queden excluidos. Por otro lado, cuando existen predictores altamente correlacionados (linealmente), Ridge reduce la influencia de todos ellos a la vez y de forma proporcional, mientras que Lasso tiende a seleccionar uno de ellos, dándole todo el peso y excluyendo al resto. Se debe tener en cuenta entonces que, en presencia de correlaciones, esta selección varía mucho con pequeñas perturbaciones (cambios en los datos de entrenamiento), por lo que, las soluciones de Lasso, son muy inestables si los predictores están altamente correlacionados. Sin embargo, como el objetivo de este trabajo es encontrar un modelo entrenado con una muestra inicial fija extraída de los datos que representa el 75% de las observaciones, Lasso es de gran utilidad en la selección de variables influyentes. De todos modos se van a aplicar y analizar los dos métodos de regularización mencionados, y luego, en función de estos resultados y los obtenidos por los métodos de selección del apartado anterior, se decidirá cuáles son las variables más importantes.

La siguiente tabla muestra los coeficientes de ambos modelos utilizando cross-validation de 5-fold, con el parámetro lambda que minimiza el error cuadrático medio en cada caso.

	Ridge		Lasso
(Intercept)	2,87E+05	(Intercept)	0.4058761599
population	2,18E+04	population	.
householdsize	-2,82E+02	householdsize	.
racepctblack	**6,19E+04**	**racepctblack**	**0.0156344042**
racePctWhite	**-5,61E+04**	**racePctWhite**	**-0.1474717545**
racePctAsian	-3,40E+03	racePctAsian	.
racePctHisp	3,93E+03	racePctHisp	.
agePct12t21	-1,50E+04	agePct12t21	.
agePct12t29	-1,76E+04	agePct12t29	.
agePct16t24	-1,51E+04	**agePct16t24**	**-0.0001709439**
agePct65up	6,22E+03	agePct65up	.
numbUrban	2,48E+04	numbUrban	.
pctUrban	1,46E+04	**pctUrban**	**0.0176903860**
medIncome	-4,88E+03	medIncome	.
pctWWage	-1,34E+04	pctWWage	.
pctWFarmSelf	-1,24E+04	pctWFarmSelf	.
pctWInvInc	**-3,79E+04**	pctWInvInc	.
pctWSocSec	5,07E+03	pctWSocSec	.
pctWPubAsst	2,60E+04	pctWPubAsst	.
pctWRetire	-1,20E+04	pctWRetire	.
medFamInc	-8,22E+03	medFamInc	.
perCapInc	-9,32E+02	perCapInc	.
whitePerCap	1,51E+04	whitePerCap	.
blackPerCap	-1,07E+04	blackPerCap	.
indianPerCap	-1,04E+03	indianPerCap	.
AsianPerCap	1,33E+04	AsianPerCap	.
OtherPerCap	1,53E+04	OtherPerCap	.
HispPerCap	9,42E+03	HispPerCap	.
NumUnderPov	2,94E+04	NumUnderPov	.
PctPopUnderPov	1,38E+04	PctPopUnderPov	.
PctLess9thGrade	4,15E+03	PctLess9thGrade	.
PctNotHSGrad	1,85E+04	PctNotHSGrad	.
PctBSorMore	-9,02E+03	PctBSorMore	.
PctUnemployed	1,53E+04	PctUnemployed	.
PctEmploy	-1,01E+04	PctEmploy	.
PctEmplManu	-1,40E+04	PctEmplManu	.
PctEmplProfServ	-5,04E+03	PctEmplProfServ	.
PctOccupManu	3,71E+03	PctOccupManu	.
PctOccupMgmtProf	-7,85E+03	PctOccupMgmtProf	.
MalePctDivorce	**4,33E+04**	**MalePctDivorce**	**0.1128718718**
MalePctNevMarr	2,07E+04	MalePctNevMarr	.

FemalePctDiv	**4,05E+04**	FemalePctDiv	.
TotalPctDiv	**4,09E+04**	TotalPctDiv	.
PersPerFam	1,38E+04	PersPerFam	.
PctFam2Par	**-5,17E+04**	PctFam2Par	.
PctKids2Par	**-5,70E+04**	**PctKids2Par**	**-0.2712687714**
PctYoungKids2Par	**-4,46E+04**	PctYoungKids2Par	.
PctTeen2Par	**-5,27E+04**	PctTeen2Par	.
PctWorkMomYoungKids	8,78E+02	PctWorkMomYoungKids	.
PctWorkMom	-1,55E+04	**PctWorkMom**	**-0.0402445149**
NumIlleg	**5,45E+04**	NumIlleg	.
PctIlleg	**6,39E+04**	**PctIlleg**	**0.2074528984**
NumImmig	8,19E+03	NumImmig	.
PctImmigRecent	4,59E+03	PctImmigRecent	.
PctImmigRec5	3,14E+03	PctImmigRec5	.
PctImmigRec8	3,99E+03	PctImmigRec8	.
PctImmigRec10	7,23E+03	PctImmigRec10	.
PctRecentImmig	5,22E+03	PctRecentImmig	.
PctRecImmig5	6,52E+03	PctRecImmig5	.
PctRecImmig8	7,80E+03	PctRecImmig8	.
PctRecImmig10	8,87E+03	PctRecImmig10	.
PctSpeakEnglOnly	-1,39E+03	PctSpeakEnglOnly	.
PctNotSpeakEnglWell	3,02E+02	PctNotSpeakEnglWell	.
PctLargHouseFam	2,17E+04	PctLargHouseFam	.
PctLargHouseOccup	1,65E+04	PctLargHouseOccup	.
PersPerOccupHous	5,56E+03	PersPerOccupHous	.
PersPerOwnOccHous	-1,25E+04	PersPerOwnOccHous	.
PersPerRentOccHous	1,69E+04	PersPerRentOccHous	.
PctPersOwnOccup	-1,96E+04	PctPersOwnOccup	.
PctPersDenseHous	2,27E+04	**PctPersDenseHous**	**0.0634871692**
PctHousLess3BR	1,79E+04	PctHousLess3BR	.
MedNumBR	-9,76E+02	MedNumBR	.
HousVacant	**3,95E+04**	**HousVacant**	**0.0918822278**
PctHousOccup	-3,14E+04	**PctHousOccup**	**-0.0301578528**
PctHousOwnOcc	-1,28E+04	PctHousOwnOcc	.
PctVacantBoarded	**4,74E+04**	**PctVacantBoarded**	**0.0569659988**
PctVacMore6Mos	-7,56E+03	PctVacMore6Mos	.
MedYrHousBuilt	3,35E+03	MedYrHousBuilt	.
PctHousNoPhone	1,91E+04	PctHousNoPhone	.
PctWOFullPlumb	9,44E+03	PctWOFullPlumb	.
OwnOccLowQuart	-2,72E+03	OwnOccLowQuart	.
OwnOccMedVal	-8,03E+01	OwnOccMedVal	.
OwnOccHiQuart	2,05E+03	OwnOccHiQuart	.
RentLowQ	-7,76E+03	RentLowQ	.

RentMedian	3,34E+02	RentMedian	.
RentHighQ	1,47E+03	RentHighQ	.
MedRent	5,48E+03	MedRent	.
MedRentPctHousInc	3,02E+04	**MedRentPctHousInc**	**0.0081665663**
MedOwnCostPctInc	1,03E+04	MedOwnCostPctInc	.
MedOwnCostPctIncNoMtg	-1,06E+04	MedOwnCostPctIncNoMtg	.
NumInShelters	**4,27E+04**	NumInShelters	.
NumStreet	**6,93E+04**	**NumStreet**	**0.1637213380**
PctForeignBorn	7,62E+03	PctForeignBorn	.
PctBornSameState	-7,60E+03	PctBornSameState	.
PctSameHouse85	-4,89E+02	PctSameHouse85	.
PctSameCity85	1,39E+04	PctSameCity85	.
PctSameState85	8,83E+02	PctSameState85	.
LandArea	9,85E+03	LandArea	.
PopDens	1,03E+04	PopDens	.
PctUsePubTrans	7,92E+03	PctUsePubTrans	.
LemasPctOfficDrugUn	2,05E+04	LemasPctOfficDrugUn	.

Se puede notar que en el modelo de Ridge los 16 coeficientes con mayor valor absoluto corresponden a las variables:
NumStreet - PctIlleg - racepctblack - PctKids2Par - racePctWhite - NumIlleg - PctTeen2Par - PctFam2Par - PctVacantBoarded - PctYoungKids2Par - MalePctDivorce - NumInShelters - TotalPctDiv - FemalePctDiv - HousVacant - pctWInvInc.

Mientras que las que poseen coeficiente no nulo según el modelo de Lasso son las siguientes 14:
PctKids2Par - PctIlleg - NumStreet - racePctWhite - MalePctDivorce - HousVacant - PctPersDenseHous - PctVacantBoarded - PctWorkMom - PctHousOccup - pctUrban - racepctblack - MedRentPctHousInc - agePct16t24.

Las variables que coinciden en ambos métodos son 8, y se encuentran resaltadas en color naranja en la tabla anterior: NumStreet, PctIlleg, racepctblack, PctKids2Par, racePctWhite, PctVacantBoarded, MalePctDivorce, HousVacant.

Resumiendo, las variables que se desprenden de alguno de los dos métodos de regularización son 22: NumStreet, PctIlleg, racepctblack, PctKids2Par, racePctWhite, PctVacantBoarded, MalePctDivorce, HousVacant, NumIlleg, PctTeen2Par, PctFam2Par, PctYoungKids2Par, NumInShelters, TotalPctDiv, FemalePctDiv, pctWInvInc, PctPersDenseHous, PctWorkMom, PctHousOccup, pctUrban, MedRentPctHousInc, agePct16t24.

Si se buscan las coincidencias de variables comparando con las 13 variables obtenidas con los métodos de selección del apartado previo, se observa que hay 12 de ellas que resaltan al aplicar Lasso, de las cuales 8 fueron variables de coeficiente alto según Ridge. La restante es PctPopUnderPov que sólo fue obtenida al aplicar el método Forward.

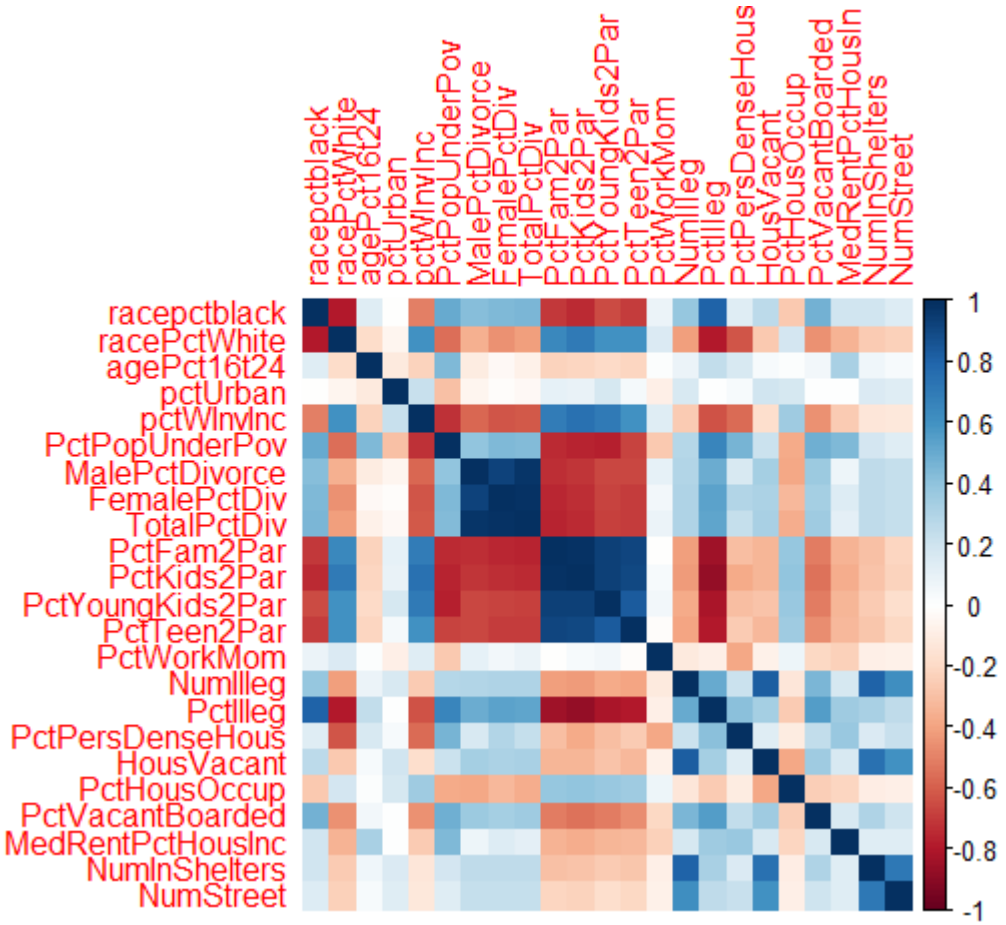
Se procede ahora a considerar todas las variables mencionadas anteriormente, es decir las 22 destacadas al regularizar, que incluye 12 variables extraídas al aplicar algún método de selección, y, a parte, la que se obtiene sólo a partir del Forward. La idea es usar todas estas 23 variables que fueron elegidas por algún método y aplicar nuevamente ambos métodos de regularización para ver qué variables tienen mayor influencia. Otra vez, se utiliza cross-validation de 5-fold, con el parámetro lambda que minimiza el error cuadrático medio en cada caso. En la siguiente tabla se muestran los coeficientes obtenidos.

	Ridge		**Lasso**
(Intercept)	0.36592597	(Intercept)	0.42907677
racepctblack	0.06918766	racepctblack	.
racePctWhite	-0.08341148	racePctWhite	-0.15696972
agePct16t24	-0.02671803	agePct16t24	.
pctUrban	0.02335627	pctUrban	0.01125387
pctWInvInc	-0.05901572	pctWInvInc	.
PctPopUnderPov	0.01800544	PctPopUnderPov	.
MalePctDivorce	0.04935234	MalePctDivorce	0.09643079
FemalePctDiv	0.04772688	FemalePctDiv	.
TotalPctDiv	0.04667606	TotalPctDiv	.
PctFam2Par	-0.06642007	PctFam2Par	.
PctKids2Par	-0.07683528	PctKids2Par	-0.28831815
PctYoungKids2Par	-0.05601749	PctYoungKids2Par	.
PctTeen2Par	-0.06289266	PctTeen2Par	.
PctWorkMom	-0.04593266	PctWorkMom	-0.02617747
NumIlleg	0.07803950	NumIlleg	.
PctIlleg	0.08863219	PctIlleg	0.20164281
PctPersDenseHous	0.07272904	PctPersDenseHous	0.05412657
HousVacant	0.05949132	HousVacant	0.10014512
PctHousOccup	-0.03820045	PctHousOccup	-0.01692841
PctVacantBoarded	0.06044115	PctVacantBoarded	0.05404898
MedRentPctHousInc	0.05346958	MedRentPctHousInc	.
NumInShelters	0.06678444	NumInShelters	.
NumStreet	0.11549461	NumStreet	0.14416900

Según Ridge los valores absolutos de los coeficientes en orden decreciente corresponden a: NumStreet, PctIlleg, racePctWhite, NumIlleg, PctKids2Par, PctPersDenseHous, racepctblack, NumInShelters, PctFam2Par, PctTeen2Par, PctVacantBoarded, HousVacant,

pctWInvInc, PctYoungKids2Par, MedRentPctHousInc, MalePctDivorce, FemalePctDiv, TotalPctDiv, PctWorkMom, PctHousOccup, agePct16t24, pctUrban, PctPopUnderPov.

Mientras que con el método de Lasso los coeficientes no nulos en orden decreciente de sus valores absolutos provienen de las variables: **PctKids2Par, PctIlleg, racePctWhite, NumStreet, HousVacant, MalePctDivorce, PctPersDenseHous, PctVacantBoarded, PctWorkMom, PctHousOccup, pctUrban**. Estas últimas fueron sombreadas en naranja en la tabla anterior. Se puede observar que las variables no resaltadas en color, es decir, que no son destacadas según Lasso, o bien toman un coeficiente pequeño comparado con los demás calculados por Ridge, o bien están altamente correlacionadas con las variables sombreadas. Por esta razón, se van a considerar las 11 variables con coeficiente no nulo según Lasso. A continuación se muestra la matriz de correlaciones de las 23 variables seleccionadas para poder verificar la explicación previa.



Regresión lineal

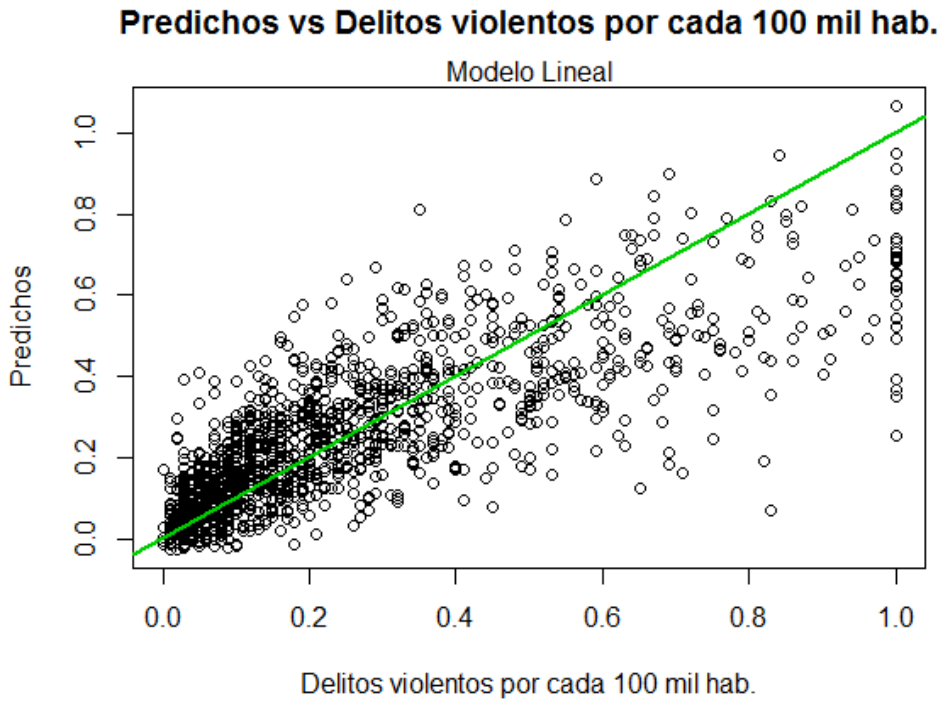
Luego del estudio realizado en la sección anterior, se utilizan las 11 covariables que fueron seleccionadas, y, basándose en la misma muestra de entrenamiento predeterminada que considera el 75% de los datos, se propone la realización de diferentes modelos de regresión a fin de encontrar aquel que permita predecir lo mejor posible la cantidad de *delitos violentos por cada cien mil habitantes*.

En primer lugar se aplica una regresión lineal, obteniendo los siguientes coeficientes, cuya significatividad en el modelo es muy alta, lo cual confirma que las variables elegidas son muy influyentes en la cantidad de *delitos violentos cada 100 mil habitantes*.

Coeficientes	Estimación	Error estándar	t-valor	p-valor
(Intercept)	0.429543	0.057096	7.523	9.22e-14***
racePctWhite	-0.161963	0.029655	-5.462	5.53e-08***
pctUrban	0.042083	0.008872	4.743	2.31e-06***
MalePctDivorce	0.146190	0.031938	4.577	5.10e-06***
PctKids2Par	-0.260581	0.050666	-5.143	3.06e-07***
PctWorkMom	-0.077176	0.023342	-3.306	0.000968***
PctIlleg	0.220826	0.041242	5.354	9.94e-08***
PctPersDenseHous	0.070527	0.024515	2.877	0.004073**
HousVacant	0.068765	0.034772	1.978	0.048157*
PctHousOccup	-0.076951	0.022814	-3.373	0.000763***
PctVacantBoarded	0.072150	0.021257	3.394	0.000706***
NumStreet	0.227787	0.047462	4.799	1.75e-06***

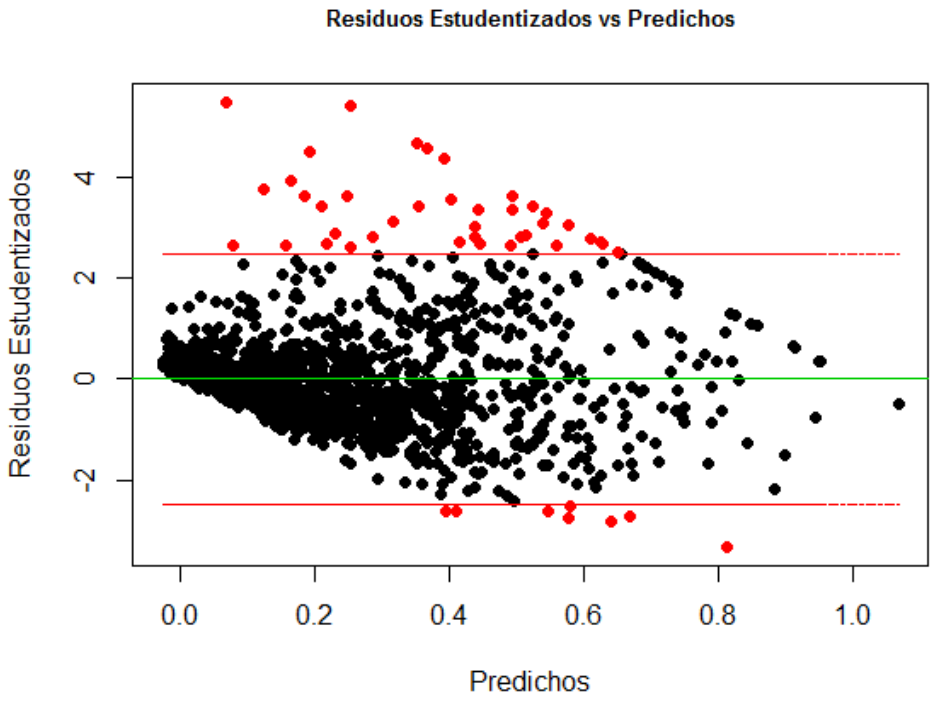
Este modelo tiene un R^2 de 0.6514 y un R^2 ajustado de **0.6488** lo que dice que en un 64,88% está explicada la variabilidad de la respuesta *delitos violentos por cada 100 mil habitantes.* Además, el p-valor del modelo es **< 2.2e-16,** muy cercano a 0, que indica que la regresión lineal ajusta bien los datos.

A continuación se observa el gráfico de valores predichos contra la variable respuesta.



En este caso el error cuadrático medio (**ECM**) es aproximadamente **0.0194**, y el error medio absoluto, mejor conocido como **MAE**, por sus siglas en inglés *mean absolute error*, se estima en **0.0974**. Ambos valores serán considerados al finalizar para comparar todos los modelos aplicados.

Por otro lado, se calculan los residuos estudentizados y se analizan en función de los valores predichos bajo la regresión lineal. El siguiente gráfico destaca en color rojo los puntos cuyo residuo estudentizado toma valor absoluto superior a 2.5. Y aquí se plantea el problema de rehacer la regresión pero quitando los datos atípicos (outliers) que se resolverá en el apartado que sigue.



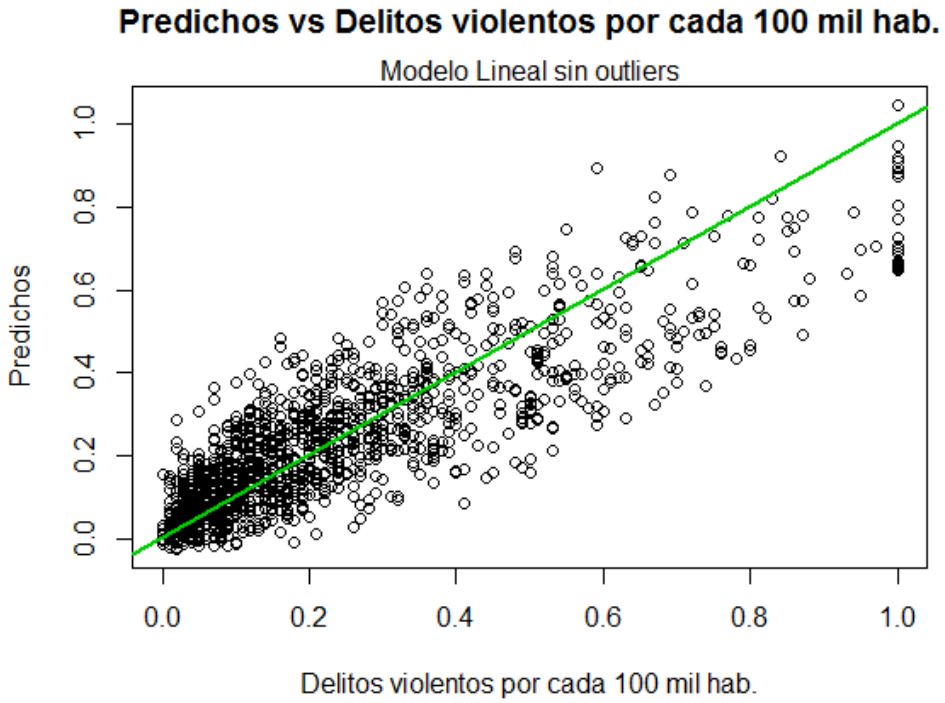
Regresión lineal sin outliers

En esta sección se consideran como outliers los puntos rojos del gráfico anterior y se remueven del conjunto de entrenamiento para realizar una nueva regresión lineal. Los resultados que se obtienen se reflejan en la tabla de abajo:

Coeficientes	Estimación	Error estándar	t-valor	p-valor
(Intercept)	0.356619	0.047785	7.463	1.46e-13***
racePctWhite	-0.141882	0.024422	-5.810	7.70e-09***
pctUrban	0.037460	0.007354	5.094	3.97e-07***
MalePctDivorce	0.144443	0.026423	5.467	5.40e-08***
PctKids2Par	-0.197147	0.042395	-4.650	3.62e-06***
PctWorkMom	-0.072457	0.019288	-3.757	0.000179***
PctIlleg	0.285939	0.034619	8.260	3.28e-16***
PctPersDenseHous	0.072420	0.020407	3.549	0.000400***
HousVacant	0.082767	0.028669	2.887	0.003948**
PctHousOccup	-0.079875	0.018729	-4.265	2.13e-05***
PctVacantBoarded	0.038063	0.017758	2.143	0.032245*
NumStreet	0.239416	0.038903	6.154	9.77e-10***

Nuevamente se destacan todas las covariables con alto grado de significatividad, siendo además el R^2 igual a 0.7235 y el R^2 ajustado de **0.7214**, ambos valores mayores que los estimados en la regresión lineal inicial. El p-valor del modelo se mantiene muy bajo: **< 2.2e-16.**

Se realiza el gráfico de predichos versus la variable respuesta que se presenta debajo, y se calculan los errores: el **ECM** estimado es de **0.0128**, y el **MAE** es **0.0832**.



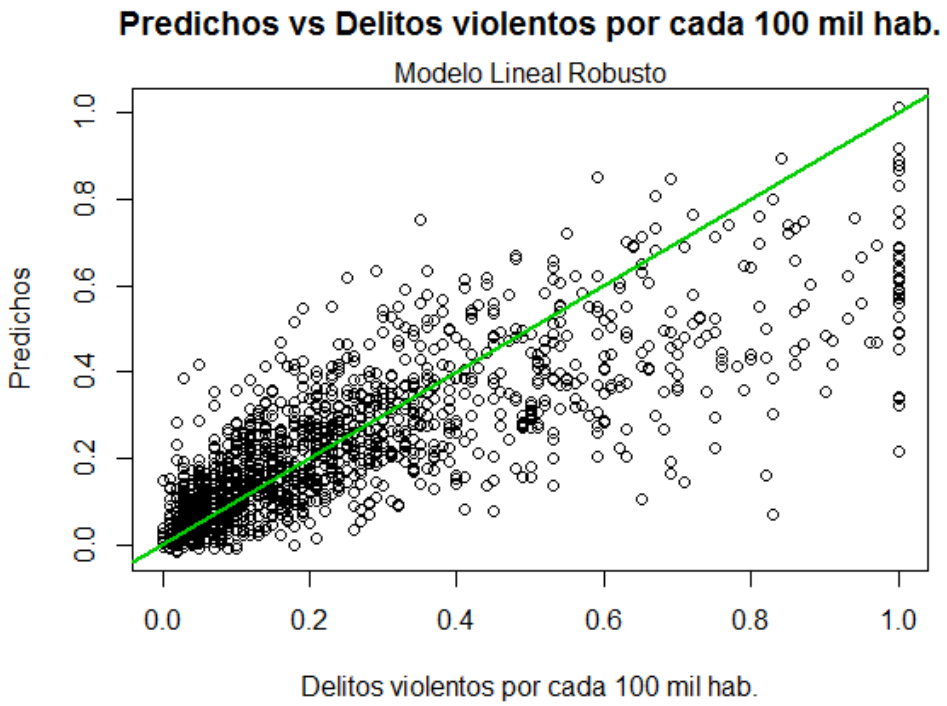
Regresión lineal robusta

La tercera propuesta es aplicar un modelo de regresión lineal robusta. Para ello utilizaremos el método robusto creado por Yohai, de tipo MM, con la función *lmrob* del paquete *robustbase* de R. Los coeficientes estimados se muestran en la siguiente tabla y todos resultan significativos:

Coeficientes	Estimación	Error estándar	t-valor	p-valor
(Intercept)	0.334918	0.045829	7.308	4.41e-13***
racePctWhite	-0.144593	0.023632	-6.119	1.21e-09***

pctUrban	0.032196	0.006937	4.641	3.76e-06***
MalePctDivorce	0.115575	0.025042	4.615	4.26e-06***
PctKids2Par	-0.175445	0.040179	-4.367	1.35e-05***
PctWorkMom	-0.063691	0.018210	-3.498	0.000483***
PctIlleg	0.286477	0.033256	8.614	<2e-16***
PctPersDenseHous	0.078465	0.019750	3.973	7.44e-05***
HousVacant	0.089993	0.027643	3.256	0.001157**
PctHousOccup	-0.061232	0.017864	-3.428	0.000626***
PctVacantBoarded	0.039665	0.017144	2.314	0.020824*
NumStreet	0.226984	0.037756	6.012	2.31e-09***

El valor de R^2 de este modelo es igual a 0.7059 y el R^2 ajustado es **0.7037.** El scatter-plot correspondiente a los valores predichos contra la variable respuesta es el que se observa a continuación.



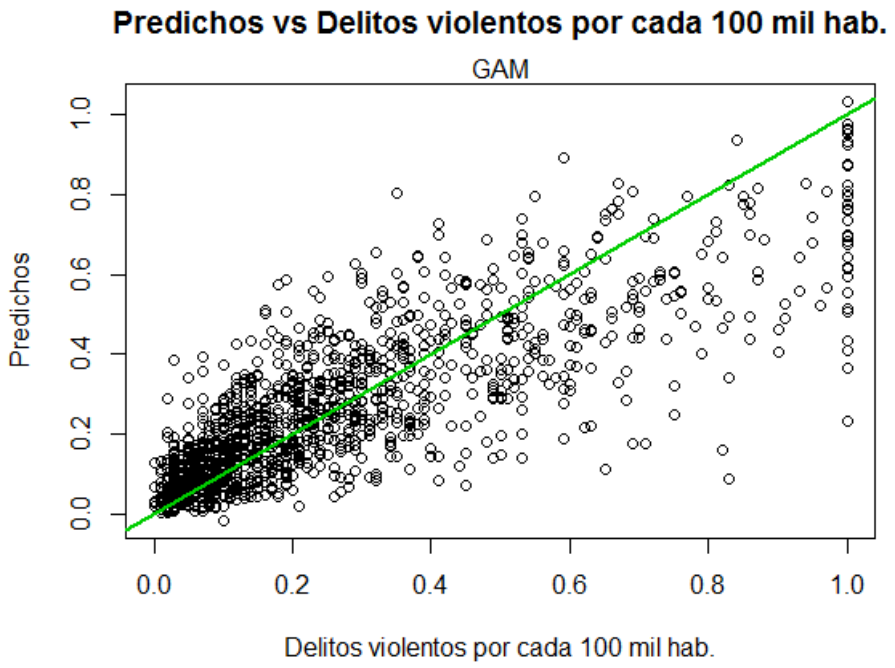
Los errores estimados son: **ECM** = **0.0201**, y **MAE** = **0.0946.**

Modelo aditivo generalizado (Generalized additived model: GAM)

El siguiente modelo aplicado es el aditivo generalizado, cuyos resultados se muestran en la tabla que se muestra debajo y donde se observa que hay dos covariables que dejan de ser significativas en este caso: pctUrban y PctHousOccup.

Coeficientes	Df	Sum Sq	Mean Sq	F-valor	p-valor
s(racePctWhite)	1	35.577	35.577	1882.4370	<2.2e-16***
s(pctUrban)	1	0.009	0.009	0.4664	0.4947593
s(MalePctDivorce)	1	9.281	9.281	491.0476	<2.2e-16***
s(PctKids2Par)	1	3.968	3.968	209.9346	<2.2e-16***
s(PctWorkMom)	1	0.744	0.744	39.3795	4.598e-10***
s(PctIlleg)	1	0.426	0.426	22.5389	2.263e-06***
s(PctPersDenseHous)	1	0.553	0.553	29.2434	7.456e-08***
s(HousVacant)	1	1.258	1.258	66.5518	7.312e-16***
s(PctHousOccup)	1	0.002	0.002	0.1217	0.7272687
s(PctVacantBoarded)	1	0.209	0.209	11.0821	0.0008935***
s(NumStreet)	1	0.485	0.485	25.6532	4.611e-07***
Residuals	1450	27.404	0.019		

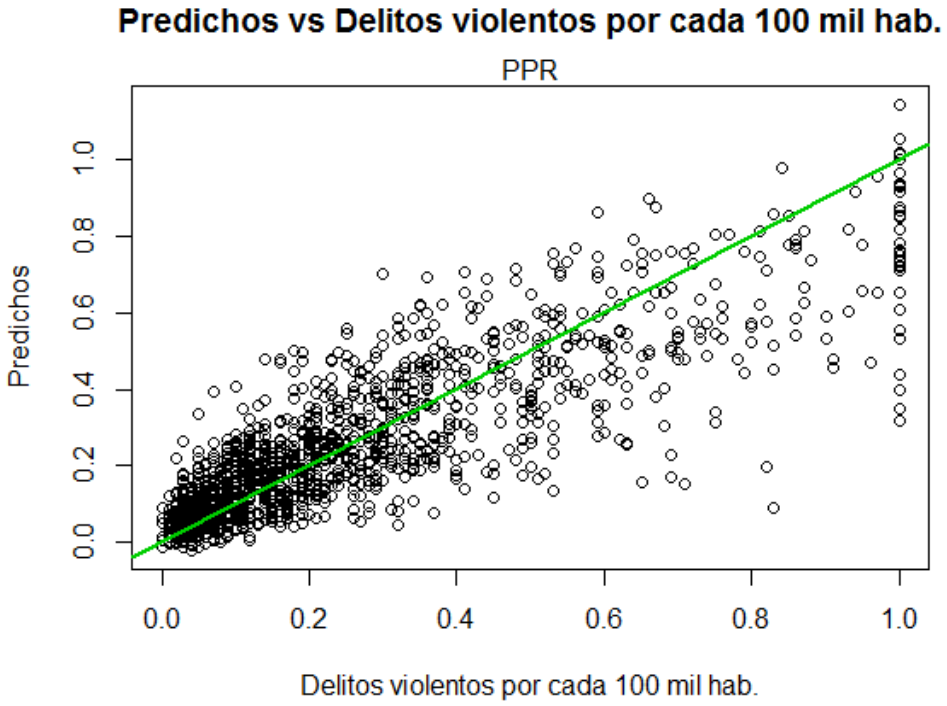
El gráfico de dispersión entre predichos y la variable respuesta se presenta a continuación.



Los errores estimados son: **ECM** = **0.0183** y **MAE** = **0.0933**.

Regresión de búsqueda de proyección (Projection pursuit regression: PPR)

En este caso la salida de R es más compleja, por lo que sólo se mostrará el gráfico de dispersión de predichos contra la variable respuesta, y se calcularán los errores.



Los errores estimados son: **ECM** = **0.0156** y **MAE** = **0.0864**.

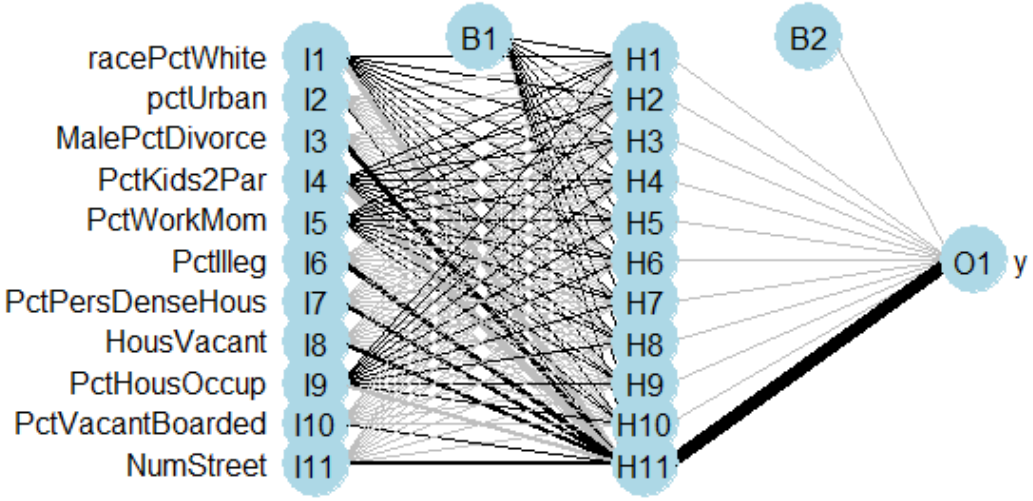
Redes neuronales (Neural networks: NN)

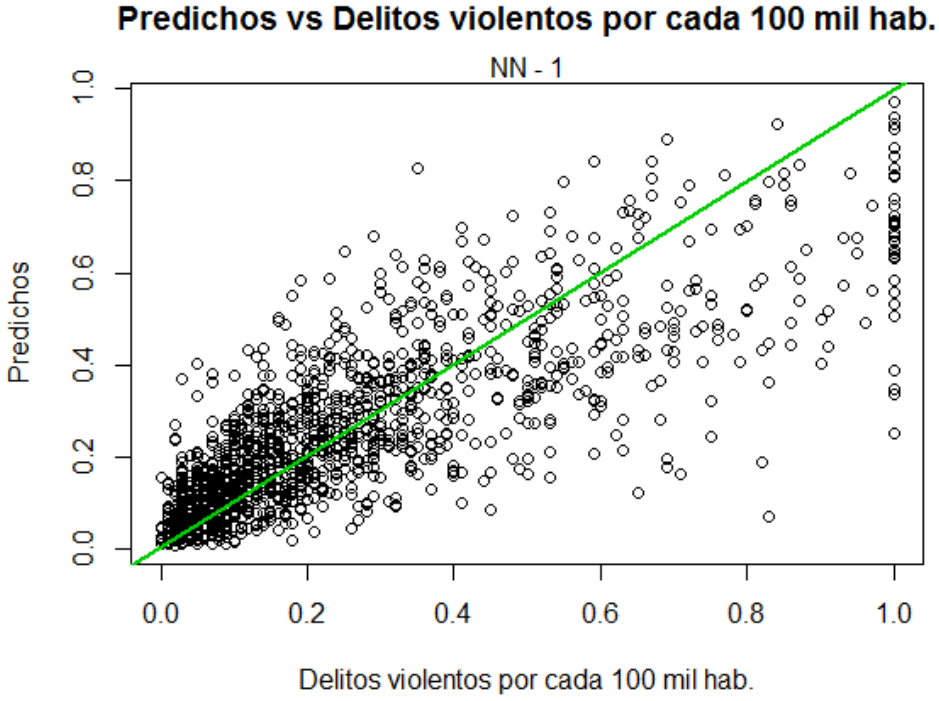
Se proponen 8 modelos de redes neuronales para evaluar su comportamiento, dada la amplia variedad de los parámetros a considerar en los argumentos de la función del paquete correspondiente a redes neuronales en R. Entre los parámetros con los que se cuenta se destacan el tamaño de las capas ocultas (*size*), el parámetro de decaimiento (*decay*), y si la salida considera funciones lineales o no (*linout*).

En cada caso, fijando una cierta semilla, se mostrará el gráfico con los nodos correspondientes, donde el grosor de las ramas que los unen indica los pesos utilizados. Siempre se parte de las 11 covariables elegidas como input (I), luego continúa con las capas ocultas (H), hasta finalmente llegar a la respuesta o salida (O).

Se presentará el scatterplot de predichos contra la variable respuesta y se calcularán los errores ECM y MAE para su posterior comparación.

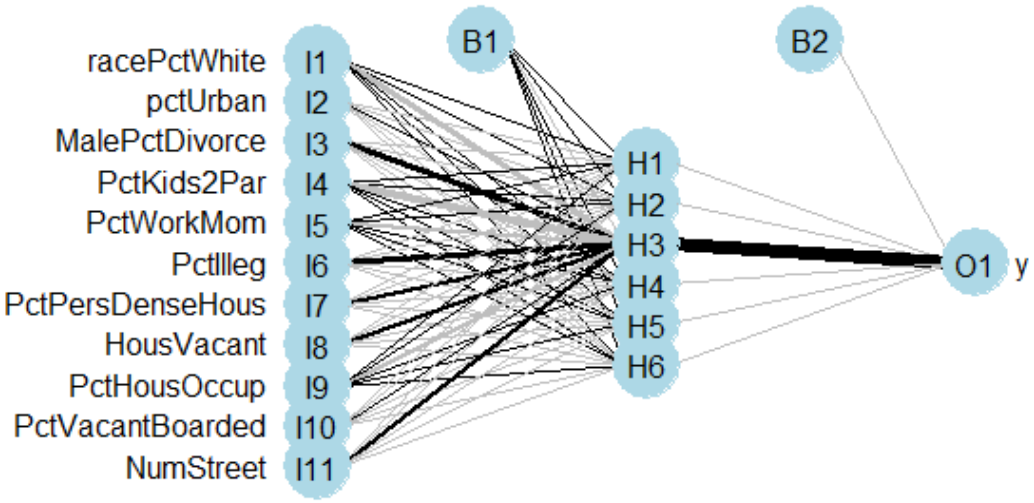
Modelo NN-1: size=11, decay=0.5, linout=TRUE

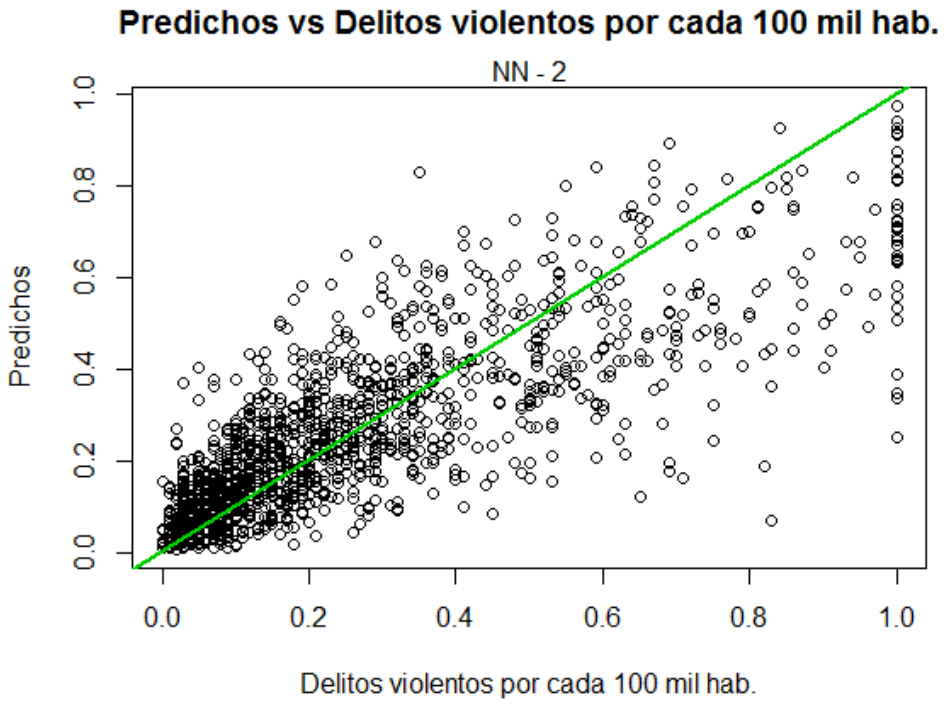




Los errores estimados son: **ECM** = **0.019116** y **MAE** = **0.095743**.

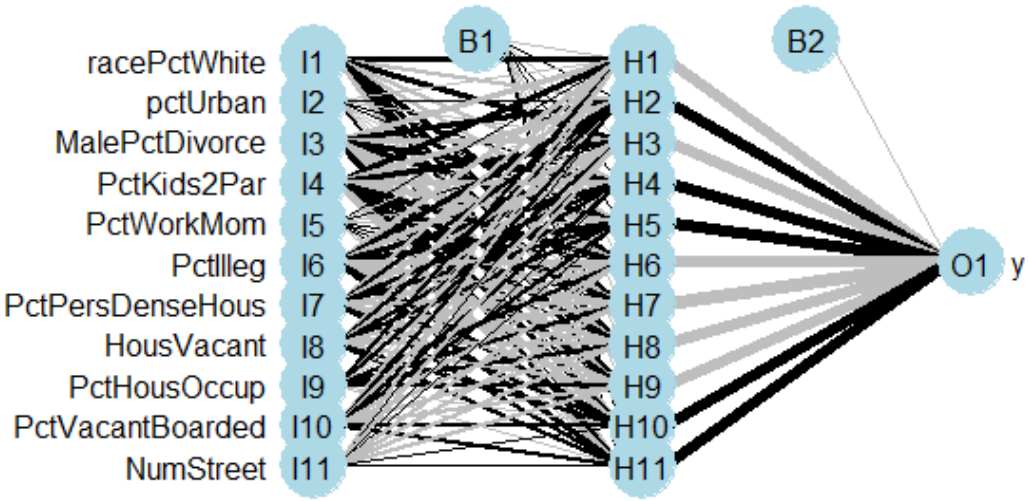
Modelo NN-2: size=6, decay=0.5, linout=TRUE

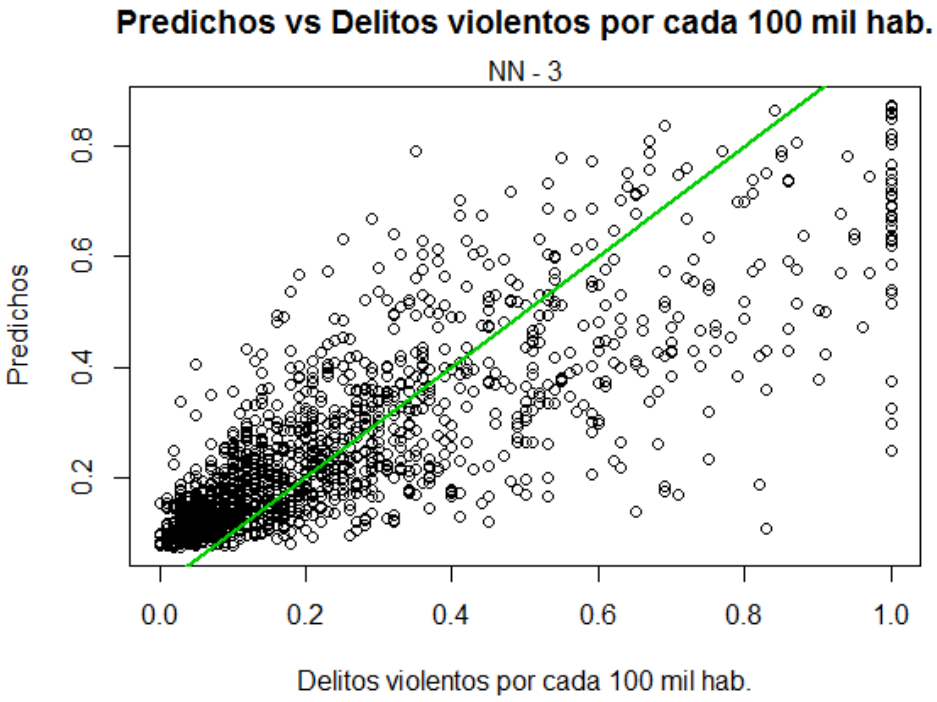




Los errores estimados son: **ECM** = **0.019118** y **MAE** = **0.095747**.

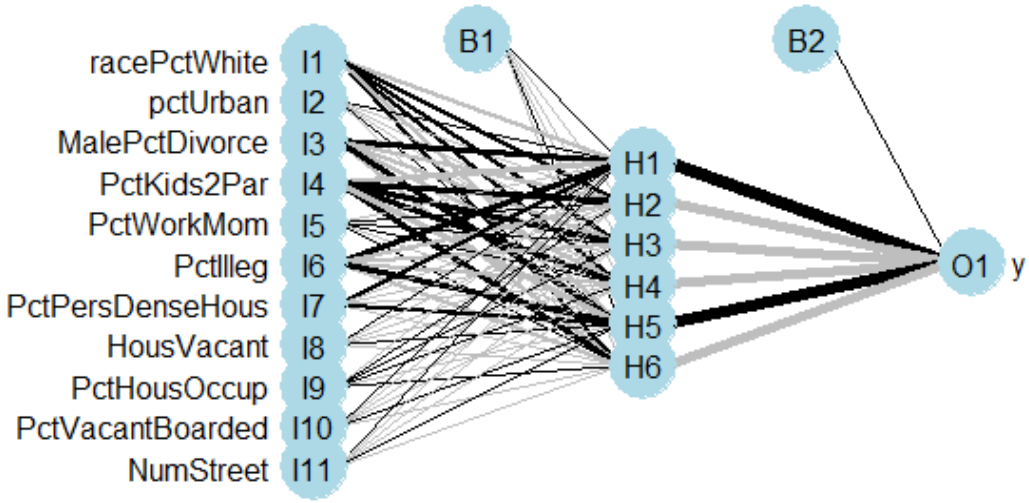
Modelo NN-3: size=11, decay=0.5, linout=FALSE

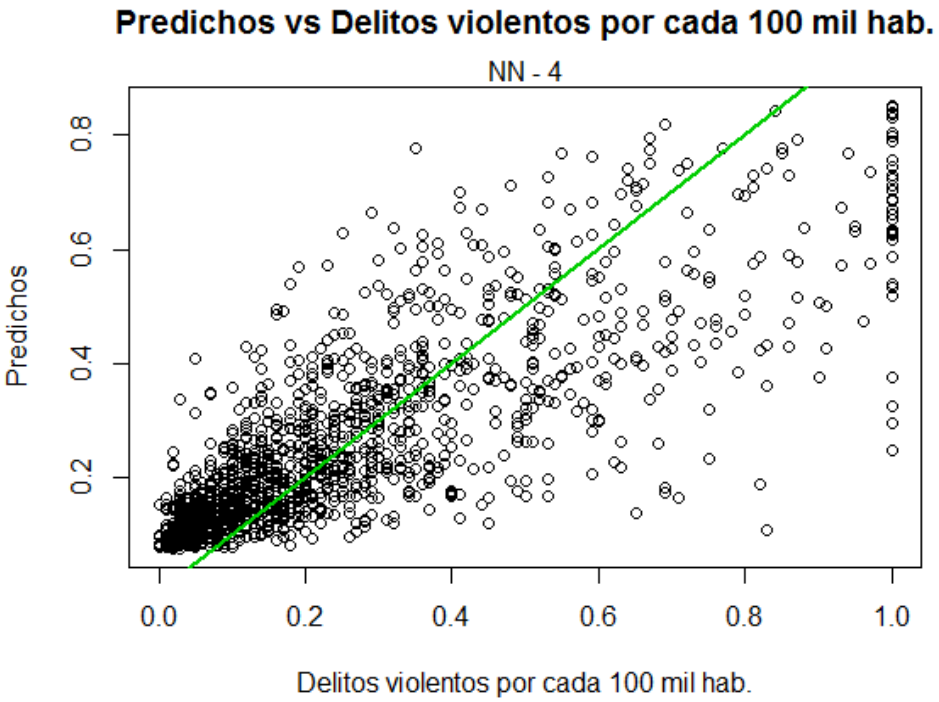




Los errores estimados son: **ECM** = **0.0198** y **MAE** = **0.1017**.

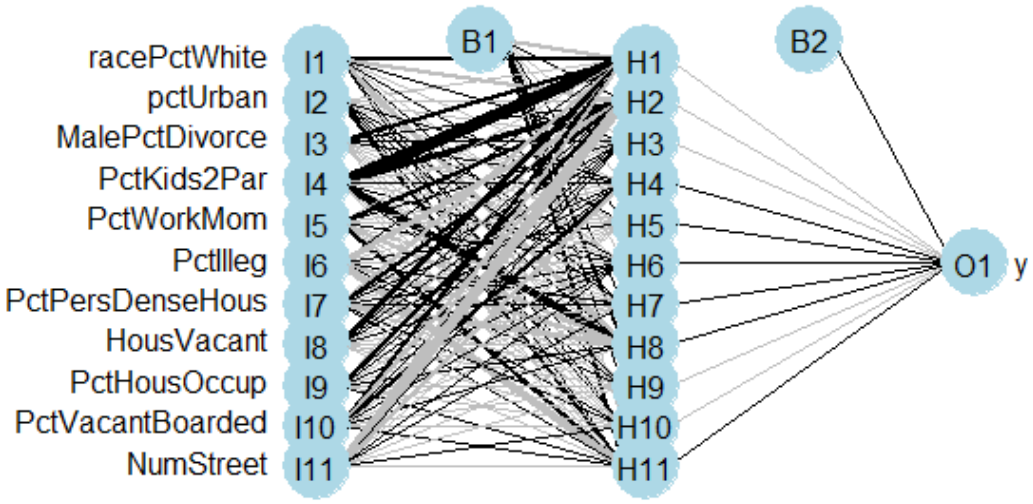
Modelo NN-4: size=6, decay=0.5, linout=FALSE

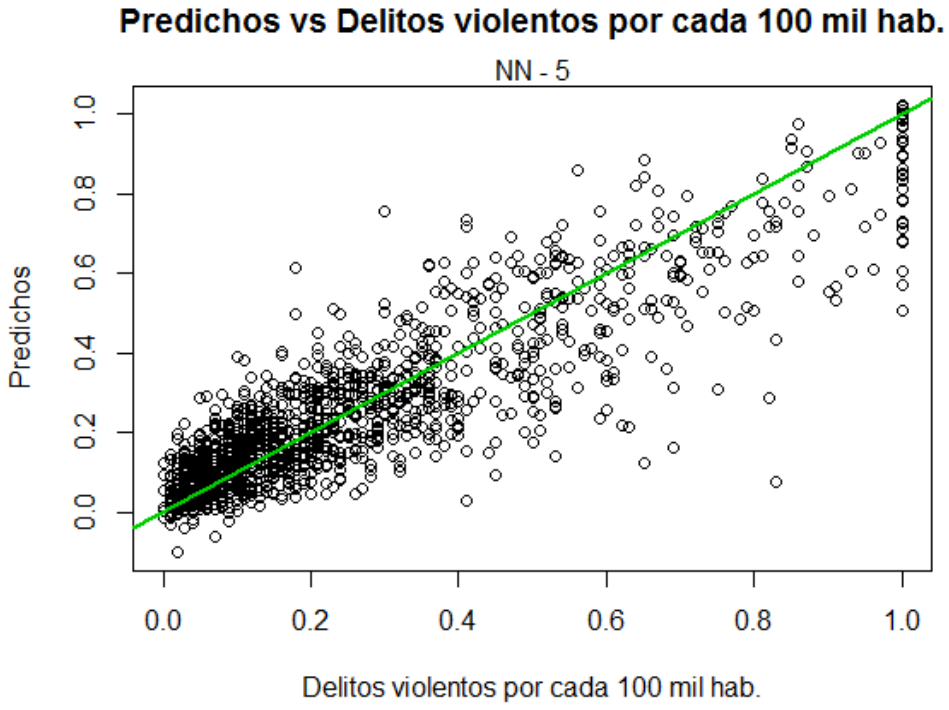




Los errores estimados son: **ECM** = **0.0198** y **MAE** = **0.1019**.

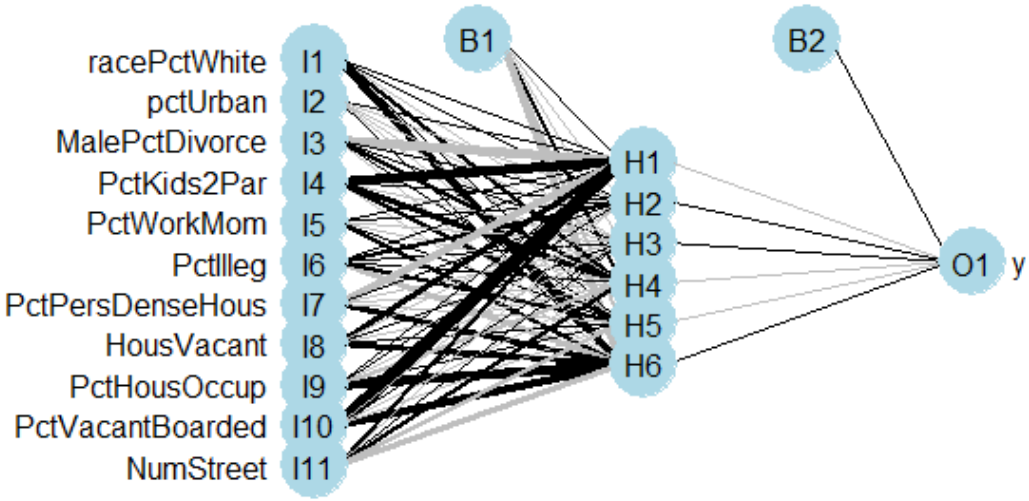
Modelo NN-5: size=11, decay=0, linout=TRUE

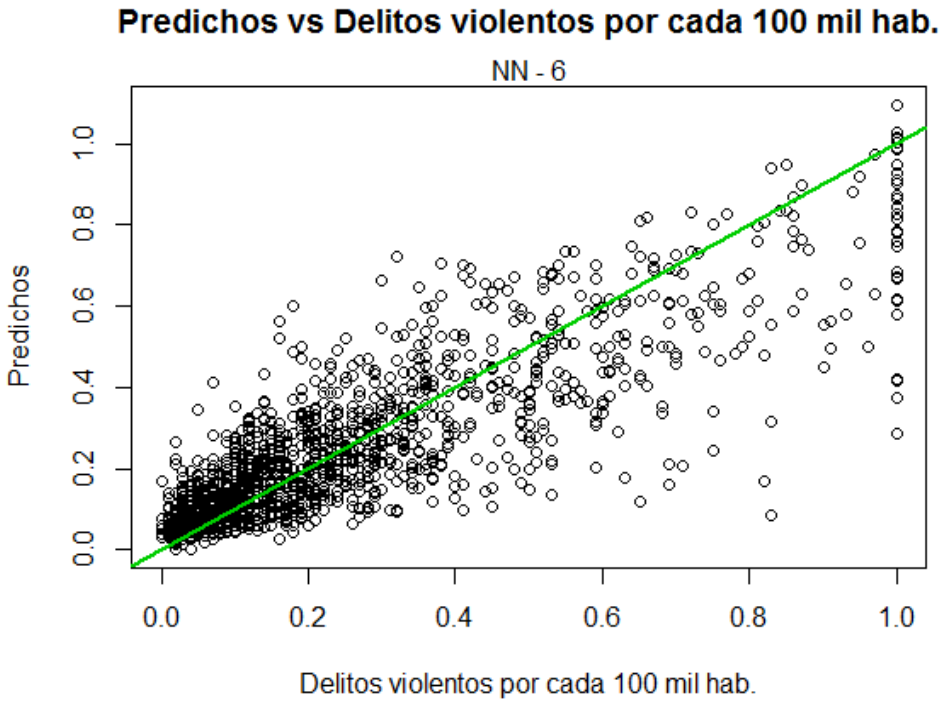




Los errores estimados son: **ECM** = **0.0121** y **MAE** = **0.0776**.

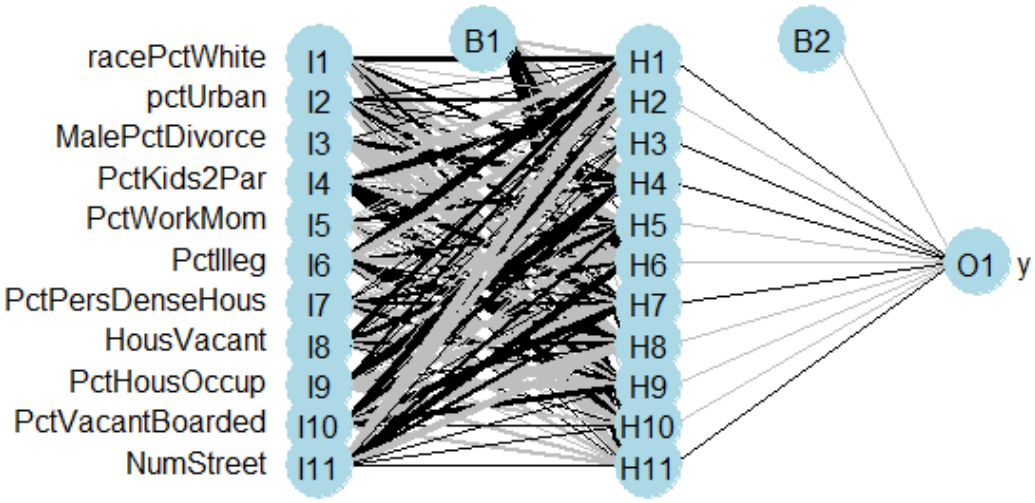
Modelo NN-6: size=6, decay=0, linout=TRUE

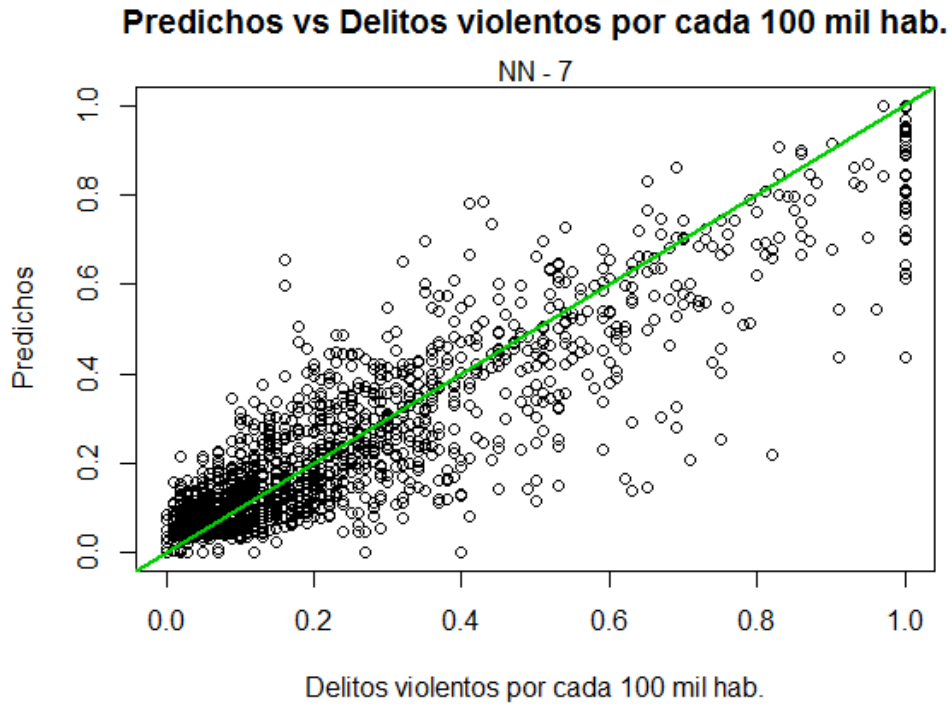




Los errores estimados son: **ECM** = **0.0155** y **MAE** = **0.0857**.

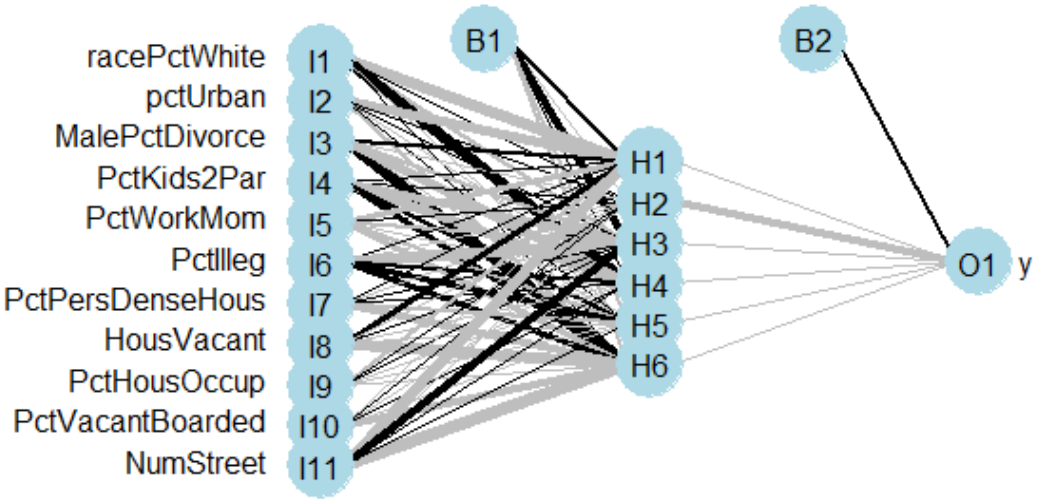
Modelo NN-7: size=11, decay=0, linout=FALSE

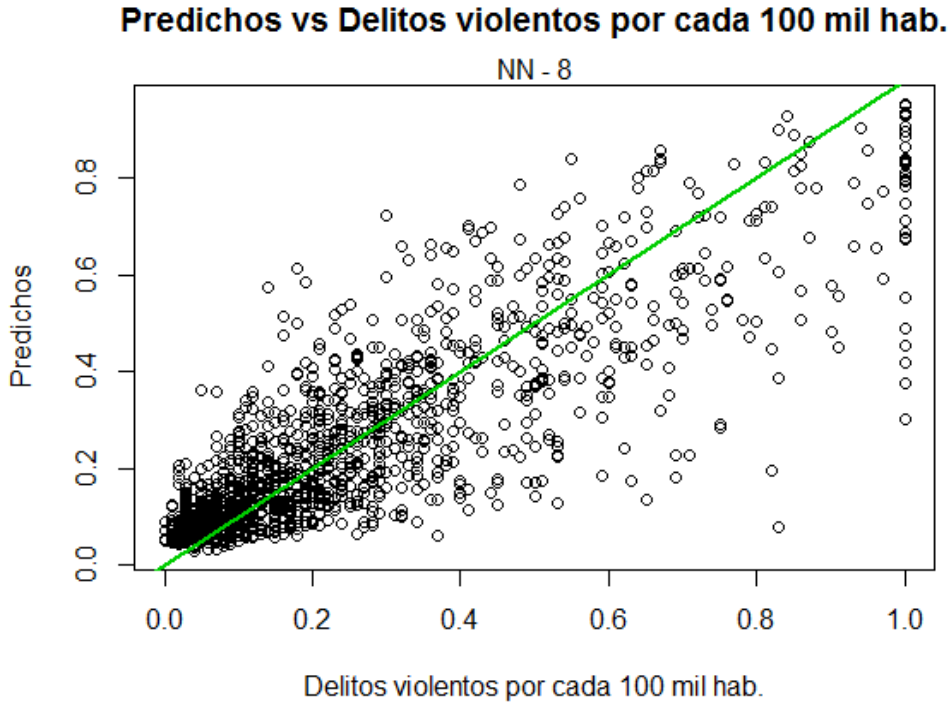




Los errores estimados son: **ECM** = **0.01197** y **MAE** = **0.07587**.

Modelo NN-8: size=6, decay=0, linout=FALSE





Los errores estimados son: **ECM** = **0.0157** y **MAE** = **0.0865**.

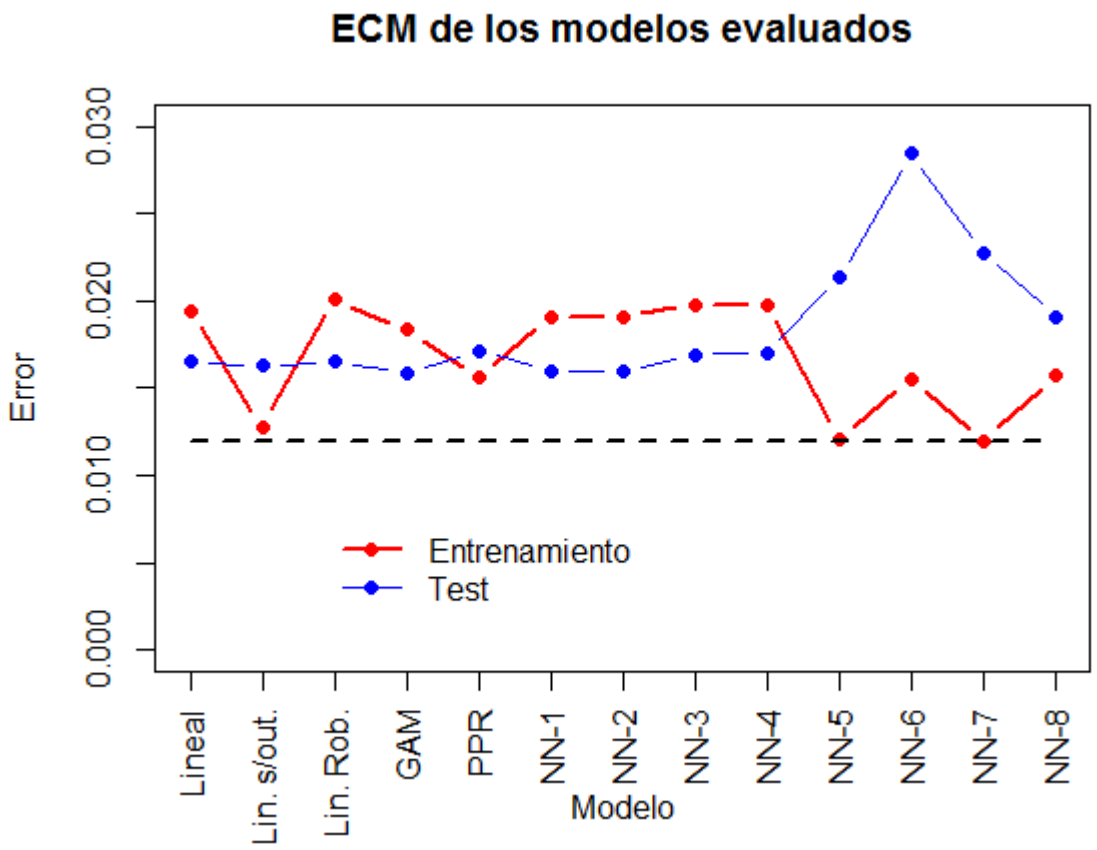
Conclusiones

El objetivo final de este trabajo es la selección de un modelo con mayor capacidad de generalización y predicción de la variable *delitos violentos por cada cien mil habitantes*, basado en las 11 covariables (*PctKids2Par, PctIlleg, racePctWhite, NumStreet, HousVacant, MalePctDivorce, PctPersDenseHous, PctVacantBoarded, PctWorkMom, PctHousOccup, pctUrban*) que fueron elegidas, luego de los estudios previos de la base de datos utilizada.

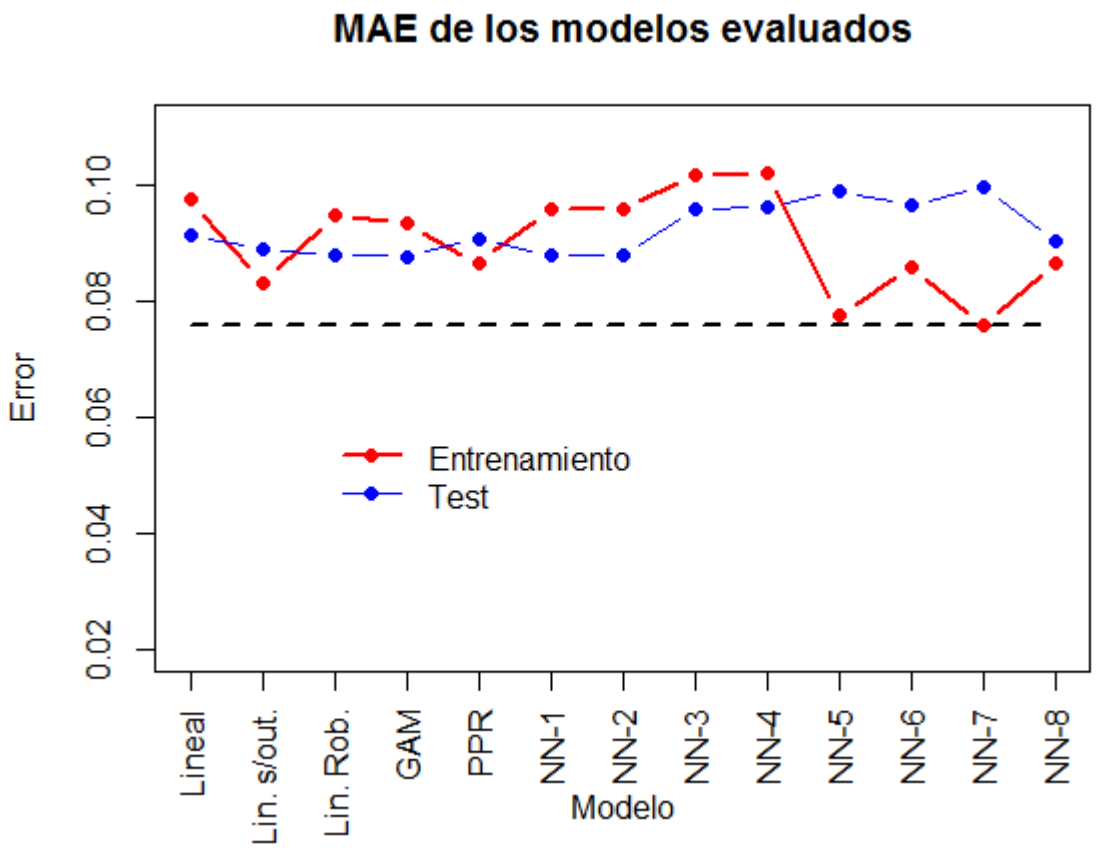
Para concluir con todo el análisis anterior, y alcanzar el objetivo de encontrar el mejor modelo se compararon los errores cuadráticos medios (ECMs) y los errores medios absolutos (MAEs) de los 13 modelos evaluados tanto con los datos de entrenamiento como con los datos reservados para test. Los resultados se muestran en la tabla que sigue, con los errores más bajos sombreados en color.

	Training		**Testing**	
Modelo	**ECM: MSE**	**MAE**	**ECM: MSE**	**MAE**
Lineal	0.01944695	0.09738091	0.01653147	0.0911468
Lineal sin outliers	0.01278311	0.08315471	0.01629022	0.0888294
Lineal Robusto	0.02011858	0.09457756	0.01654393	0.08775212
GAM	0.01833056	0.09325264	**0.01586487**	**0.08767052**
PPR	0.01558196	0.08641591	0.01714897	0.09057935
NN-1	0.01911682	0.09574331	0.01596287	0.08788813
NN-2	0.01911812	0.09574787	0.01596527	0.08789708
NN-3	0.01975137	0.1016627	0.01693102	0.09571736
NN-4	0.01980422	0.1018857	0.01696239	0.09605159
NN-5	0.01208429	0.07757507	0.02130364	0.09871859
NN-6	0.01551066	0.08574288	0.02844519	0.09649167
NN-7	**0.01197029**	**0.07587229**	0.0227542	0.09952095
NN-8	0.01569115	0.08645105	0.01908987	0.09023818

A continuación se presentan los gráficos de dichos errores, donde se puede observar que si bien el comportamiento de las curvas de error sobre los datos de entrenamiento difiere del que corresponde a las curvas de error sobre los datos de test, hay modelos con error menor, es decir, que predicen mejor la variable respuesta al aplicarse sobre las 11 covariables seleccionadas.



Si se analiza la curva roja de ECMs con los datos de entrenamiento, se destacan dos modelos de redes neuronales, uno de ellos con el mínimo ECM (NN - 7) y otro que le sigue en segundo lugar (NN – 5). Una situación análoga se da en el caso de la curva roja de MAEs con los datos de entrenamiento. Por lo que entonces se elegiría como **mejor modelo** a **NN–7**. Se puede notar que no ocurre lo mismo con los errores calculados a partir de la muestra de test, donde se puede observar en los gráficos que los errores son superiores en varios de los modelos de redes neuronales, y que el mínimo error se encuentra bajo el modelo GAM, y en segunda instancia bajo el modelo **NN-1**.



Sin embargo, esto que se produce en esta muestra particular de test no significa que se reproducirá de la misma manera en otra evaluación del mismo modelo pero con otros datos. Aplicados en esta muestra de test, los modelos de redes neuronales de menor ECM bajo la muestra de entrenamiento (NN-7 y NN-5) no parecen ser los mejores, pero con otros datos sería probable que los errores sean inferiores, dado el comportamiento de ambos métodos bajo la muestra de entrenamiento comparada al comportamiento del resto de los modelos.

Para finalizar, según este análisis se podría elegir como mejor método, en el sentido de que tenga la mayor capacidad posible de generalización y de predicción, y basándose en las 11 covariables elegidas, al modelo de redes neuronales NN-7.