Is there a relation between the condition of the driver and factors such as the characteristics of the vehicle, characteristics of the driver, and date of occurrence, at the time of an accident?

¿Existe relación entre la condición del conductor y factores como las características del vehículo, características del conductor y fecha de ocurrencia, al momento de un accidente?

Luis Estuardo Delgado Ordoñez [1] , María Fernanda Estrada Cornejo [2] , Christopher Kevin Sandoval García [3]
[1,2,3] Departamento de Ciencias de la Computación, Facultad de ingeniería, Universidad del Valle de Guatemala
[1] del17187@uvg.edu.gt, [2] est14198@uvg.edu.gt, [3] san13660@uvg.edu.gt

Abstract

The use of transportation has become vital in people's lives, either it's public or personal type of transportation. This last alternative represents a great responsibility, since not only must the characteristics of the vehicle be taken into account, the characteristics of the driver or the day in which he will drive will also have to be considered. The hypothesis we will work on is that the condition of the driver at the time of an accident does have a relationship with the characteristics of his vehicle, his own characteristics - age and sex - and with the date of occurrence. To test this hypothesis, an exploratory analysis of the traffic events data from the years 2009 to 2014 was performed. Then, after determining the best algorithms that could be applied, three different models for predicting the condition of the pilot at the time of the accident were obtained based on the already mentioned factors. As a result, the three models had an accuracy greater than 63%, showing that there is a relation between these characteristics, therefore showing that the condition of the driver can be predicted.

Keywords: neural networks, SVM, Caret, PCANNet, traffic event

Resumen

El uso de un medio de transporte se ha vuelto vital en la vida de las personas, ya sea un medio público o personal. Ésta última alternativa representa una gran responsabilidad, ya que no solo se debe tener en cuenta las características del vehículo, también se deben considerar las características del

piloto o el día en que saldrá a manejar. La hipótesis entonces es que la condición del piloto al momento de un accidente sí tiene relación con las características de su vehículo, sus propias características -edad y sexo-, y con la fecha de ocurrencia del hecho de tránsito. Para comprobar esta hipótesis, se realizó un análisis exploratorio de los datos de hechos de tránsito de los años 2009 a 2014. Luego, al determinar los mejores algoritmos que podían aplicarse, se obtuvieron tres modelos distintos de predicción de la condición del piloto al momento del accidente en base a los factores mencionados. Como resultados, los tres modelos tuvieron una precisión mayor al 63%, mostrando que, al sí existir una relación de estas características, se puede predecir la condición del conductor.

Palabras clave: redes neuronales, SVM, Caret, PCANNet, hechos de tránsito

Introducción

Los hechos de tránsito en vehículos particulares dependen de muchos factores, tales como la condición del vehículo, condición del piloto, hora del día, día de la semana, género del piloto, entre otros. Por ejemplo, según el boletín No. 12-2019 de la ONSET, la hora en la que se da la mayor cantidad de hechos de tránsito es 7:00 PM. En este mismo boletín, una gráfica muestra que los días con mayor cantidad de hechos de tránsito son viernes, sábados y domingos. La población general puede intuir estos resultados por lo que ven en las calles, incluso sin ver el boletín u otros reportes. Al existir tantos factores que considerar, en el siguiente artículo se tomaron específicamente el estado y las características del conductor, características del vehículo y fecha de ocurrencia del hecho. El objetivo general fue determinar si la condición del conductor tenía relación con estos factores para poder crear un modelo de predicción de su estado (bajo efectos de alguna sustancia o no). Primero se mencionan los materiales y métodos utilizados. Luego, se discutirán los resultados obtenidos y cómo se cumplieron los objetivos. Por último, se presentan las conclusiones más importantes del proyecto, un agradecimiento especial y las referencias utilizadas.

Materiales y métodos

El contexto del problema son las calles de Guatemala, específicamente cuando se da un hecho de tránsito (colisión, choque, atropello, etc.). Las variables utilizadas son: condición del piloto (si se encontraba bajo efectos de alguna sustancia), día de la semana y hora en que ocurrió el hecho, sexo y edad del piloto, y el tipo y color del vehículo en el que se transportaba. Las herramientas utilizadas para el análisis de estos datos fueron R y RStudio. Se utilizó la base de datos[1] del INE, la cual poseía

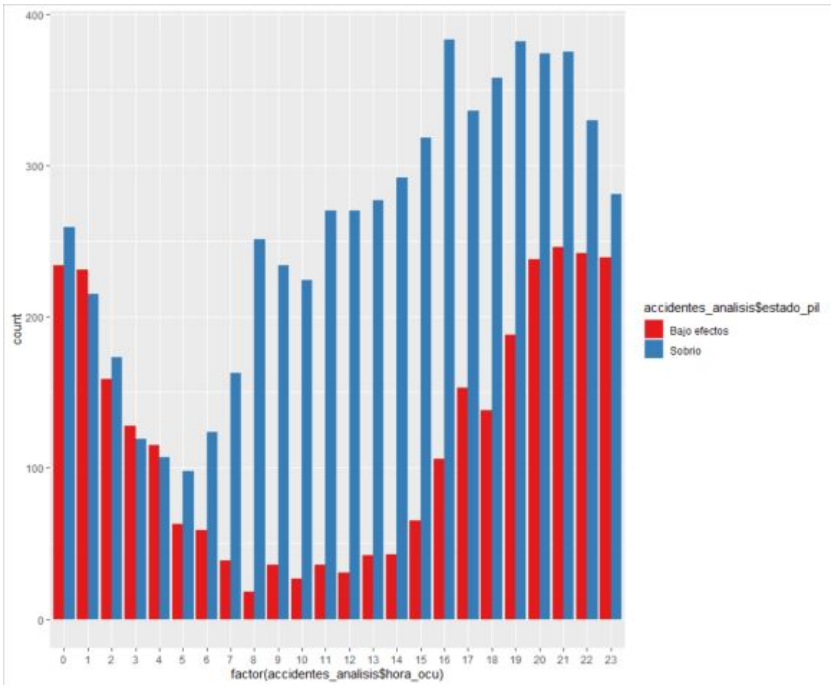
archivos .sav de los hechos de tránsito ocurridos en los años 2009-2019. Con estos datos se debe tener mucho cuidado, ya que los nombres de las columnas varían por año (aunque representen la misma información) y para los últimos años ya no se encuentra la variable que representa la condición del piloto; es por esto que para fines del artículo, solamente se utilizaron los años 2009-2014. Otros aspectos a considerar antes de trabajar con los datos, es que se deben normalizar para que se encuentren entre el rango de 0-1; solamente debe aplicarse la normalización a variables numéricas y no las categóricas. También, se debe observar que los datos que se colocaron como "vacíos" o "NA" están representados por 99 o 999, por lo que deben eliminarse antes. Por último, se debe estandarizar la hora de ocurrencia, ya que unos datos tienen formato de 0-23 horas y otros de 1-24; en este proyecto se trabajó con formato 0-23.

Antes de aplicar cualquier algoritmo de predicción, es importante saber en qué consiste cada uno. El algoritmo de la red neuronal[2] es un tipo de algoritmo de ML el cual puede ser supervisado o no-supervisado. Esta busca encontrar la respuesta más adecuada entre sus opciones de salida por medio de una serie de "capas" de neuronas las cuales son afectadas de manera secuencial hasta obtener una salida. Cada capa incide en la siguiente por medio de "pesos" los cuales sirven para deducir a qué neurona de cada capa se apega más cada capa ya evaluada. La primer capa o input son todos los datos únicos que se ingresarán en la red. La capa final es donde se evalúa qué possible output es más adecuado de acuerdo a los parámetros definidos en la red, ya sea manualmente o por algoritmos de entrenamiento. Por otro lado, las SVM[3] es un algoritmo utilizado para clasificación de clases supervisado. Es utilizado primariamente para la clasificación binaria o la separación en 2 grupos. Su objetivo es encontrar un hiperplano en cual separa los grupos de datos y utiliza sus ubicaciones para clasificar nuevos datos. Estos hiperplanos se definen en base a diversas variables en los datos. Esto es utilizado para definir qué datos pertenecen a un grupo específico o no entonces limita la cantidad de grupos a clasificar.

Entonces, se decidió utilizar redes neuronales por por su complejidad y eficiencia, y la cantidad de capas puede mejorar considerablemente la clasificación de la variable respuesta. Se decidió utilizar SVM porque permite específicamente una clasificación binaria (el piloto está bajo efectos de una sustancia o no). Además, este último algoritmo es bastante preciso y sus predicciones se hacen con rapidez.

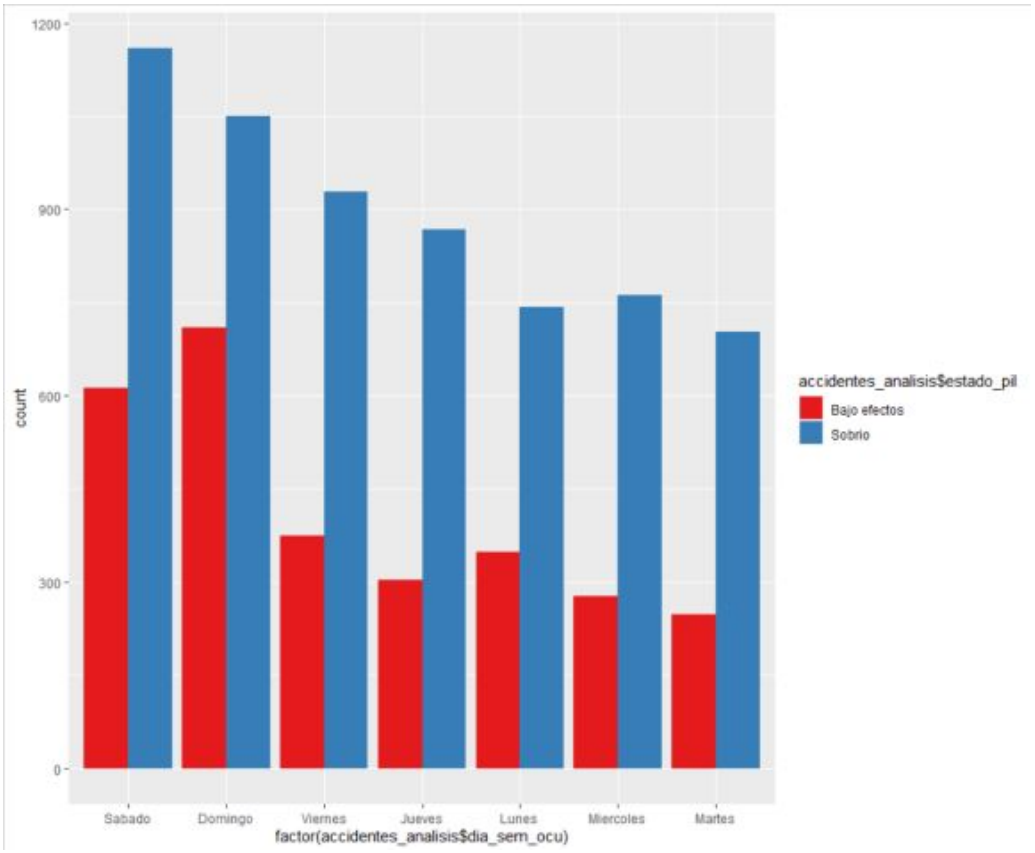
El análisis exploratorio permitió conocer más acerca de las variables y su comportamiento. Se realizó un histograma con todas las variables, pero comparadas con la variable respuesta "condición del piloto". De esta forma, se observa el comportamiento y la posible relación de variables. Primero, en la gráfica 1 se observa que el horario con mayor cantidad de accidentes con conductores bajo efecto de alguna sustancia es de 8 PM a 1AM. En cuanto a conductores sobrios, las horas con mayor cantidad de accidentes son de 4 PM a 9 PM.

Gráfica 1. Histograma horas de ocurrencia y estado del piloto.



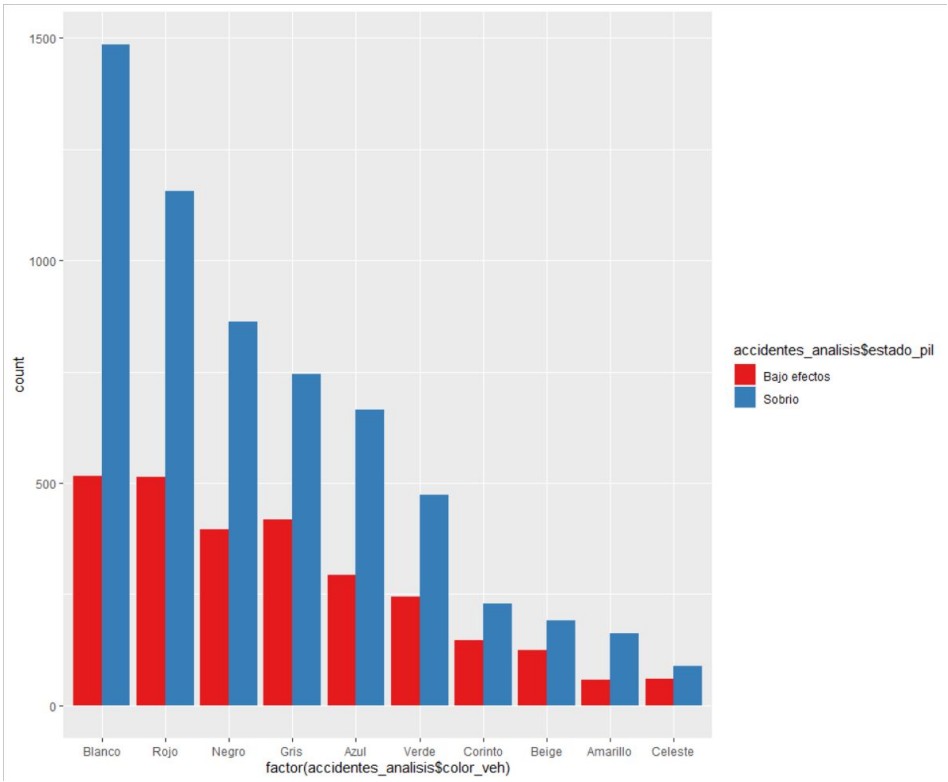
En la gráfica 2, se observa que el día con mayor cantidad de accidentes involucrando conductores sobrios es el sábado. El día con mayor cantidad de accidentes que involucran a un conductor bajo efectos de alguna sustancia es el domingo.

Gráfica 2. Histograma día de la semana del hecho y estado del piloto.



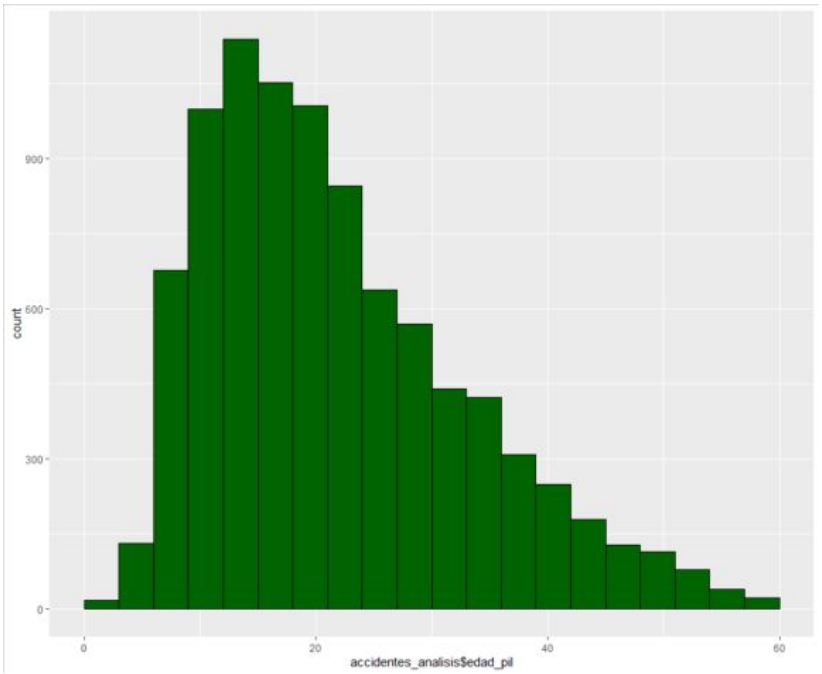
En la gráfica 3, se compara el color del vehículo con el estado del piloto. Se observa que, tanto con conductores bajo efectos de alguna sustancia como los que no, los colores con mayor cantidad de accidentes es el blanco y el rojo

Gráfica 3. Histograma color del vehículo y estado del piloto.



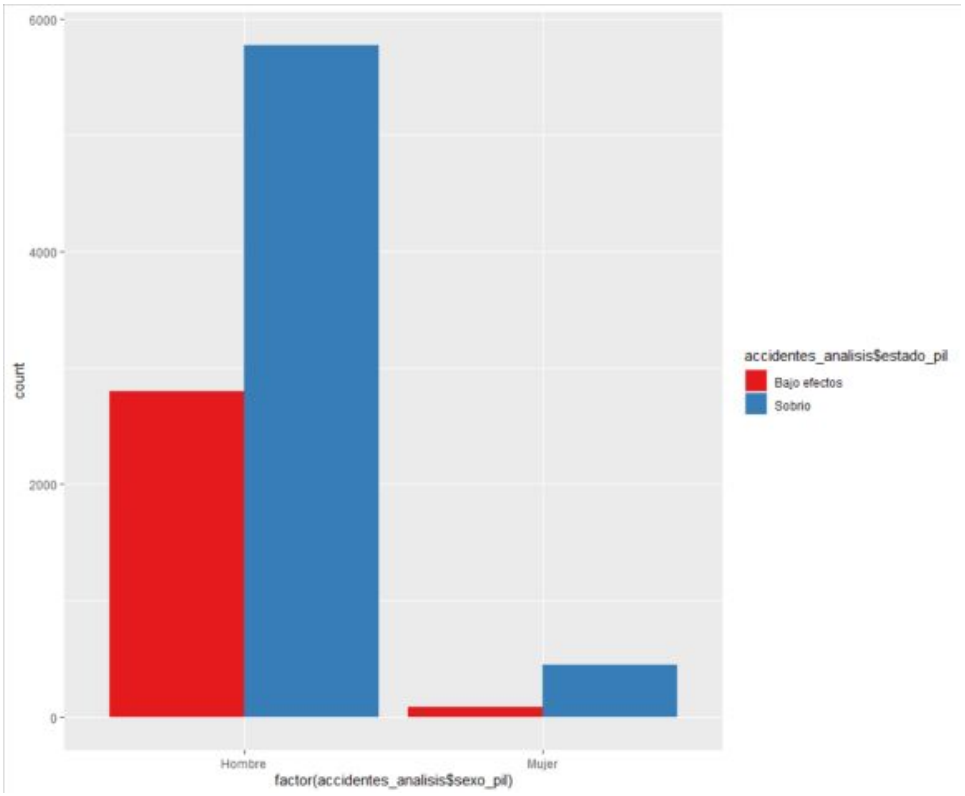
En la gráfica 4, se muestra un histograma de las edades, el cual indica que la mayor cantidad de pilotos tienen una edad entre 20-30 años.

Gráfica 4. Histograma edad del piloto.



En la gráfica 5, se observa que hay mayor cantidad de pilotos hombres (sobrios y bajo efectos) que de pilotos mujeres (sobrias y bajo efectos).

Gráfica 5. Histograma sexo del piloto.



Como se observa en todas las gráficas, la cantidad de conductores sobrios supera considerablemente a la cantidad de conductores bajo efectos de alguna sustancia. Es por eso que se debe tener cuidado en la distribución del set de entrenamiento y set de test, ya que la desigualdad podría resultar en modelos incorrectos.

Resultados y discusión

Ya con todos los datos juntos, se decidió que el set de datos sería del año 2009-2013, y el año 2014 como test. Se procuró que ambos sets estuvieran balanceados. Las variables que se tomaron en cuenta para predecir el estado del conductor fueron día de la semana, hora de ocurrencia, sexo y edad del piloto, tipo y color del vehículo. Se utilizaron los datos transformados para generar modelos de redes neuronales con Caret (en figura 1), redes neuronales con PCANNet (figura 2) y support vector machines (figura 3).

Figura 1. Modelo de redes neuronales utilizando paquete Caret.

```
> modeloCaret
Neural Network

5247 samples
   6 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 5247, 5247, 5247, 5247, 5247, 5247, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0e+00  0.5870879  0.1521243
  1     1e-04  0.5972358  0.1696290
  1     1e-01  0.6029742  0.1715781
  3     0e+00  0.6218592  0.2314694
  3     1e-04  0.6232986  0.2371078
  3     1e-01  0.6495046  0.2814021
  5     0e+00  0.6249130  0.2359392
  5     1e-04  0.6297215  0.2471321
  5     1e-01  0.6515465  0.2856175

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 5 and decay = 0.1.
```

Figura 2. Modelo de redes neuronales utilizando paquete PCANNet.

```
> modeloCaretPCANNet
Neural Networks with Feature Extraction

5247 samples
   6 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 5247, 5247, 5247, 5247, 5247, 5247, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0e+00  0.5974872  0.1890315
  1     1e-04  0.6004846  0.1955654
  1     1e-01  0.6053958  0.1906227
  3     0e+00  0.6045070  0.2002140
  3     1e-04  0.6078887  0.1928804
  3     1e-01  0.6141905  0.2081362
  5     0e+00  0.6143975  0.2080963
  5     1e-04  0.6127904  0.2024479
  5     1e-01  0.6343525  0.2476310

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 5 and decay = 0.1.
```

Figura 3. Modelo de SVM.

```
> modeloSVM

Call:
svm(formula = estado_pil ~ sexo_pil + dia_sem_ocu + edad_pil + color_veh + tipo_veh, data = accidentes_train)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  4423
```

Para determinar qué tan precisa fue la predicción de estos modelos, se realizaron sus matrices confusión correspondientes. Las matrices de confusión comparan los datos teóricos (dataset de test)

con los datos de la predicción, para obtener el porcentaje de accuracy al que se llegó. La matriz de confusión del modelo de redes neuronales con Caret, la del modelo de redes neuronales con Pcannet y la del modelo con SVM se muestran en las figuras 4 a 6, respectivamente.

Figura 4. Matriz de confusión de redes neuronales utilizando paquete Caret.

```
> cfmCaret
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2295  265
         1  913  361

               Accuracy : 0.6927
                 95% CI : (0.6779, 0.7073)
    No Information Rate : 0.8367
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2062

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7154
            Specificity : 0.5767
         Pos Pred Value : 0.8965
         Neg Pred Value : 0.2834
             Prevalence : 0.8367
         Detection Rate : 0.5986
   Detection Prevalence : 0.6677
      Balanced Accuracy : 0.6460
```

Figura 5. Matriz de confusión de redes neuronales utilizando paquete PCANNet.

```
> cfmCaretPCANNet
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2237  266
         1  971  360

               Accuracy : 0.6774
                 95% CI : (0.6623, 0.6921)
    No Information Rate : 0.8367
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1874

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6973
            Specificity : 0.5751
         Pos Pred Value : 0.8937
         Neg Pred Value : 0.2705
             Prevalence : 0.8367
         Detection Rate : 0.5835
   Detection Prevalence : 0.6528
      Balanced Accuracy : 0.6362
```

Figura 6. Matriz de confusión de SVM.

```
> cfmSVM
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2710  467
         1  498  159

               Accuracy : 0.7483
                 95% CI : (0.7343, 0.762)
    No Information Rate : 0.8367
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0968

 Mcnemar's Test P-Value : 0.3342

            Sensitivity : 0.8448
            Specificity : 0.2540
         Pos Pred Value : 0.8530
         Neg Pred Value : 0.2420
             Prevalence : 0.8367
         Detection Rate : 0.7068
   Detection Prevalence : 0.8286
      Balanced Accuracy : 0.5494
```

Al ejecutar los algoritmos en el set de datos la primera vez, se obtuvo que siempre predecía que el conductor iba sobrio. Se investigó qué podría ser, ya que el set de training sí contenía datos de conductores sobrios. El problema era que habían demasiados conductores sobrios en el set de datos de training (aprox. 1% contra 99%). Entonces, se procedió a equilibrar mejor entre conductores bajo efectos o no.

Ya con ambos lados equilibrados, se ejecutaron nuevamente los algoritmos y éstos ya lograban predecir ambos casos. En el modelo de redes neuronales usando Caret se obtuvó un accuracy del 69.3%, con más errores al predecir que el conductor iba bajo efectos de una sustancia cuando no era así. En el modelo de redes neuronales usando PCANNet se obtuvó un accuracy del 67.8%, igualmente con más errores al predecir que el conductor iba bajo efectos de una sustancia cuando no era así. En el modelo de SVM se obtuvó un accuracy del 74.8%, igualmente con más errores al predecir que el conductor iba bajo efectos de una sustancia cuando no era así.

Analizando los tres casos en general, el mayor error que tuvieron fue clasificar el estado del conductor como bajo efectos de una sustancia cuando no era así. Aunque es un error grave, es mejor aclararlo que no notar que sí lo está y dejarlo ir cuando podría ocasionar otro accidente. En conclusión el mejor modelo fue el de SVM, ya que tuvo el mejor accuracy y fue el más rápido en ejecutar.

Conclusiones

Luego de la ejecución de los algoritmos de entrenamiento y prueba, y analizando los resultados se puede decir que se puede predecir, con una precisión moderada, si un conductor estaba en estado de ebriedad de acuerdo con su edad, sexo, hora de axidente, dia de la semana donde ocurrió el accidente, tipo de vehículo que manejaba y color del mismo. El porcentaje de precisión puede haberse visto reducido por el hecho que se cuenta con un muy pequeño porcentaje de accidentes en los que se reportó el estado del conductor y este estaba ebrio.

El algoritmo de support vector machines es el algoritmo que tuvo mayor tasa de éxito general, mas fue el algoritmo con menor cantidad de éxitos en predecir conductores ebrios. Mientras que el modelo de redes neuronales en caret fue el más exitoso de los dos algoritmos de redes neuronales ambos algoritmos predecían que el conductor estaba ebrio más que el algoritmo de SVM aún cuando dicha predicción era errónea.

La investigación se realizó con un número muy reducido de casos los cuales a su misma vez estaban desbalanceados con una sobre-representación de casos donde el conductor se encontraba sobrio. Esto llevó a una mayor reducción de los casos para balancear los tipos de casos por lo cual se usaron muestras relativamente pequeñas para los algoritmos de entrenamiento.

Agradecimiento

Un especial agradecimiento a la profesora de Minería de Datos, Lynette García Pérez, por haber descrito y explicado los algoritmos de predicción utilizados en este artículo. También se le agradece el haber proporcionado códigos de ejemplo y presentaciones de dichos temas, así como los documentos de creación de artículos.

Referencias y bibliografía

[1] Instituto Nacional de Estadística de Guatemala. Accidentes de Tránsito. Consultado de la página web https://www.ine.gob.gt/ine/estadisticas/bases-de-datos/accidentes-de-transito/

[2] Sanderson, G. 2017. Backpropagation calculus | Deep learning, chapter 4. Consultado de la página web https://www.3blue1brown.com/neural-networks

[3] Manning, C. Raghavan, P. Schütze, M. 2008. Introduction to Information Retrieval. Consultado de la página web https://nlp.stanford.edu/IR-book/

Deng, H.; Miao D.; Lei, J;Lee Wang, J. 2011. Artificial Intelligence and Computational Intelligence. 1era edición. Springer Science & Business Media. China. 718 pp.

Graupe, D. 2019. Principles Of Artificial Neural Networks: Basic Designs To Deep Learning. 4ta edición. World Scientific. 440 pp.