Estudi de les comunicacions del bus de dades d'un vehicle

Raül Mola Escudero

Resum— Avui en día tots el vehicles moderns estan controlats per varis processadors integrats que reben el nom de ECU(Electronic Control Units), els quals es comuniquen mitjançant el bus de dades CAN(Controller Area Network), aquest ademés esta conectat a tots els sensors i dispositius del vehicle.
Per accedir a l'informació d'aquest aquest bus, els vehicles disposen del protocol OBD(On-Board Diagnostics), que ens permet entendre el tràfic de dades que circula per el bus CAN. Així podem detectar o diagnosticar errors al sistema d'un cotxe que potser a simple vista no son apreciables.
La finalitat d'aquest projecte es estudiar el bus CAN i el protocol OBD2 amb el qual es llegueix aquest i donar una possible solució utilitzant una red neuronal per evitar atacs que pot rebre aquest.

Paraules clau— Bus CAN , OBD, OBD2 , vehicle , Red Neuronal, ECU.

Abstract— Nowadays, every modern car have controled with many integrated microprocessors which receive the name of ECU(Electronic Control Units), this microprocessors can comunicate between them thank you to data CAN(Controller Area Network) bus which connect all the sensors and devices of the vehicle.
For see the information that are in the CAN bus, the vehicles have the OBD(On-Board Diagnostics)protocol which allow understand the data traffic in the bus. Because of this is more easy detect errors in the car System that maybe it could be hard to see.
The finallity of this project is go deep in all of this parts of the car for understand better how they works and investigate about what atacs could violate the integrity of the system to final create a neural network able to predict this atacs in the bus.

Keywords — CAN bus ,OBD, OBD2 , vehicle , Neural. Network, ECU

1 Introducció

La tecnologia al mon automovilistic a crescut molt més del que ens pesem, de manera que fins i tot als vehicles més moderns trobem microprocessadors integrats per tot el cotxe els quals controlen un o més d'un dispositiu electronic, aquest microprocessadors reben el nom de ECU i es comuniquen entre si amb el bus de dades CAN.
Per aquest bus pasen dades de tots els dispositius i parts del cotxe desde una finestra fins al motor del vehicle, fet que ho fa molt atractiu per a possibles atacs. Aquest sistema no s'ha estudiat gaire i de fet les companyies punteres fabricants d'automovils no tenen en compte la seguretat informàtica en els seus vehicles en quant a possibles atacs exteriors, motiu que encara fa més atractiu el fet de poder controlar un vehicle o cents d'ells ja no disposen de cap protecció. Per això en aquest article parlem sobre que és el bus de dades CAN, el protocol OBD2 amb el qual llegim les dades que van per aquest bus, els microprocessadors (ECU) i com tot aquest sistema funciona i les seves propietats. Ademés veurem com amb una red neuronal trobem una possible solució a evitar atacs i com hem arribat a aquesta conclusió a través del estudi del bus de dades del vehicle i les seves caracteristiques.
Dir que per la realitació d'aquest projecte s'ha utilitzat la metodologia Kanban, mètode que esforça visual i ajuda a orientar-se amb un cop d'ull al tauler d'organització.

- *E-mail de contacte: raul.mola@e-campus.uab.cat*
- *Menció realitzada: Enginyeria de Computació*
- *Treball tutoritzat per: Jordi Serra (CVC)*
- *Curs 2019/20*

2 Bus CAN

En aquest apartat presentarem el bus CAN(Controller Area Network), el qual permet als dispositius i microcontroladors comunicarse entre ells mitjantçant un protocol sòlid de baix cost basat en missatges. Veurem el seu origen, la motivació perque va ser creat i els beneficis del seu ús.

2.1 Historia

A mesura que els vehicles i la tecnologia en ells han anat evolucionan, s'han anat controlant més paràmetres d'aquest per tal de que els vehicles fossin més confortables a l'hora de conduir, més segurs , més eficients, etc. Per això sorgeix la necessitat de tindre una unitat de control a les diferents parts del coche, lo que al mateix temps sorgeig un altre problema per conectar tots aquells mòduls per a que es comuniquen de forma individual. D'aquesta necessitat sorgeig la tecnologia CAN que va ser desarrollada per Robert Bosch GmbH i presentada juntament amb intel al 1985. L'objectiu original d'aquesta tecnologia era reduir la cantitat de cables als vehicles degut a que en alguns casos había arribat a necessitarse fins a 2 kilometres de cables en un d'aquest i reduïa el pes dels cotxes fins a 45 kilograms. Per això desde 1993 l'industria automotriu va adaptar ràpidament CAN i es va convertir en un estadart internacional conegut com ISO 11898.
Més endavant aquest estàndard afavoreixeria l'implantació del protocol OBD.

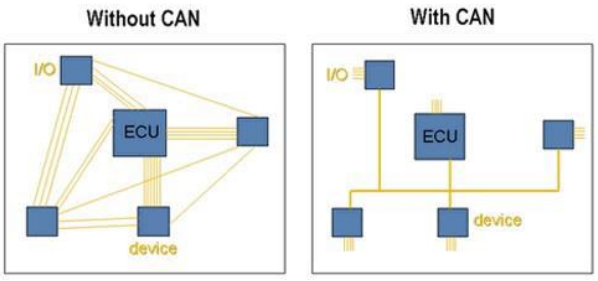

Fig.1: Comparativa conexió amb CAN i sense CAN.

2.2 Avantatges de CAN

Les principals avantatges del bus CAN ón:

1. Baix cost : el dispositius es comuniquen mitjançant una única interfície CAN, no mitjançant senyals analògiques directes, fet que redueix el cost i els errors del cablejat.
2. Centralitzat : el sistema de bus CAN permet la configuració central d'errors en tots els dispositius.
3. Sòlid : el sistema compta amb una sòlida capa física, dissenyada per ambients amb molt de soroll. De tal forma que garanteix la consistència de les dades.
4. Eficiencia : els missatges CAN es prioritzen mitjançant l'us de ID, de manera que els ID amb major prioritat no s'iterrumpeixen.
5. Flexible : cada dispositiu compte un chip per rebre tots els missatges transmesos per decidir la rellevància i actuar seegons com correspongui. Flexibilitat en la disposició dels mateixos, permet afegir o treure de forma dinàmica. Poden conectar-se un màxim de 110 nodes a una red CAN.

2.3 Arquitectura de CAN

L'arquitectura de CAN esta composta per 2 elements els quals són:

- Bus : per a aconseguir la tipologia de bus sutilitzen dos cables trençats amb una impendancia de 120Ω, aquest uneix totes les unitats de control del sistema i serà per on circuli l'informació. Aquesta informació es transmeteix per diferencia de tensió entre els dos cables, de manera que un valor molt alt de tensió representa un 1 i un valor de baixa tensió representa un 0. La combinació adecuada de uns i ceros formen en missatge a transmetre.
  A un cable els valors de tensió oscil·len entre el 0V i 2,25V per això es denomina cable L(low), mentres que a l'altre cable H (high) els valors oscil·len entre 2,75V i 5V.

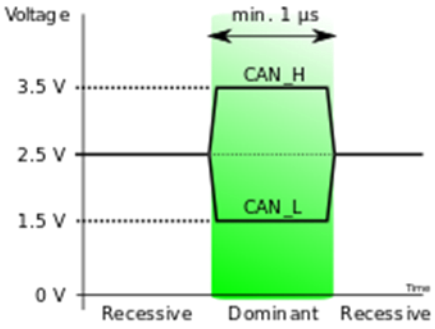

Fig.2: Nivells de tensión del bus CAN

- Nodes: els cuals están formats per dos elements:
  - Controlador: és l'ement encarregat de la comunicació entre el microprocesador de la unitat de control y el transmisor-receptor. Gestiona les trames CAN, comproba errors en la transmisió entre altres nodes, detecció de col·lissions.
    El controlador esta situat a la unitat de control, per això existeixen tants com unitats hi haguin al sitema. Aquest element treballa amb nivells de tensió molt baixos i es el que determina la velocitat de transmisió dels missatges, la qual serà mes o menys elevada segons el compromís del sistema.
  - Transmissor/Receptor: Aquest mòdul es l'encarregat de la codificació i decodificació dels missatges al bus, sincronització i control dels nivells de la senyal o control d'acces al medi.
    El transmissor-receptor es bàsicament un circuit integrat que esta situat a cada unitat de control pertanyent al sistema, treballa amb instensistats pròximes a 0.5 A i en ningún cas intervé modificant el contingut del missatge. Funcionalment esta situat entre els cables que formen el bus CAN y el controlador, tal i com podem observar a la Fig.3 ques està acontinuació.

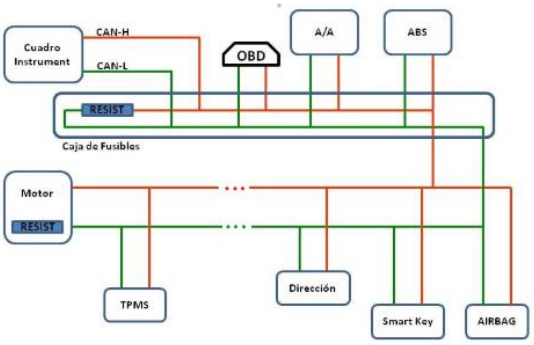

Fig.3: Topologia de la red CAN i els seus nodes.

Comentar que també existeixen una variació del sistema d'organització de l'estructura del bus CAN on tots aquest nodes estan connectats a un node central anomenat Gateway que fa d'enllaç entre els bus CAN de diferents velocitats amb el conector OBDII.

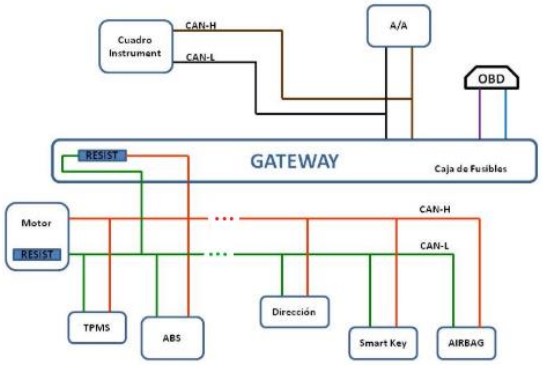

Fig.4: Topologia de la red CAN i els seus nodes amb Gateway.

2.4 Funcionament de CAN

El procés de transmissió de les dades es desarrolla seguint una sèrie de fases les quals son les següents:

- Subministament de dades:
  Una unitat de comandament rep informació dels sensors que te associats com per exemple : revolucions del motor, velocitat, temperatura del motor, porta oberta, etc. Llavors els microprocessadors que tenen incorporat pasa l'informarció al controlador on es gestionada i acondicionada además de ser pasada al transmissor-receptor on es transforma en senyals elèctriques.

- Transmisió de les dades:
  El controlador propi de cada unitat transfereix les dades i el seu identificador juntament amb la petició d'inici de transmissió, assumint la responsabilitat de que el missatge sigui correctament transmès a totes les unitats de comandament associades. Per transmetre el missatge ha tingut que trobar el bus lliure, en cas de col·lisió amb una altra unitat de comandament intentant transmetre simultàniament, tindre un prioritat més gran. A partir del moment en que això pasa, la resta d'unitats de comandament es converteixen en receptores.

- Recepció del missatge:
  Un cop totes les unitats de comandament reben el missatge, verifiquen el identificador per determinar si el missatge serà fet servir per elles. Aquestes unitats necessiten les dades del missatge que processen, si no ho necessiten el missatge es ignorat.

2.5 Trama de CAN

La trama de CAN esta formada per diferents camps que informen de diferents aspectes del missatge. Aquestes trames utilitzen bit stuffing, es a dir, quan succeixen 5 bits iguals s'introdueix un bit extra de valor contrari per evitar la desincronització. Existeixen dos tipus de trames l'estàndard i l'extesa en base al número de bits de l'identifcador.

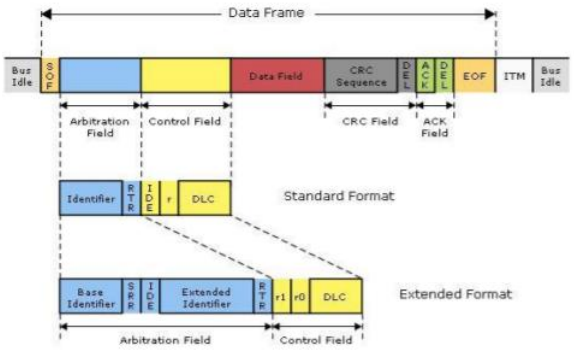

Fig.5: Trama del bus CAN i els seus tipus.

Aconinuació s'explica les diferents camps que conformen la trama en l'ordre en que aparaeixen en aquesta:

- SOF (Start of Frame bit): marca l'inici del missatge i permet la sincronització dels nodes conectats a la red.

- Arbitration Field / Camp d'arbitratge:
  esta format per 12 bits en el cas de la trama estàndard o 32 bits en el cas de la trama extesa. Dins d'aquest camp es troba l'dentificador, el qual indica la prioritat del node. El node amb mayor pioritat es aquell que té l'identificador més baix. El bit RTR que s'utilitza per distingir entre trama remota o trama de dades.

- Control Field / Camp de control: Format per 6 bits. El bit IDE indica amb un estat dominant que la trama enviada és estàndard i el bit RB0 esta reservat i s'estableix en estat dominant perel protocol CAN.

  Per el que fa Data Length Code(DLC) indica el

nombre de bytes de dades que conte el missatge. La trama extesa te el bit adicional RB1.

- Data Length Code / Camp de dades:
  pot estar format fins per a 8 bytes, depenent del que especifique el DLC, compte les dades del misatge.

- CRC Field / Camp de verificació per redundància cíclica:
  Camp de 15 bits, detecta errors en la transmisió del missatge, es delimita amb un bit final en estat recesiu.

- ACK Field / Camp de reconeixement:
  El últim camp de la trama, está format per 2 bits. El node ttransistor mana una trama amb el bit ACK (acknowledge) es estat reseciu, mentres que els receptor si reben el missatge correctament, manaran un missatge es estat dominant.

- EOF / Camp de fi de trama :
  Conjunt de 7 bits que indiquen el fi de trama.

3 Protocol OBD

En aquest apartat presentarem el procol OBD (On Board Diagnostics) que és un sistema a bord dels vehicles(cotxes i camions). Veurem el seu origen, la motivació perque va ser creat i els beneficis del seu ús.

3.1 Historia

Als anys 50 la mayor part del descobriment de problemes en els vehicles es realitzava utilitzant pocs mesuradors i una gran part es feia escoltan el soroll del motor, olors estranys, o d'alguna altra manera utilitzant el sentits per detectar problemes.

A mesura que la tecnologia va anar millorant i els vehicles es van tornar més complexos, amb l'interació de l'ectronica al sistema, els ordinadors abord es fan més accessibles això permet que altres mesures com el buit, la presio de l'oli i algunes temperatures puguin ser integrades a un sistema de diagnòstic.

La primera generació de On Board Diagnostics (OBDI) , va ser desenvolupada al estat de California, Estats Units d'Amèrica , per la Junta de Recursos de l'Aire de California (California Air Resources Board (CARB)) e implementada al 1988 per controlar algunes emissions dels components del vehicle, però no va ser fins al 1994 impulsat per alertes smog a Los Angeles que CARB va requerir que tots el cotxes de 1996 en andavant portessin equipats OBDII.

Respecte a Europa aquest estàndard no va entrar en vigor fins al 13 d'Octubre de 1998. Pero a diferencia de EEUU s'utilitza el protocol EOBD que és una variació del OBDII. Aquest protocol es comença a aplicar a tots els cotxes gaslina desde l'any 2000 I diesel desde el 2003.

3.2 OBDII

El sistema OBDII és un sistema de diagnòstic, la seva funció primordial es la monitorització es per això que la gran part d'eines que funcionen per a OBDII estan enfocades en aquest sentit.

OBDII va suposar una gran revolució, ja que ademès de les funcions que ja impartia OBDI, es van afegir una gran varietat de noves aplicacions que permetien un control pràcticament total del motor, capacitat de monitorització del chasis i altres funcions de control dels sistemes electrònics, el conjunt d'aquestes millores va suposar un salt de millora en el sector automovilistic i va suposar que aquest sistema de diagnòstic es col·loques a la vanguardia tecnològica.

Cal comentar que avui en dia aquest sistema no s'utilitza tant per el que va ser disenyat en un principi, sino que es fa server més a tallers, on dona un gran support i facilita el treball als mecànics.

3.3 Funcionament OBDII

El sistema OBDII esta dissenyat per controlar les emissions dels sistemes de control i alguns components específics del vehicle. Existeixen 3 registres que ajuden a detectar averies

1. Indicador d'averia (MIL) :
   Quan es detecta un error, el sistema OBDII encén un indicador lluminós d'avís d'averia (MIL) situat al panell d'intruments del vehicle i d'aquesta forma avisar al conductor mitjançant la frase Check Engine o Service Engine Soon.

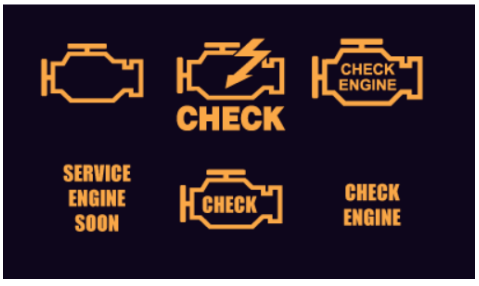

Fig.6: Tipus d'indetificador lluminics MIL

2. Codis d'error (DTCs):
   Els codis d'error de diagnosis son emmagatzemats per el sistema OBDII en resposta a un problema trobat al vehicle. Aquets codis estan formats per 5 digits alfanumerics, el primer caracter sempre es una lletra que identifica el sistema de control on s¡ha produit l'error. Els cuatre digits restants son números, que proporcionen informació adicional sobre el DTC que s'ha originat i les condicions de funcionament que l'han causat.

Fig.7: Esquema DTCs

3. Estat dels monitors:
   Una part important dels vehicles amb sistema OBDII son els monitors de lectura, els quals son indicadors que s'utlitzen per trobar les emissions dels components evaluats per el sistema OBDII. Aquets funcionen per una comprovació periodia en sistemes i components específics per assegurarse que estan treballan sota valors normals.

   Actualment hi han 11 monitors de lectura del OBDII definits per la Agencia de Protecció Mediambiental (EPA). Aquets es diferencien en dos tipus:

   - Monitors continus: format per aquells components o sistemes que son comprovats/monotoritzats contínuament per el sistema OBDII dels vehicles, entre els que podem trobar fallada d'arrancada, sistema de combustible, Components detallats (CCM)

   - Monitors no-continus: son aquells components que necessiten que el vehicle compleixi une condicions de funcionament abans que el monitor pugui funcionar. Aquets monitors son el sistema EGR, sensors d'O2, Catalitzador, Sistema d'evaporació, calentador del sensor d'O2, Injecció d'aire secundari, catalitzador calefactat, sistema d'aire condicionat.

3.4 Missatges OBDII

En termes simplificats un missatge OBDII es composa d'un identificador i dades. Aquestes dades es dividiexen en número de bytes, mode , pid i bytes de dades Ah,Bh,Ch,Dh.

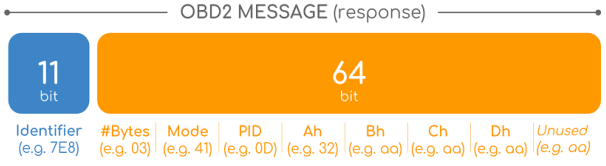

Fig.8: Estructura missatge OBDII.

- Identificador: per als missatges OBD2, l'identificador es estàndard d'once bits i s'utilitza per distingir entre missatges de sol·licitud i missatges de resposta, ademés indiquen el lloc del vehicle del qual prove.

- Longitud: A la Fig.9 surt representat com #Bytes, això fa referència a la longitud en número de bytes de les dades restants.

- Mode: per solicituts el valor és de 01-0A. Mentres que per a respostes, el 0 es remplaçat per 4 es a dir 41-4A. Hi ha 10 modes com es descriu al estàndard SAE J1979 OBDII.
  Els modes son els següents:

  - 0x01: Mostra les dades actuals.
  - 0x02: Mostra el frame de dades congelat.
  - 0x03: Mostraels codis d'error de diagnosis emmagatzemats.
  - 0x04: Neteja els codis d'error de diagnosis i els valors emmagatzetmats.
  - 0x05: Resultats test, monitorejament del sesor d'oxigen.
  - 0x06: Resultats test, altres components/sistema de monitorejament.
  - 0x07: Mostra de codis d'error de diagnosis pendents.
  - 0x08: Control d'operacions en component/sistema.
  - 0x09: Sol·licitar información del vehícle.
  - 0x0A: DTC's permanents.

- PID: per cada mode existeix una llista de PID OBDII estàndard, per exemple en el mode 01 PID 0D es la velocitat del vehicle. Per consultat la llista completa es pot veure al seguent link: https://en.wikipedia.org/wiki/OBD-II_PIDs .

- Ah, Bh, Ch, Dh: Aquets son bytes de dades en hexadecimal que es tenen que convertir a decimal abans d'utilitzar-se en calculs de la fórmula de PID.

4 Treball Realitzat

En aquesta part explicaré el treball realitzat amb les bases de dades proporcionades. I com s'ha arribat a la conclusió d'utilitzar una red neuronal.

4.1 Analisis Exploratori de Dades (EDA)

Per aquest projecte se'ns proporciona una base de dades en la qual el primer que es farà és un anàlisis de les dades, per tal de compendre millor aquestes dades proporcionades. El analisis estarà estructurat de la següent forma:

1. Comprendre el problema
2. Estudi univariable
3. Estudi multivariable
4. Neteja básica de les dades
5. Comprobació de suposicions

Per poder manipular millor les dades es pasa a un archiu les dades anomenat "tr.csv" el cual queda estructurat amb 11 atributs i un total de 39483 casos, que serán la cuantitat de trames de la base de dades. Acontinuació s'explica cada un d'aquets atributs que es poden trobar al arxiu.

- Tiempo: es el temps de lectura de cada trama, el cual está en milisegons.

- tiempoR: és el temps que tarda en llegir entre trama i trama.

- Pid (identificBador) : permet diferenciar de quin sector del cotxe provee la trama (motor, acelerador , frens, porta …)

- Nb (número de bytes): informa del número de bytes que contendrá la trama de dades.

- Trama de dades: es desglosa en 8 parts les quals reven el nom de dt1, dt2, dt3, dt4, dt5, dt6, dt7, dt8.

Comentar que tant el atribut pid com els atributs que formen la part de dades es pasen de hexadecimal a decimal ja que així en resultara molt més efectiu treballar en python que es el llenguatge que utilitzarem.

A continuació es començarà a explicar l'anilisis fet i els resultats obtinguts.
La capacita de poder predir que quina zona proveirà la seguent trama de dades del bus can, resulta particularment útil com a suport sobre en que fundamentar les decisicions en quant a la prevenció de possibles fallades d'alguns del elements del cotxe o avui en dia com s'ha comentat anteriorment que el vehicles puguin ser hackejats. Per aquest motiu, el atribut objectiu de la base de dades serà l'identificador (pid), ja que és l'atribut que es pretén predir.

Un cop decidit això el primer pas que fem es veure la distribució de les dades,això ho podem fer amb un matriu de dispersió, com la següent.

Fig.9: Matriu de dispersió

Com es pot observar les dades estan molt sesgades cap al 0. Això te dos explicacions, en el cas de les gràfiques de la primera columna i la primera fila la majoria de dades estan cap al 0 ja que una gran part de les trames provenen del motor i aquest te identificador 0. En el cas de la base de dades no es trobara un identificador amb 0 ja que aquest correspondria a les dades en hexadecimal però tots els identificadors am un valor menor que 180 corresponen a identificadors 0. Per veure la millor la relació es pot consultar la taula de canvi de base al annex.
L'altre motiu per la resta de de files i columnes es que al utilitzar la funció fillna per eliminar valors nan, elnombre de dades en valor 0 augnmenta. Però aquest agument no es tan significatiu per a que doni problemes a l'hora de fer l'anàlisis.
Aquesta visualització de les dades no ens aporta gran claritat ja que resulta difícil veure a simple vista que pot ser un outlier o no i tampoc s'intueix ningún tipus de distribució. La qual ens podria ajudar a escollir un mètode de regressió per predir els identificadors, però el que si podem observar es que hi han molts pocs punts per a la quantitat de dades que disposa la base de dades, això es degut a que els identificadors amb un valor solen tindre la mateixa trama amb valors indetics.

Com els atributs amb més correlació tenen amb el atribut a predir PID són els pertinents a la trama de dades que van des de el dt1 fins al dt8 descartarem la resta de atributs per fer aquets estudi.

Ara veiem una sèrie de visualitzacions entre els diferents atributs i el atribut a predir per veure com es relacionen.

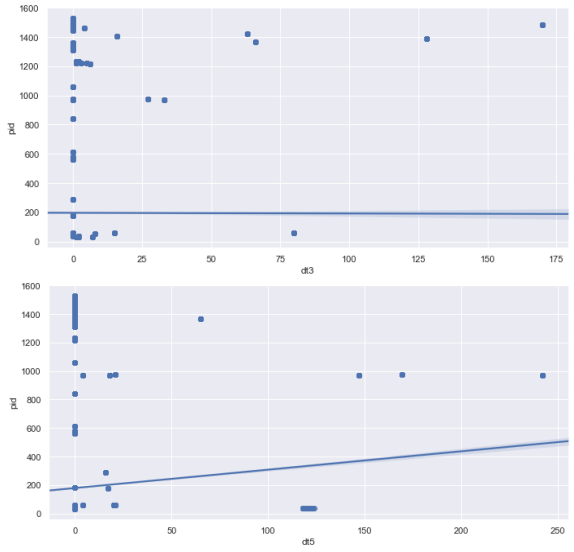

Fig.10: Gràfiques de regresió lineal dt3 i dt5

Només es mostren les dos gràfiques més representatives, en les quals podem veure com la regressió es desvia, tant per al atribus dt3 com per el dt5. Por lo que podemos afirmar que aquets atributs tenen algun valor que fa que la regresió lineal es desviï. Però aquets no es poden veure només oservant aquestes gràfiques.

Utilitarem diagramesde caixes i bigotis, els quals ens ajudaran a veure com estan distribuïdes les nostres de dades de forma gràfica. Aquesta representació funciona de la següent forma, entre el percentil 25 al puntatje més baix trobem el 25% de les dades, del percentil 25 al percentil 75 trobem el 50% de les dades i la linea que divideix aquesta caixa representa la mitjana. Per útim esta el percentil 75 i el puntatje més alt trobme el 25%de les dades restants. Aquessta explicació la trobem resumida a la següent imatge.
Comentar que aquest diagrama nosaltres l'utilitzarem en horitzontal però funciona totalment igual.

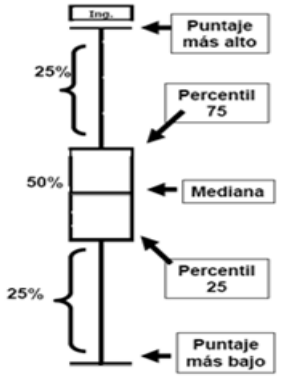

Fig.11: Parts diagrama de caixes i bigotis

Ara veurem per cada atribut quins outliers podem identificar.

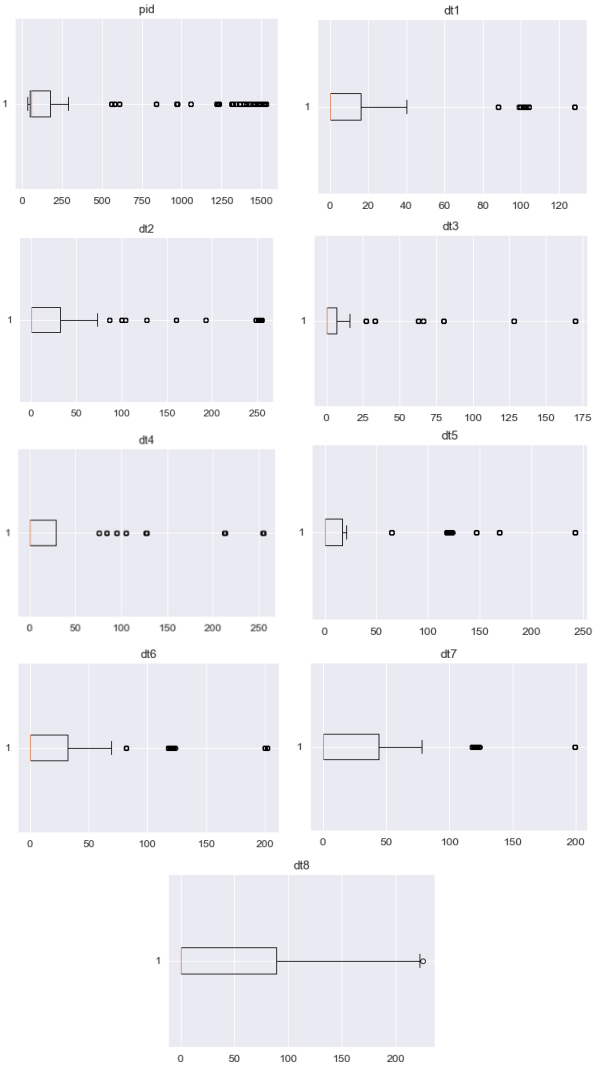

Fig.12: Resultats diagrama de caixes i bigotis

Dels resultats de la Fig.13 podem observar que tots els atributs tenen outlies, menys dt8 que pràcticament no en té.
Un cop trobat tots aquest putliers es va pensar en eliminar tots ells però el dataset quedava molt reduït, pasaba de 32984 elements a poc més de 5000. Per això sol s'eliminen outliers d'alguns atributs. Els atributs seleccionats van ser els que tenien més correlació amb l'atribut PID.
Això s'obte amb una sencilla consulta que és la següent

```
corr = data.corr()
corr[['pid']].sort_values(by = 'pid',ascending = False).style.background_gradient()
```

Amb la qual obtenim una taula ordenada dels atributs amb més correlació amb l'atribut a predir pid, que és la següent.

	pid
pid	1
dt5	0.15489
dt1	0.00162352
tiempo	0.000192611
dt3	-0.00332859
dt6	-0.0408608
nb	-0.0777812
dt8	-0.0810759
dt4	-0.0988937
tiempoR	-0.124083
dt2	-0.128004
dt7	-0.187384

Fig.13: Correlació dels atributs amb l'atribut PID.

Un cop obtinguda aquesta informació sol agafarem els tres primers atributs amb més correlació dt5,dt1 i dt3 ja que l'atribut temps l'em descartat anteriorment per aquest estudi. A continuació el que farem serà eliminar els outliers d'aquest 3 atributs. Per això calcularem el primer quartil Q1 = data[i].quantil(0.25), el tercer quartil Q3 = data[i].quantile(0.75), el Rang Interquartil IQR = Q3 - Q1 i un cop calculat això calcularem els bigotis inferiors BInferior = (Q1 - 1.5 * IQR) i el superior BSuperior = (Q1 + 1.5 * IQR). Aquest valors es guarden en una llista per cada atribut i es procedeix a eliminar els outliers de la següent forma:

```
sin_out = (((data['pid'] >= bigotes_pid[0]) & (data['pid'] <= bigotes_pid[1])) & ((data['dt1'] >= bigotes_dt1[0]) & (data['dt1'] <= bigotes_dt1[1])) & ((data['dt3'] >= bigotes_dt3[0]) & (data['dt3'] <= bigotes_dt3[1])) & ((data['dt5'] >= bigotes_dt5[0]) & (data['dt5'] <= bigotes_dt5[1])))
sin_outliers = data[sin_out]
```

Així ens queda una base de dades lliure d'outliers, amb un tamany de 21399 files * 12 col·lumnes. Per tant s'han eliminate 11585 files, una mica més d'un terci del total de dades.
Aquest canvi el podem observer com anteriorment en una matriu de dispersió la qual podem trobar al annex. I es pot observar que una gran cantitat de punts que abans apareixien ja no hi són. Això succeix per a tots els atributs no sols per aquells que em eliminat outliers. Per el que respecta a la distribucció tampoc s'obté un resultat que ens faci sortir de dubtes. I pel que fa els atributs amb més correlació tampoc permet arribar a una conclusió que ens permeteixi sortir de dubtes.

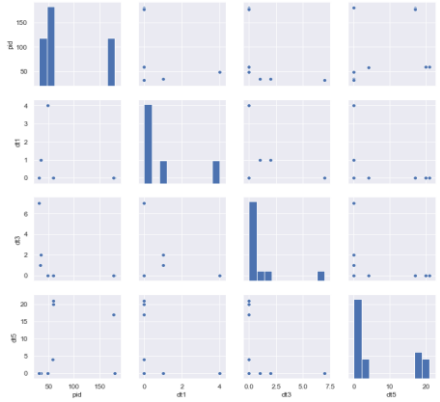

Fig.14: Matriu de dispersió atributs amb major correlació

A continuació s'entrena un regressor múltiple lineal utilitzant tots els atributs amb l¡objectiu de predir l'atribut PID. La predicció del model entrenat es mostra amb línies vermmeles. Comentar que abans de procedir a realitzar l'entrenament les dades s'han estandarditzat amb la funció standarize(), la qual retorna les dades restades per la mitja i divides per la desviació estàndard.

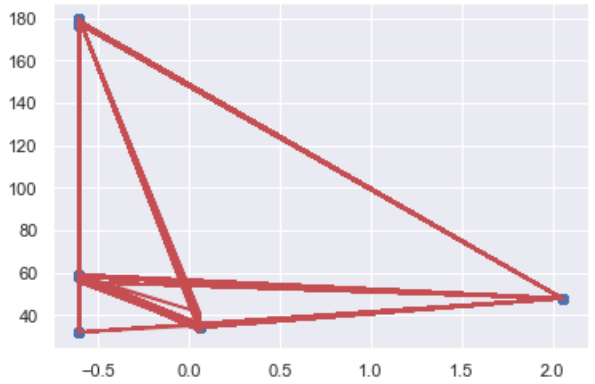

Fig.15: Regresor Múltiple Lineal amb tots els aributs

Els resultats obtinguts per al MSE son de 0,163399658038 i per a R2 score de 0.9999. Aquest resultats son molt bons, encara que aquest regresor no es el que més s'adecua a la distribució de les dades. Però al eliminar outliers els atributs tenen una relació lnea més pronunciada amb el atribut a predir ja que les dades estan ubicades en 5 punts per aixó aquest regressor predeix bé el valor i si que es podria utilitzar en aquest cas.

A continuació s'entrena un regresor multiple lineal amb les dades originals sense eliminar outliers, solament amb les dades estandaritzades.
Els resultats que obtenim per al MSE son de 60082.3161816 i per al R2 score 0.3100172432, això representa que aquest regressor no funciona bé per a la nostra representació de les dades.

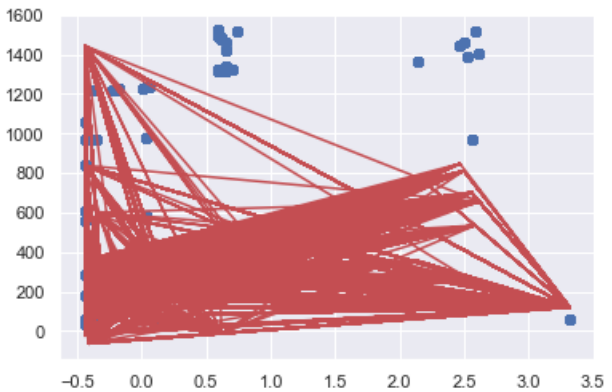

Fig.16: Regresor multiple lineal amb tots els atributs amb outliers

Ara bé, el que si cambia es la correlació dels atributs al eliminar outliers. Això es pot observar en la matriu de correlació del apèndix A2.
Veient la matriu de correlació veiem que ha cambiat considerablement. Ara tenim que dt5,dt6 i dt8 son els atributs

en més correlació respecte el atribut a predir i ademés aquest grau de correlació ha augmentat força. Per exemple la del atribut dt5 tenia una correlació de 0.15 mentres que ara a augmentat a 0.42, el que suposa una millora de 2,8 vegades més. Mentres que dt6 i dt8 tenien una correlació de -0.04 i -0.08 respectivament i a augmentat en el cas de dt6 fins a 0,73, el que suposa una millora del coeficient de 18,25 i 0,26 en el cas de dt8 que representa una millora del coeficient de 3,25.

Per això es decideix entrenar dos regresors més, per veure si es podria trobar una alternativa al regresor lineal múltiple. Els quals s'entrenaran per predir el valor de PID en funció dels atributs anterior amb correlació més forta. Els regresors escollits seran regressió logística i una maquina de vector de suport amb diferents kernels els quals son linear, polinòmica i gaussiana. El resultats es mostren amb diferents mides per a les dades de training.

Per poder determinar quina és la millor opció que hi ha que entendre l'informació que representa cada mètrica.

- Accuracy : indica en termes generals el percentatje d'encert de les mostres classificades be respecte del total.

- Precisió : te en compte sol el que es clasifica i en aquest sentit mesura si el que s'ha classificat es correcte. En aquest cas si tot el que classifica esta ben classificat aquesta mètrica tindra valor 1, encara que hi haguin més casos que s'han deixat de classificar. Per tant, sol té en compte el accert sobre la classificació feta.

- Recall : mesura els positius classificats respecte el total de positius. Per tant, aquesta mètrica penalitza el cas dels positius que no s'han classificat de forma correcta.

- F1-score : utilitza les mètriques de precisió i recall per generar aquesta mesura, fer que comporta un bon compromís i balanceig d'ambdues.

A continuació es mostren els resultats en forma de taula:

Regresión Logistica	Precision	Recall	F1-score	accuracy
0,5 - 0,5	1.0	1.0	1.0	1.0
0,7 - 0,3	1.0	1.0	1.0	1.0
0,8 - 0,2	1.0	1.0	1.0	1.0

Tabla.1: Regresión Logística

SVM Kernel Linear	Precision	Recall	F1-score	accuracy
0,5 - 0,5	0.83	0.89	0.85	0.9067
0,7 - 0,3	0.83	0.89	0.85	0.9067
0,8 - 0,2	1.0	1.0	1.0	1.0

Tabla.2: SVM Kernel Linear

SVM Kernel poly	Precision	Recall	F1-score	accuracy
0,5 - 0,5	0.83	0.89	0.85	0.9098
0,7 - 0,3	0.83	0.89	0.85	0.9071
0,8 - 0,2	0.83	0.89	0.85	0.9085

Tabla.3: SVM Kernel Poly

SVM Kernel rbf	Precision	Recall	F1-score	accuracy
0,5 - 0,5	0.83	0.89	0.85	0.9098
0,7 - 0,3	0.83	0.89	0.85	0.9116
0,8 - 0,2	0.83	0.89	0.85	0.9085

Tabla.4: SVM Kernel rbf

Veïent els resultats obtinguts del entrenament, els millors resultats s'obtenen amb SVM de kernel polinómic i de kernel rbf. Aquest resultat era d'esperar ja que aquest dos tipus de kernels son molt més flexibles que la regressió logística i que el kernel lineal. Encara que tenen el perill de overfitting, ja que son molt flexibles, i això faci que no prediguen bé els valors de pid.
Encara així si haguéssim d'escollir dos models dels anteriors, escolliria els dos esmentats SVM de kernel polinomic i kernel rbf, ja que SVM linear i la regressió logística donen bons resultats però son ficticis i por la distribución de los datos no s'adapaten gaire ni a una funció sigmoide ni a una funció fàcil per separar amb un sol hiperplà.

4.2 Red Neuronal

Un cop realitzat el EDA s'arriba a la conclusió que es difícil saber que es una anomalia i que no, per falta d'informació sobre les dades dt1 – dt8, ja que cada fabricant té les seves dades i també cambia per diferents models amb mateix fabricant. Per això s'ota a realitzar una red neuronal amb totes les dades disponibles considerant que aquestes fossin totes bones, així entrenar aquesta red neuronal de foma que pugui predir si arriba una trama amb diferents valors que aquella es una anomalia.
S'ha fet servir una red neuronal feedforward, aquesta s''ha entrenat amb un mètode d'activació de la tangent hiperbòlica ja que entrenarem amb valors transformats entre -1 i 1, cosa que facilita l'aprenentatge a les red neuronals.
La red neuronal estarà entrenada amb el 80% de les dades i el 20% restant serà per validació. També es transforma l'entrada en arrays amb forma (31585,1,8), això vol dir que tindrem 31585 entrades amb vector de 1x8. L'arquitectura neuronal tindra 8 inputs degut a que predeiim l'atribut pid amb 8 atributs (dt1 fins a dt8), comptarà amb 3 capes ocultes de 8 neurones , la sortida serà d'una sola neurona, utilitarem com optimizador Adam i com a mètrica de perdida Mean Absolute Error i finalment com la predicció seran valors continus per calcular l'Accuracy utilitzarem Mean

Squared Error. La red neuronal estarà formada per 4 capes ocultes la primera amb 72 neurones, la següent amb 48 neurones, al tercera amb 32 i la quarta amb 8 neurones.
Ara veurem alguns dels resultats obtinguts d'executar la red neuronal.
En la següent gràfica podem veure els valors que la red nueronal predeix que son els de color vermell i el verds representen els originals de la base de dades.

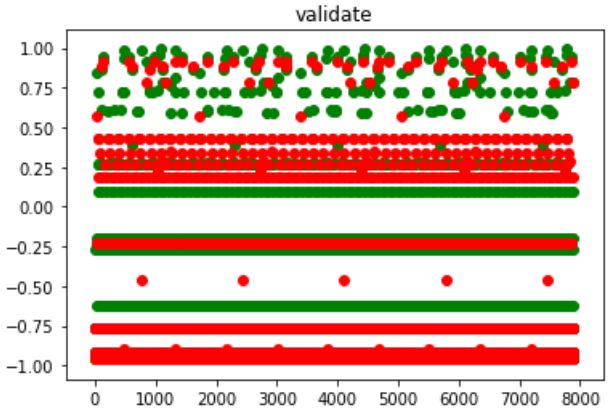

Fig.18: Gràfica de validació dels resultats de la red neuronal

Podem dir que aquesta red neuronal predeix força bé els valors de pid. Això també ho podem veure a les següents gràfiques. Podem observar que predeix força bé tot , encara que hi ha zones que no funciona del tot bé com els valors més alts, aquí es on te més errors.

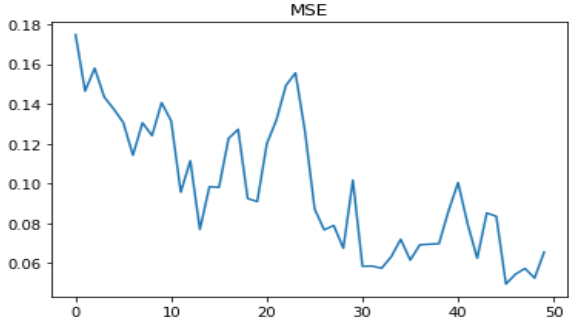

Fig.19: MSE de la predicció de la red neuronal

Aquesta gràfica ens monstre que el error MSE es molt baix en quant les prediccions encara que hi ha alguns pics. Això es degut a que no acaba de predir be del tot valors del pid molt alts com es pot veure a la Fig.19.
Aquesta red neuronal funcionaria força bé per predir anomalies, però seria difícil poder integrar-la directament al a alguna ECU, ja que aquestes careixen de potencia.
(Més informació al apèndix A2)

3 Conclusions

La conclusió extreta després d'haber realitzat totes les entregues i havet fet l'estudi de les comunicacions del bus de dades del vehicle, és que resulta difícil poder hackejar un cotxe fàcilment. Primer s'hauria de dur a terme un gran estudi de totes les trames possibles i com aquell cotxe actua amb elles. Per això necessitaries també un cotxe per poder fer proves i aquell estudi sol valdria per aquell cotxe depenent de la marca i model. Pot ser per això es que les marques no inverteixen en la seguretat informàtica dels vehicles.
Però si una cosa esta clara es que aquest fet aviat canviarà degut a que el futur avança cap els vehicles autònoms, els quals porten cada cop més també dispositius electronics i tard o d'hora hi haurà alguna forma encarà més fàcil per malmetre la seguretat d'un vehicle.
Els resultats obtinguts de la red neuronal utilitzada són força bons per a la prevenció de atacs, el problema d'aquesta sol·lució seria com poder executar aquesta en el sistema delsvehicles, ja que aquets no disposen de una CPU suficientment potent per això. Aquest fet seria un treball futur que s'hauria d'estudiar detingudament per no ocasionar vulnerabilitats del sistema amb les modificacions per poder executar la red neuronal.

Agraïments

Voldría agrair a la meva família el seu suport durant aquests anys i sobretot en aquest últim el qual ha sigut força difícil.

Bibliografia

[1] National Instruments, CAN BUS.
URL:https://www.ni.com/es-es/innovations/white-papers/06/controller-area-network--can--overview.html

[2] Aficionados a la mecánica CAN BUS:
URL: http://www.aficionadosalamecanica.net/canbus.html

[3] Global Information Assurance Certification Paper :
URL:https://www.giac.org/paper/gcia/9927/hacking-bus-basic-manipulation-modern-automobile-bus-reverse-engineering/133228

[4] Abdenour LABED y Aomar SERIR Zakaria SAHROUI. FIP over Ethernet/IP for Real Time Distributed Systems Implementation.
URL: http://www.iaeng.org/publication/WCE2010/WCE2010_pp515-520.pdf

[5] Estudio para simular una Red CAN con aplicación en comunicación de dispositivos electrónicos en el automóvil.
URL: https://studylib.es/doc/8355154/estudio-para-simular-una-red-can---dspace-de-la-universid...

[6] Centraleta electrònica (ECU) configurable per al control de l'encesa i sistema d'injecció d'un motor de 4 temps.
URL:https://upcommons.upc.edu/bitstream/handle/2099.1/10997/PFC_MARC_SUBIRANA.pdf?sequence=1&isAllowed=y

[7] EVOLUCIÓN DE LOS PROCESOS DE DIAGNOSIS ELECTRÓNICA EN EL AUTOMÓVIL.
URL: http://zaguan.unizar.es/record/5851/files/TAZ-PFC-2011-216.pdf

[8] El CAN-Bus de datos
URL: http://www.steinbock.cl/wp-content/uploads/2015/04/9-Documento-Dise%C3%B1o-y-Funcionamiento-AUDI-CAN-Bus-de-Datos.pdf

[9] In-vehicle network intrusion detection using deep convolutional neural network.
URL: https://www.sciencedirect.com/science/article/pii/S2214209619302451

[10] El protocolo OBD/OBD2/EOBD.
URL: http://www.grupocircuit.com/el-protocolo-obd2-eobd/

[11] On-board diagnostics.
URL: https://en.wikipedia.org/wiki/On-board_diagnostics

APÈNDIX

A1.Bibliografia Complementaria

[12] Todo lo que debes saber sobre el puerto OBD-II.
URL:https://noticias.autocosmos.com.mx/2016/06/08/todo-lo-que-debes-saber-sobre-el-puerto-obd-ii

[13] On the Development and Implementation of the OBD II Vehicle Diagnosis System
URL: http://www.ijeijournal.com/papers/Vol.7-Iss.4/D07041927.pdf

[14] DEVELOPMENT OF OBD-II DRIVER INFORMATION SYSTEM
URL:https://www.researchgate.net/publication/266176657_DEVELOPMENT_OF_OBD-II_DRIVER_INFORMATION_SYSTEM

[15] CANAuth - A Simple, Backward Compatible Broadcast Authentication Protocol for CAN bus
URL:https://www.researchgate.net/publication/235323481_CANAuth_-_A_Simple_Backward_Compatible_Broadcast_Authentication_Protocol_for_CAN_bus

[16] API Keras
URL: https://keras.io/guides/functional_api/

[17] API Sklearn
URL: https://scikit-learn.org/stable/modules/classes.htm

[18] API Matplotlib.pyplot
URL: https://matplotlib.org/api/pyplot_api.html

A2. MATRIU DE CORRELACIÓ

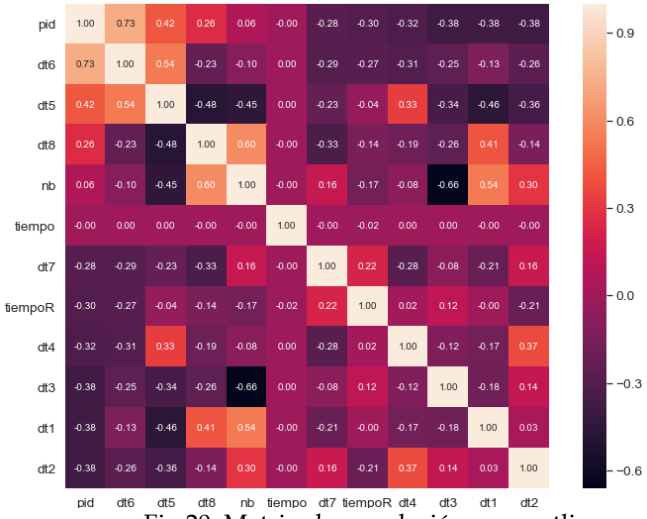

Fig.20: Matriu de correlació sense outliers.

Matriu de correlació, representa quins atributs tenen major correlació amb l'atribut pid abans de treure outliers.

A3. RED NEURONAL

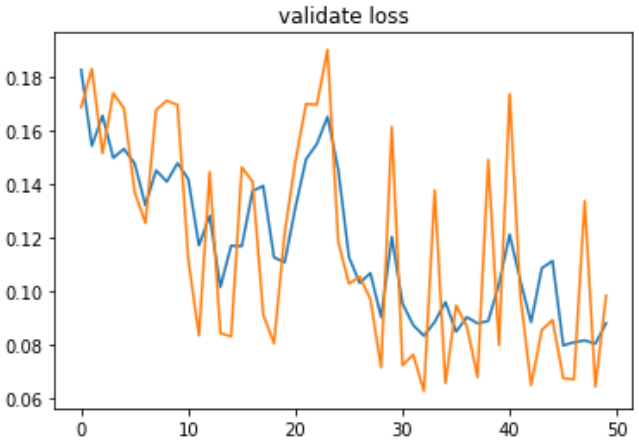

Fig.21: Gràfic de perdues

A partir del gràfic de perdues, podem veure que el model té un rendiment comparable tant al entrenament com al conjunct de dades de validació, pel conjunct d'entranament gráfica blava i per al conjunt de validació gràfica taronja. Encara que aquest s'en va una mica, ja que hi alguns valors que no acaba de predir be o si hi ha algún canvi molt brusc d'un valor alt a un baix també falla una mica però pel que fa la resta de valors es predeixen força bé.
Això també es pot observar a la gràfica compartiva entre el valors reals i predit, els relas la gràfica de color blau I predits gràfica e color taronja.

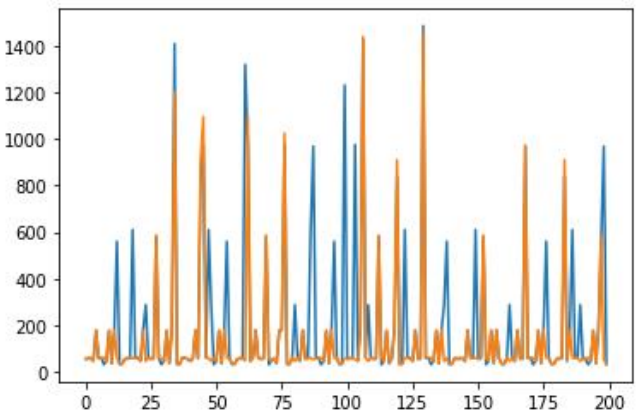

Fig.22: Gràfic comparatiu valors reals vs predits

En aquesta gràfica s'aprecia millor com alguns valors no acaben de ser predits correctament. Tot I així aquesta red neuronal funciona força bé i podria ser utilitzada com a metode per detecta anomalies.
Aquesta es podria millorar si el fabricant dones informacio més detallada de les dades dt1 – dt8, d'aquesta manera es podria tinder en compte algun factor que ara es impossible de veure.

A4. DICCIONARI CONVERSIÓ PIDS

'423': 1059, '30': 48, 'b3': 179, '3cf': 975, '38': 56, '244': 580, '4c6': 1222, '3a': 58, '3b': 59, '348': 840, '39': 57, '20': 32, 'b1': 177, '25': 37, '3e': 62, 'b4': 180, '262': 610, '22': 34, '23': 35, '120': 288, '230': 560, '3c9': 969, '3cb': 971, '3ca': 970, '3c8': 968, '3cd': 973, '527': 1319, '5c8': 1480, '528': 1320, '5ec': 1516, '5a4': 1444, '4ce': 1230, '4d0': 1232, '591': 1425, '521': 1313, '5cc': 1484, '5ed': 1517, '4c8': 1224, '526': 1318, '529': 1321, '5b6': 1462, '52c': 1324, '5d4': 1492, '5f8': 1528, '554': 1364, '553': 1363, '4c1': 1217, '540': 1344, '56d': 1389, '4c7': 1223, '4c3': 1219, '57f': 1407, '4d1': 1233, '5b2': 1458

En aquest diccionari es mostra el canvi de hexadecimal a decimal el dígit de l'esquerra representa el número que trobem a la base de dades original en hexadecimal i el dígit de la dreta representa el dígit obtingut de fer la conversió a decimal.

A5. DETECTOR D'ANOMALIES

També es va implementar un detector d'anomalies segons el temps de resposta. Aquestes bases de dades utilitzades en aquest cas estaben separades per pid es a dir nomes contenien trames provinents de la mateixa zona d'un vehicle. Observant la distribució d'aquestes dades segons le temps de resposta es podia apreciar que tenien una bona distribució normal, la qual s'utilitzara per aquest detector d'anomalies.

El que fa aquest detector es calcular un sostre i un terra, es a dir aprofitant la distribució normal de les dades es calcula la mitja de 15 en 15 dades utilitzant la funció Rolling() i aquesta dada se li resta o suma la distribució estandard * sigma, de cada dada segons sigui el terra o el sostre respectivament, d'aquesta forma s'adapta a la representació de les dades, així obtenim uns parametres resultants finals que els pendrem com a mesura per saber si un temps és una anomalia o no.

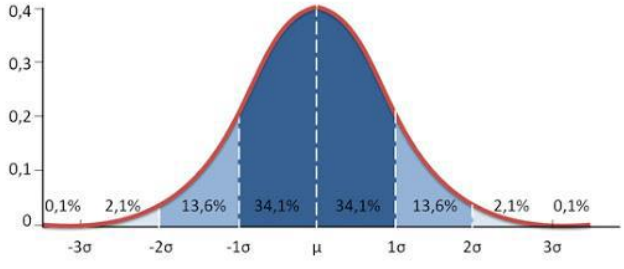

Fig.23: Distribució Normal

Com podem veure a l'anterior imatge Fig.23, segons el parameter sigma que es declari aquest detector serà més o menys estricte. En el nostre cas per veure el funcionament vam assignar sigma = 1 la qual cosa significa que el nostre detector es força estricte. Cal dir que lo normal seria posar sigma = 2 i considerer les dades for a d'aquestes quotes com anomalies.

Alguns dels resultats obtinguts son els següents:

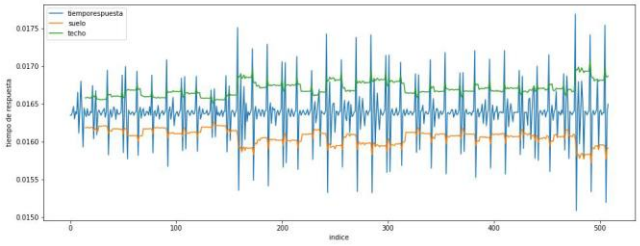

Fig.24: Gràfica de les dades de 100.csv sense mostrar anomalíes

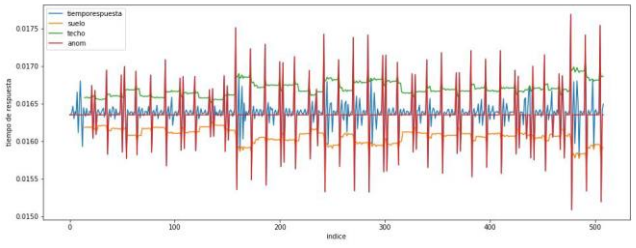

Fig.25: Gràfica de les dades de 100.csv amb anomalíes

En aquestes imatges podem observar el sostre i el terra calculats amb els colors verd i taronja en aquest ordre respectivament. Al haver assignat sigma = 1 veïem que es detecten força anomalies, però com s'ha dit abans a l'hora d'implementar-lo a un vehicle s'hauria de deixar com a miniím sigma = 2. A la imatge de la Fig.25 es pot veure en la gràfica de color vermell les anomalies detectades que sobrepassen o el sostre o el terra definits com límit per a la detecció d'anomalies.