**Dissertation Summary**

Propunerea mea pentru lucrarea de dizertatie face referire la procesul de analiza si predictie a consumului de combustibili si noxe, dorindu-mi astfel demonstrarea potentialului unor arii ca Data Mining si Machine Learning in industria automotive.

Cateva articole au aratat ca acest process de analiza se poate face din doua mari puncte de vedere:

* Analiza in raport cu parametri “constructivi” ai vehiculului (masa acestuia, capacitatea cilindrica, anul de productie, clasa de poluare din care face parte (Euro1 – Euro6) samd) – exista dataseturi speciale, ca de exemplu cel de la Vehicle Cerification Agency;
* Analiza in raport cu comporatamentul in timpul sofatului (driving behaviour) caracterizat de factori ca viteza de deplasare, numarul de rotatii pe minut al motorului, modul de accelerare etc – acesta necesita o monitorizare mai atenta a vehiculelor cu ajutorul unor dispozitive cum ar fi tahograful.

In cadrul lucrarii mele imi propun doua mari componente:

* Una teoretica de analiza a datelor;
* Cealalta practica care ar consta intr-o aplicatie mobila de tracking a datelor de la soferi si predictie a consumului de combustibil/noxe;

Intrucat colectarea de date privind comportamentul in timpul sofatului este mai dificila, in cadrul analizei mele ma voi focusa pe prima categorie, folosindu-ma de datasetul oferit de Vehicle Cerification Agency. In cadrul analizei “teoretice” voi incerca sa urmez metodele folosite in articolul “Application of data mining methods for analyzing of the fuel consumption and emission levels”, cu posibilitatea imbunatatirii/diversificarii lor in masura timpului disponibil, si anume:

* Prima etapa va fi aceea de design si analiza statistica datelor culese. Se vor elimina valorile aberante, inlocui valorile lipsa (NA-s), anumite variabile continue numerice ar putea fi transpuse in variabile categoriale pe intervale numerice. De asemenea, desi articolul mentionat nu precizeaza acest lucru, in aceasta etapa s-ar putea realiza o analiza explanatorie a datelor (univariata - descrierea variabilelor prin metode ale statisticii descriptive: tendinta centrala, imprastierea, forma distributiei (skewness, kurtosis), frecvente pt atribute categoriale, grafice (boxplot si histograme), dar si bivariata – determinarea dependentei variabilelor prin calcul de corelatii si teste de independenta, vizualizarea rezultatelor). Un alt lucru care ar putea fi facut in aceasta prima parte, ar fi extragerea unor reguli de asociere relevante din punct de vedere al suportului si increderii, care ar putea fi utile in gasirea unor relatii intre variabile si determinarea factorilor cei mai relevanti din anumite puncte de vedere.
* O urmatoare etapa se refera la analiza datelor prin clasificare nesupervizata (clusterizarea lor) cu scopul indentificarii anumitor grupuri naturale. Algoritmii de clusterizare mentionati a fi folositi in articolul de mai sus sunt K-means si Two-Step. Despre cel din urma se considera ca este mult mai potrivit a fi folosit pentru dataset-uri mari formate atat din variabile categoriale cat si continue. Este indicata folosirea mai multor metode de clusterizare din dorinta de a gasi numarul optim de clustere si de a gasi metoda care se potriveste cel mai bine in raport cu datele.
* Dupa aceasta procedura, va avea loc un proces de invatare supervizata (clasificare). In comparative cu clusterizarea care ofera informatii cu privire la anumite similaritati si deosebiri din datele noastre, clasificarea este aceea pe care se pot prezice output-uri pe baza unor noi date. Metodele de clasificare mentionate de studio sunt cele bazate pe Retele Bayesiene, Retele neuronale si algoritmul C5.0.
* Un aspect important nementionat in articol este procesul de feature selection.
* O ultima etapa a procesului de analiza presupune evaluarea rezultatelor obtinute. Se face o comparatie a algoritmilor de clusterizare privind numarul de grupuri obtinute si centroizii acestora, iar in ceea ce priveste algoritmii de clasificare metrica definitorie este acuratetea obtinuta. De asemenea, pe baza weight-urilor obtinute (la reteaua neuronala de exemplu) se poate evalua importanta factorilor.

De asemenea, se pot realiza anumite grafice de evaluare a diverselor criteriu.

In principiu, partea de analiza a datelor as dori sa o realizez folosindu-ma de modulele special oferite de limbajul Python si de tool-urile oferite de AWS (Amazon Web Services), ca mai apoi algoritmii de predictie sa-i pot expune sub forma unor servicii Web in cloud.

Pentru partea practica, imi propun aplicatia unei aplicatii Android dedicate in principal soferilor. Acestia vor introduce datele cu privire la autoturismul lor, apeland astfel un serviciu Web care va realiza o predictie asupra consumului de combustibil si noxe aferent. Aceasta ar putea functiona si ca un tool de predictie si ca unul de tracking si statistica (un “jurnal de bord” al soferilor), putand imbogati astfel dataset-ul oferit de Vehicle Certification Agency cu noi date de la conducatorii auto.