Математичка гимназија, Београд

МАТУРСКИ РАД
из физике

Примена метода машинског учења у физици животне средине

Ученик:
Лазар Златић IVa

Ментор:
др Андреја Стојић

Београд, мај 2019.

САДРЖАЈ

1. УВОД

Повећање потребе за енергијом, пораст броја становника, економски развој, урбанизација и транспорт, проблем загађења ваздуха постављају у фокус савременог друштва, првенствено због штетних ефеката на здравље људи, животну средину и климатски систем. Међународна агенција за истраживање рака је идентификовала бензен као канцерогену материју повезану са хромозомским аберацијама и великим бројем обољења у која спадају анемија и малигна обољења, пре свега леукемија, лимфом и мијелом. Краткорочна изложеност високим концентрацијама бензена може узроковати главобољу, опијеност, дрхтање и губитак свести, док изложеност бензену током трудноће делује токсично на плод и узрокује смањену телесну масу новорођенчета.

Са порастом рачунарске моћи и развојем метода машинског учења, отвара се могућност да се велике количине података о физици атмосфере, прикупљаних током низа претходних година, сагледају на нов начин и повежу у заједничку слику. Ове нове методе омогућавају да се открију међусобне повезаности унутар података и да се ти резултати ефективно представе. У оквиру овог рада анализираћемо податке везане за нашу атмосферу, фокусирајући се на утицај њених физичких својстава и концентрација једињења присутних у ваздуху на концентрацију бензена у атмосфери, као пример, и затим интерпретирати резултате нашег програма користећи савремене методе машинског учења и интерпретације добијених модела.

2. МАШИНСКО УЧЕЊЕ

Машинско учење је подобласт истраживања **вештачке интелигенције** која се бави алгоритмима и статистичким методама које рачунарски системи користе за решавање сложених проблема или доношење одређених предвиђања без директних инструкција, на основу налажења међусобних веза унутар података.

2.1. Врсте метода машинског учења

Основни типови машинског учења су надгледано учење (*supervised learning*) и ненадгледано учење (*unsupervised learning*). Надгледано учење подразумева предикцију резултата (исход циљне променљиве) за приложене податке датог проблема на основу доступне базе података (*dataset*) која садржи инстанце (*instance, datapoint*) са одређеним својствима, односно променљивама (*features, atributes*), и већ познатим резултатима (нпр. идентификовање предмета на слици на основу претходних слика за које је решење дато). На основу тога да ли резултат алгоритма узима вредност из дискретног или континуираног скупа вредности, проблем машинског учења може бити класификациони (претходни пример са одређивањем објекта на слици) или регресиони (нпр. одређивање вредности некретнине на основу релевантних података). Код проблема ненадгледаног учења, пак, нису познати излази за дате податке, већ ова врста учења углавном подразумева проблеме груписања инстанци података (*clustering*) или откривања аномалија у приложеним подацима. Такође, постоји и учење поткрепљивањем (*reinforcement learning*), које се не групише у претходна два типа и које подразумева оптимизацију понашања програма машинског учења у одређеној проблемској средини (нпр. учење оптималних стратегија одређених компјутерских игара). Ми ћемо се за потребе овог рада фокусирати на **регресионим методама надгледаног учења.**

2.2. Појмови надгледаног учења

2.2.1. Модели и параметри

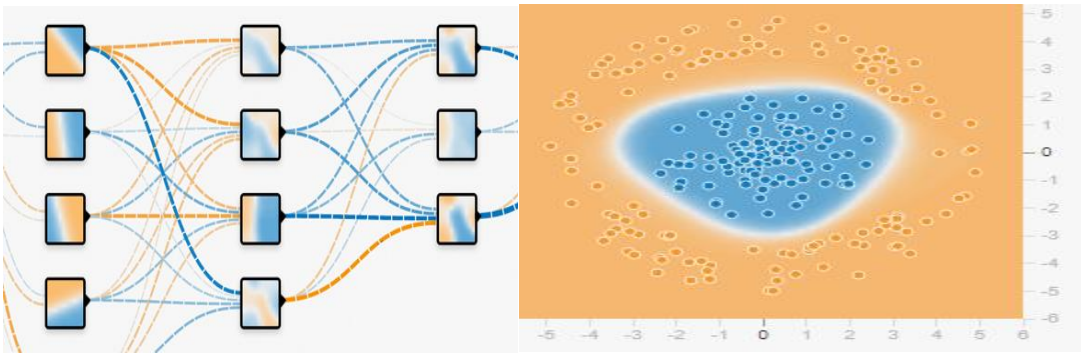
Модел машинског учења је обучени програм који пресликава улазне податке у одговарајући резултат, односно предикцију. Он дакле представља функцију која враћа предикцију тражене вредности y_i за инстанцу на основу улазних података за ту инстанцу x_i. Најједноставнији пример био би линеарни модел где је предикција дата са:

$$\hat{y}_i = \sum_j \theta_j x_{ij}$$

Вредности θ представљају **параметре** (мада у сложенијим моделима параметри могу бити знатно компликованијег облика и има их знатно више). Тренирање неког модела у ствари подразумева учење параметара θ за које модел даје добру предикцију.

2.2.2. Класе модела

У оквиру машинског учења постоји више **класа** модела (неки од најпознатијих су неуронске мреже, стабла одлучивања, метода потпорних вектора, итд.). Касније ћемо се у овом раду фокусирати на стабла одлучивања.



Слика 1. Илустрације метода машинског учења[1]

2.2.3. Функција евалуације

Да би успешно тренирали модел, морамо дефинисати **функцију евалуације** (функција ризика) која ће бити мера поклапања модела и података над којима модел тренира. Она се представља као збир два члана:

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta)$$

где је L **функција грешке** (*training loss function*), а Ω **регуларизациони члан** (*regularization term*). Функција губитка говори нам колико је модел добар у предвиђању резултата из скупа података за тренинг. Чест избор за ову функцију је функција средње квадратне грешке:

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$$

или функција логистичког губитка:

$$L(\theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})]$$

[1] Преузето са http://playground.tensorflow.org, приступљено априла 2019. године

Регуларизациони члан је функција чијим се додавањем у израз за функцију евалуације смањује могућност да дође до превелике прилагођености модела (*overfitting*). ***Overfitting*** представља проблем превелике прилагођености модела на конкретан скуп тренинг података чиме се смањује квалитет његових предикција на непознатим скуповима података, односно грешка генерализације расте. Регуларизациони члан се понекад изоставља ради једноставности, али коришћење одговарајуће функције за овај члан углавном доноси велика побољшања у квалитету предикције. Одабир те функције зависи од природе проблема и самог модела који користимо.

2.3. Стабла одлучивања

Модели **стабала одлучивања** функционишу по принципу вишеструке поделе базе података на подскупове на основу вредности својстава инстанци. У сваком кораку (при свакој подели) делимо скуп података на подскупове тако да свака инстанца из почетног скупа припада само једном од подскупова. Крајњи подскупови називају се терминални чворови или листови, а чворови између листова и почетног скупа називају се унутрашњи чворови. Постоји више алгоритама на основу којих можемо креирати стабло одлучивања (на основу броја подела по чвору или критеријумима за поделу података), али вероватно најпопуларнији је **CART** (*Classification and regression trees*) алгоритам. Код регресионих проблема, CART алгоритам врши поделу на подскупове тако што за дато својство одређује преломну вредност за коју је дисперзија предикција вредности*у*по чворовима максимална (дакле најбоља подела је она која чини добијене подскупове што више међусобно различитим у односу на тражени резултат). Овај поступак се даље понавља на добијеним подскуповима све док се не достигне одређени критеријум за заустављање (нпр. минимални број инстанци у подскупу да би он могао бити унутрашњи чвор).

Тада је формула која описује зависност предвиђања за вредност *y* од улаза *x* дата са:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

Где свака инстанца припада једном од листова, односно подскупова R_m и $I\{x \in R_m\}$ представља функцију индикатор која враћа 1 уколико $x \in R_m$, односно 0 у супротном. Уколико је инстанца у листу који одговара подскупу R_m, онда је предикција $\hat{y} = c_m$, где је c_m средња вредност свих резултата за тренинг инстанце у листу R_m.

2.4. Ансамбли метода машинског учења

Ансамбл методи подразумевају коришћење више алгоритама машинског учења и комбиновања (агрегације) њихових резултата да би се дошло до бољих предикција.

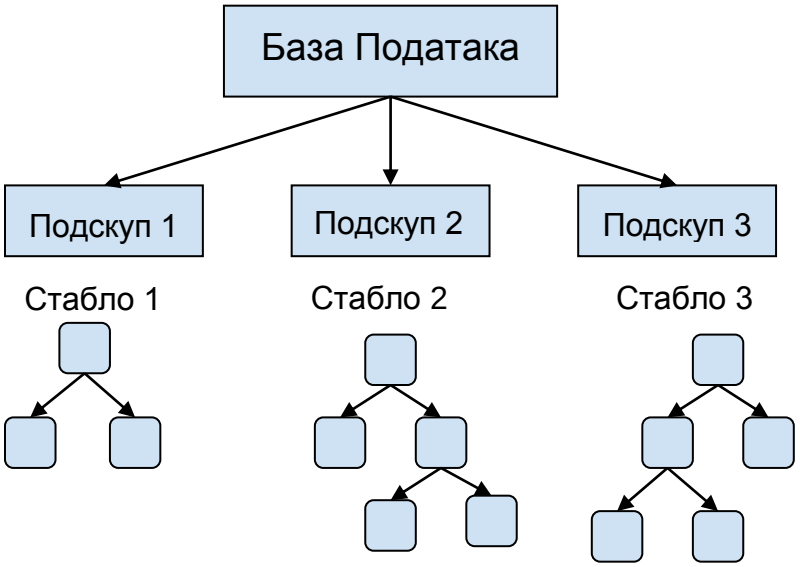
2.4.1. Случајне шуме

Случајне шуме (*Random forests*) представљају метод учења који се базира на конструисању ансамбла стабала одлучивања и формирања финалне предикције на основу агрегације резултата појединачних стабала. Ово се математички може представити као:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

где је K број стабала, а f_k функција из скупа F свих CART-ова.

Насумичност у овој методи уноси чињеница да се у сваком унутрашњем чвору подела врши на основу насумично одабраног подскупа својстава. Дакле, подела у чвору се врши не на основу најважнијег својства (као код обичног стабла одлучивања), већ на основу најбољег својства из датог насумичног скупа. Овакав приступ резултира већом разноврсношћу и већим ансамблом за агрегирање резултата, што продукује већу тачност предикције. Још једна предност овог модела је да омогућава лако препознавање релативне важности одређеног својства за предвиђање, што постаје веома важно касније за интерпретацију резултата.



Слика 2. Илустрација методе случајне шуме (Random forests)

2.4.2. *Gradient Tree Boosting* и *Extreme Gradient Boosting*

Gradient tree boosting представља једну од техника ***gradient boosting*** концепта коришћену углавном са стаблима одлучивања. Алгоритам *gradient boosting* подразумева

иницијализацију почетне вредности функције предикције $f_0 = \gamma$, тако да је вредност суме функција губитака:

$$\sum_{i=1}^{n} L(y_i, \gamma)$$

где је n број инстанци у бази података, минимална. Затим се жељени број пута (M) итеративно понавља (за $m = 1, 2, 3, \ldots, M$) следећи поступак:

1. Прво се за сваку инстанцу i из базе података врши израчунавање псеудо-остатака r_{im} где је:

$$r_{im} = -\frac{\partial L(y_i, \hat{f}_{m-1}(x_i))}{\partial \hat{f}_{m-1}(x_i)}$$

2. Затим се базични модел (у нашем случају стабло одлучивања) тренира на бази података у којој су тражени резултати y_i замењени псеудо-остатком r_{im} и добија се предикција аналогно оној код обичних стабала одлучивања:

$$h_m(x) = \sum_{j=1}^{N_m} c_j I\{x \in R_{jm}\}$$

3. Потом израчунавамо фактор γ_{jm} тако да је вредност суме функција губитака:

$$\sum_{i=1}^{n} L(y_i, \hat{f}_{m-1}(x_i) + \gamma_m h_m(x_i))$$

минимална.

4. Последњи корак је ажурирање модела:

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m h_m(x)$$

На крају тражено предвиђање једнако је крајњем резултату нашег итеративног процеса $f_M(x)$.

Extreme gradient boosting (XGBoost) је једна од имплементација методе *gradient boosting* за стабла одлучивања која је дизајнирана за високе перформансе и доступна је у библиотекама за популарне програмске језике за машинско учење, те ће стога она бити наш избор при изради овог рада.

3. ИНТЕРПРЕТАЦИЈА МОДЕЛА МАШИНСКОГ УЧЕЊА

Након конструисања одговарајућег модела који треба да нам изнесе предикцију везану за наш проблем, поставља се питање интерпретације поступака којим наш модел долази до предвиђања. Нарочито у научним истраживањима ово питање је кључно, јер нам даје представу о реалном смислу поступака нашег модела.

Иако постоји много метода које нам у општем случају могу бити корисне, ми ћемо се осврнути на неколико агностичних метода (метода које не подразумевају познавање самог модела машинског учења и могу се применити у идентичном облику на све моделе) за интерпретацију модела машинског учења.

3.1. Важност променљивих

Важност променљиве представља пораст грешке предикције нашег модела након пермутовања вредности својства. Што је оваквим поступком већи пораст грешке предикције, то је већа важност тог својства односно те променљиве за нашу предикцију. Израчунавање важности променљиве *j* врши се израчунавањем почетне грешке нашег модела $e_0 = L(y, \hat{f}(x))$, затим израчунавањем грешке модела са пермутованим вредностима за циљану променљиву *j*: $e_{perm} = L(y, \hat{f}(x_{perm}))$, и на крају је тражена важност циљане променљиве једнака:

$$F_j = \frac{e_{perm}}{e_o}$$

Понекад се за важност променљиве користи и израз $F_j = e_{perm} - e_0$.

3.2. Парцијалне зависности

Током интерпретације наших резултата занима нас такође и **парцијална зависност** нашег предвиђања од једне или неколицине променљивих.

Функција парцијалне зависности за регресију дефинише се као:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

где x_S представља променљиву из скупа S својстава за које одређујемо парцијалну зависност предвиђања, а x_C представља променљиве из скупа C свих осталих својства коришћених у нашем моделу. Како нам је циљ углавном да ове парцијалне зависности

наше предикције од променљивих графички представимо, у скупу *S* ће најчешће бити само једно или два својства.

Функцију парцијалне зависности можемо проценити као:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})$$

где су $x_C^{(i)}$ вредности из наше базе података за својства из скупа *C*. Сада коришћењем дате формуле можемо графички представити функције парцијалне зависности нашег предвиђања од одређених својстава.

3.3. Интеракције

Због могућих међусобних **интеракција својстава** у нашем предикционом моделу, зависност наше предикције од неког скупа својстава може бити веома компликована. Стога нас интересује процена међусобних интеракција својстава у нашем моделу.

Један од начина да извршимо ову процену јесте да измеримо колико дисперзија вредности наше предикције зависи од интеракције наших својстава (*H-statistic*).

Приликом мерења интеракције својстава посебно нас интересују две вредности. Првенствено мера интеракције два својства, тј. да ли и у којој мери та својства интерагују, а затим и укупна мера интеракције неког својства са свим осталим својствима.

Уколико два својства x_j и x_k не интерагују за њихове функције парцијалне зависности важиће:

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k)$$

где је $PD_{jk}(x_j, x_k)$ функција парцијалне зависности предвиђања за оба својства, а $PD_j(x_j)$ и $PD_k(x_k)$ функције парцијалне зависности за свако својство појединачно.

Аналогно се наша предикциона функција може записати у облику:

$$\hat{f}(x) = PD_j(x_j) + PD_{-j}(x_{-j})$$

где је $PD_{-j}(x_{-j})$ функција парцијалне зависности за сва својства осим x_j.

Следећи корак је да израчунамо разлику између посматране вредности функције парцијалне зависности за оба својства (у случају мерења интеракције између две променљиве), односно функције предикције које добијамо (у случају мерења укупне интеракције неког својства са осталим својствима) и вредности које добијамо када дате функције разложимо на раније приказан начин, као у случајевима када интеракција не постоји. Дисперзија објашњена овом разликом се користи као статистика снаге интеракције и узима вредности између 0 (нема интеракције) и 1.

Вредности *H*-статистике су:

$$H_{jk}^2 = \sum_{i=1}^{n} \left[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right]^2 / \sum_{i=1}^{n} PD_{jk}^2(x_j^{(i)}, x_k^{(i)})$$

за интеракцију између својстава j и k, односно:

$$H_j^2 = \sum_{i=1}^{n} \left[\hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)}) \right]^2 / \sum_{i=1}^{n} \hat{f}^2(x^{(i)})$$

за интеракцију својства j са свим осталим својствима.

3.4. SHAP вредности

SHAP (*SHapley Additive exPlanations*) **вредности** интерпретирају важност одређене вредности својства за предикцију за дату инстанцу. SHAP вредност се израчунава као разлика предвиђања за дату инстанцу узимајући у обзир дато посматрано својство, односно искључујући га. Ово је сличан поступак израчунавању важности променљивих, али доноси одговарајућу предност, јер док приликом израчунавања важности променљивих посматрамо својство глобално, SHAP вредност нам даје важност те променљиве на нивоу инстанци. Такође *SHAP Values* техника је погодна за коришћење заједно са XGBoost моделом и такође има своју имплементацију у библиотекама за програмске језике популарне за машинско учење.

4. ПРИМЕНА МЕТОДА У ФИЗИЦИ ЖИВОТНЕ СРЕДИНЕ

4.1. Загађење ваздуха

Од почетка индустријске револуције, па до данас, концентрација штетних материја у нашој атмосфери је у константом порасту, али се овом проблему тек од недавно приступа са већом пажњом.

Као главне загађујуће супстанце у ваздуху јављају се оксиди угљеника (угљен-моноксид и угљен-диоксид), оксиди азота и сумпора, амонијак, неки токсични метали, испарљива органска једињења, итд.

4.1.1. Испарљива органска једињења

Испарљива органска једињења (*volatile organic compounds - VOCs*) представљају органске материје које услед веома ниске температуре кључања лако испаравају или сублимују. Велики број испарљивих органских једињења има веома штетан утицај на здравље човека и животне околине.

Ми ћемо се у овом раду фокусирати на **бензен**, као испарљиво органско једињење веома штетног дејства на човеково здравље, и методама машинског учења одредићемо утицај других фактора атмосфере на концентрације бензена у ваздуху.

4.2. Анализа бензена у метеоролошком контексту

Бензен (C_6H_6) је испарљиво органско једињење карциногеног дејства. Такође бензен може загадити храну или воду и довести до тровања које може резултирати повраћањем, вртоглавицом и убрзаним откуцајима срца, а у екстремним случајевима и смрћу.

Бензен у атмосферу доспева из дуванског дима, издувних гасова аутомобила и других превозних средстава, или као отпад при производњи пластике, боја и слично, а ређе и при вулканским ерупцијама или шумским пожарима.

На основу доступних података о концентрацијама различитих материја у атмосфери, као и физичких услова који су важили у тренуцима мерења (температура,

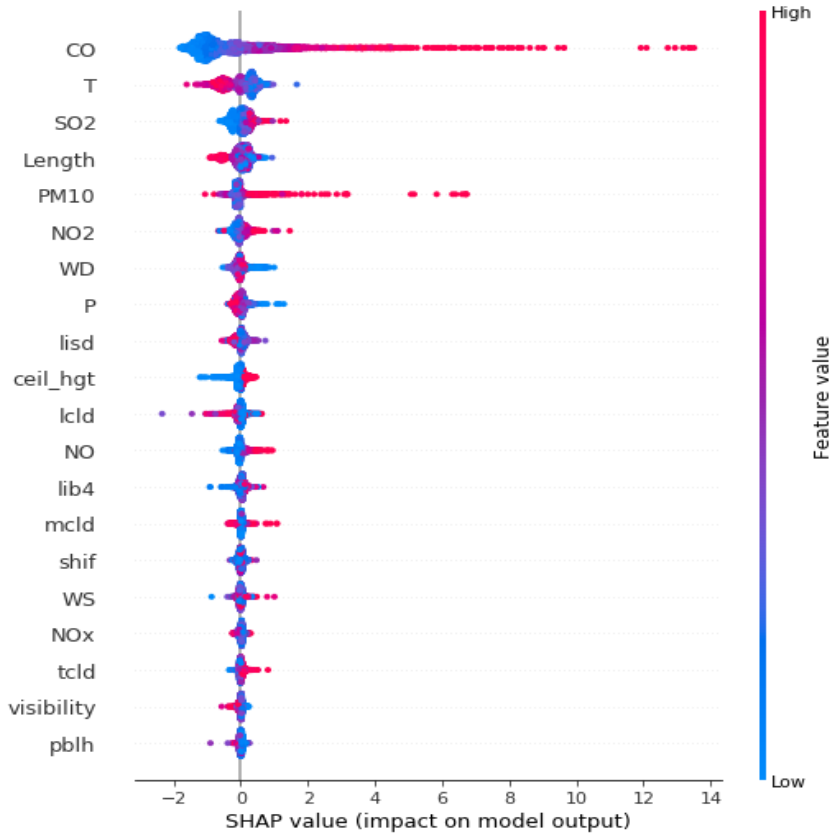
притисак, итд.), можемо, користећи претходно описане методе, одредити зависност измерене концентрације бензена у ваздуху од ових фактора.

Програм у оквиру кога примењујемо методе машинског учења и интерпретације података је урађен у програмском језику ***Python***, у окружењу ***Jupyter***.

Слика 3. Комплексност фактора животне средине који обликују еволуцију концентрација бензена у атмосфери

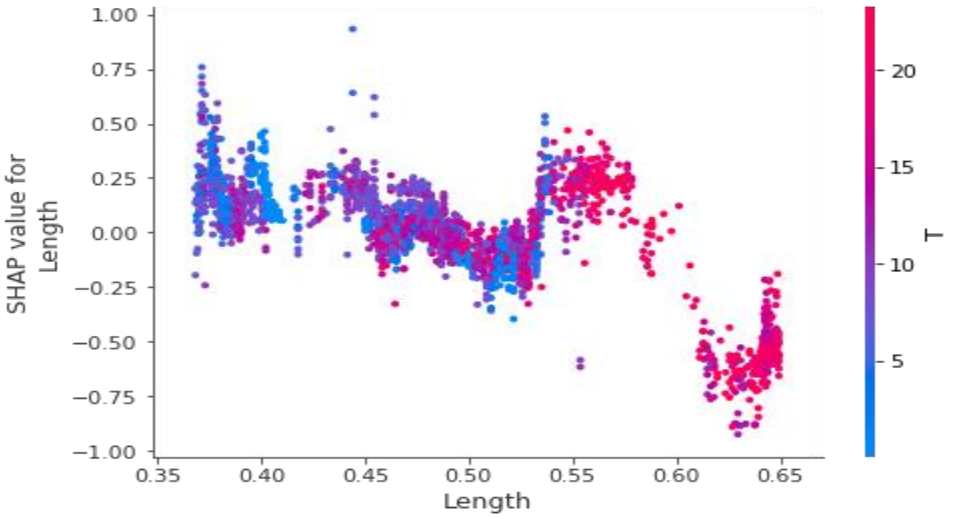
Након учитавања базе података у којој се налазе измерене вредности атмосферских фактора, затим поделе података на *тренинг* и *тест* податке и подешавања одговарајућих параметара претраге тако да наш алгоритам за обучавање постиже најбоље резултате, обучавамо наш модел XGBoost методом и добијене резултате затим можемо графички представити коришћењем SHAP библиотеке.

Добијени резултати нам откривају важности одређених својстава на концентрацију бензена у атмосфери. Тако видимо да ће највећи утицај на концентрацију бензена имати концентрација угљен-моноксида у ваздуху, што значи да ово једињење доминанто потиче из процеса сагоревања, али да ће значајан утицај имати и температура, концентрација сумпор-диоксида као и дужина трајања обданице (*Length* у табели).

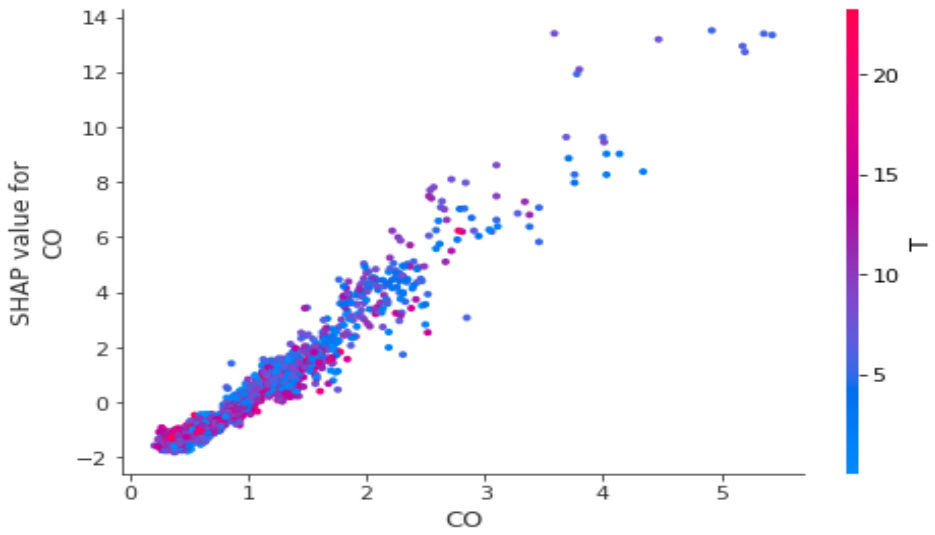


Слика 4. Збирни приказ вредности SHAP

Такође можемо узети у обзир и међусобни утицај одређених фактора.



Слика 5. Парцијална зависност концентрација бензена од температуре (T) и дужине обданице (length)



Слика 6. Парцијална зависност концентрација бензена од концентрације угљенмоноксида и дужине обданице

Овим добијамо увид у зависности које би теоријски било тешко предвидети, а помоћу којих можемо сагледати ширу слику нашег проблема. На пример, док се снажна корелација између концентрација бензена и присутности угљен-моноксида у ваздуху могла претпоставити због сличних извора обе загађујуће супстанце, снажна и изразито нелинеарна веза између концентрације бензена и дужине трајања обданице није очигледна, као ни конкретни облици зависности ових својстава које би теоријски било захтевно извести.

Наравно уколико желимо да повећамо поузданост и прецизност нашег модела можемо применити наш алгоритам над све већим и већим базама података или користити компјутере веће рачунарске снаге. Међутим, чак и на мањим скуповима података уз ограничену рачунарску моћ, брзо се долази до кључних зависности које нас највише интересују при анализирању нашег проблема.

5. ЗАКЉУЧАК

На примеру бензена видели смо како велике количине доступних података можемо, применом савремених метода машинског учења, брзо искористити за добијање конкретних резултата, чијим интерпретирањем добијамо бољи увид у физику наше атмосфере.

Машинско учење пружа нове могућности за разумевање сложених физичких процеса наше атмосфере, које пре није било могуће у потпуности сагледати класичним методама. Услед даљег напретка вештачке интелигенције, али и додатног развоја рачунарске моћи, методе машинског учења сигурно ће донети бројне резултате значајне за разумевање света око нас.

ЛИТЕРАТУРА

[1] https://christophm.github.io/interpretable-ml-book/, приступљено априла 2019. године.
[2] http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/, приступљено априла 2019. године.
[3] https://github.com/slundberg/shap, приступљено априла 2019. године.
[4] https://medium.com/@gabrieltseng/interpreting-complex-models-with-shap-values-1c187db6ec83, приступљено априла 2019. године.
[5] https://medium.com/civis-analytics/demystifying-black-box-models-with-shap-value-analysis-3e20b536fc80, приступљено априла 2019. године.
[6] https://xgboost.readthedocs.io/en/latest/, приступљено априла 2019. године.
[7] https://www.who.int/ipcs/features/benzene.pdf, приступљено априла 2019. године.
[8] https://en.wikipedia.org/wiki/Volatile_organic_compound, приступљено априла 2019. године.
[9] http://bpm.ipb.ac.rs/, приступљено априла 2019. године.