# ПОДГОТОВКА НА ДАННИТЕ ЗА АНАЛИЗ

**Даниела Орозова**
*Бургаски свободен университет*

# DATA PREPARATION FOR ANALYSIS

**Daniela Orozova**
*Burgas Free University*

***Abstract:*** *В тази статия се представят основните стъпки, за подготовка на данните, чрез системата Orange, в етапа на предварителен анализ. Процесът на подготовка на данните, включва идентифициране на грешки, коригиране, изтриване, подредба или друга обработка, като се запазят само потенциално полезните данни. Разглеждат се основни проблеми и техни решения, приложени са обучаващи примери.*

***Key words:*** *Data Analytics, Data Science, Virtual Education Space, E-learning.*

**Въведение**

Науката за данните (Data science) позволява да се съчетаят множество подходи като включва техники, свързани с анализ на данни от областта на статистиката, дейта майнинг и откриване на знания, машинно обучение, изкуствен интелект, програмиране, комуникация др. Науката за данните включва и процесите по изчистване и интеграция на данни, избор и трансформация на данни, извличане на знания, техния анализ, оценяване и представяне [1]. Може да се каже, че *Data Science* е „сплав“ от различни дисциплини и технологии.

*Orange* [11] е платформа, която може да ни помогне да решаваме проблеми в областта на *Data Science*. Но често в данните се срещат грешни стойности, пропуски в колоните, грешен формат и др. Това налага допълнителната им обработка, за да могат данните да станат с нужното за анализа им качество и да се използват в процеса на работа. Преди да бъдат анализирани данните е необходимо те да бъдат предварително изчистени: да се премахнат ненужни характеристики, да се кодират категорийните променливи, да се премахнат отличителни (екстремни) стойности и т.н.

Основни задачи за подготовка на данни са [3]: почистване на данните (Data Cleaning), интеграция на данните (Data Integration), трансформация на данните (Data Transformation), редуциране на данните (Data Reduction). В тази статия се представят основните стъпки в етапа на предварителна обработка на данните чрез средствата на системата *Orange*.

Почистването на данните е процесът за гарантиране, че данните са правилни, последователни и използваеми. *Дублиращи* се данни могат да възникнат при комбиниране на набори от данни от множество източници. *Неподходящи* данни са такива, които не отговарят на решавания проблем. *Структурни грешки* в данните могат да възникнат по време на измерване или прехвърляне на данни. *Outliers* (отличителни стойности) могат да причинят проблеми с някои видове модели. В някой случаи имаме *липсващи данни* (Missing Data*)*. Процесът на подготовка на данни, включва иден-

тифициране на грешки, коригиране, изтриване, подредба или друга обработка, като се запазят само потенциално полезните данни.

Когато се обработва огромно количество данни, в някой случаи се налага преобразование, за да се редуцира представеното множество данни, без това да промени резултата от работата с тях. *Редукцията* [2] на данните може да се реализира по различни начини като: редукция на размерностите – премахване на някои несъществени за анализа размерности (характеристики на данните); заместване на данните с по-малки по размер алтернативни данни, чрез параметризация; дискретизация и създаване на концептуални йерархии, чрез представяне на данните в групи (клъстери) на различни йерархични нива; компресиране на данните, с цел намаляване на физическия им размер и други подходи.

**Проблеми при подготовка на данните за анализ и подходи за решаване**

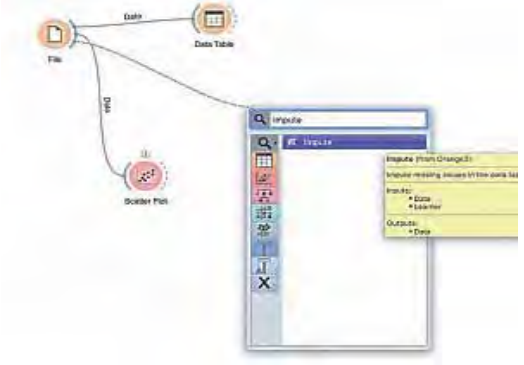
Процесът на подготовка на данните за анализ включва взимане на решения за обработка в редица ситуации. Следва представяне на основни проблеми при подготовка на данните за анализ и по какви начини могат те да бъдат обработени чрез системата *Orange*.

**1. Проблем – *липсващи стойности***

Когато част от данните липсват, това предизвиква проблеми в работата на моделите за машинно обучение и се отразява негативно върху точността на получените резултати. В практиката се използват различни подходи, чрез които данните да се преработят във вид, позволяващ преодоляването на тези проблеми.

*1.1. Изтриване* – радикалният подход е да се изтрият редовете или колоните, в които има липсващи стойности. При големи обеми от данни [8], ефектът от изтриване на няколко реда няма да има голямо влияние.

Чрез системата Orange се избира инструмента *„File*“, а след това инструмента „*Impute*“ (фигура 1).




*Фигура 1. Прилагане на инструменти „File“ и „Impute“.*

При липсата на прекалено много стойности в една колона (например над 50%) и слаба корелация с останалите, тя може да бъде премахната. Но ако има висока корелация между колоната с липсващи стойности и останалите характеристики, то най-добре е тя да се остави и да се потърсят начини те да се заместят.

*1.2. Заместване*. Изборът на подходящ метод за заместване на липсващите стойности се определя от типа на данните. Статистиката разделя данните на категорий-

ни/качествени (пол, степен на образование, семейно положение и др.) и количествени/числови (възраст, заплата и др.).

При *категорийните данни* се използва заместване с най-често срещаната стойност – мода, а при количествени данни едни от най-често използваните методи са:

- заместване със средна стойност или медиана;
- заместване с произволни стойности;
- заместване с константа (например 0);
- изчисляване на стойността чрез регресия;
- определяне на липсващата стойност въз основа на създаден модел на данните и други техники.

Когато щракнете двукратно върху инструмента *„Impute*“ (фигура 2), се представят различни методи за вмъкване, които могат да се използват. Ако се остави опцията *„Remove the rows with missing values*“, то редовете се изтриват. Други възможни опции са: *Distinct Value*, *Random Values*, *Model-Based*.


*Фигура 2. Прилагане на различни методи на инструмент „Impute“.*

**2. Проблем – *мащабиране (нормализация) на данните***

Целта на нормализирането е да промени стойностите на числовите колони в набора от данни до общ мащаб, без да изкривява разликите в диапазоните от стойности. Например, ако имаме набор от данни, съдържащ двете характеристики, възраст и доход. Където възрастта варира от 0-100, докато доходът варира от 0-20 000 и повече. Доходът е около 1000 пъти по-голям от възрастта и тези две характеристики са в много различни граници. Когато се прави допълнителен анализ, (например линейна регресия) дохода ще повлияе по същество на резултата поради по-голямата му стойност, без да е по-важна характеристика.

Нужно е да се извършат необходимите трансформации, така че стойностите да бъдат от еднакъв порядък. Тази техника се нарича мащабиране на характеристики (*Feature scaling*) и се прилага в етапа на предварителната обработка на данните (*Data*

*preprocessing*). За машинно обучение не всеки набор от данни изисква нормализиране. Но при работа на определени модели с немащабирани данни, например при такива, които изчисляват коефициенти, се получават резултати с голяма грешка. За всяка характеристика трябва да се изчислят теглови коефициенти и се прилага мащабиране.

Алгоритми за машинно обучение, които изискват данните да бъдат мащабирани са например: алгоритмите, изчисляващи разстояние [7]: k най-близки съседи (*k-Nearest Neighbors*); метод на опорните вектори (*Support Vector Machines*); k-means клъстеризация (*k-means clustering*). Също така алгоритми, изчисляващи коефициенти: регресионни алгоритми (логистична регресия, линейна, нелинейна и др.), невронни мрежи и др.

Прилагат се различни методи, като например [5]:

- *Zscore* – данните са трансформирани чрез изваждане на средната стойност и разделяне на стандартното отклонение.
- *MinMax* – изважда се минималната стойност на характеристиката да бъде 0, а стойността се получава като се раздели на новата максимална стойност, която е разликата между първоначалните максимални и минимални стойности (max-min).
- *Robust* – изважда се медианата (Q2) и полученият резултат се разделя на интерквартилния размах (Q3 - Q1).
- *MaxAbs* – мащабира данните за всяка променлива по нейната максимална абсолютна стойност.
- *LogNormal* – преобразува всички стойности в логаритмична скала.
- *TanH* – всички стойности се преобразуват в хиперболична тангента.

Когато данните са мащабирани, всички характеристики са равно поставени, т.е. липсва склонността алгоритъмът да възприема една променлива за по-значима от друга. Това въздейства положително върху работата на модела, като се подобряват финалните резултати и възможността да се направи по-точна оценка на качеството.

**3. Проблем – *отличителни стойности (outliers)***

В някой случаи в данните може да има стойности, които значително се различават от останалите. Те се наричат отличителни или екстремни и могат да бъдат необичайно високи или ниски. В някой случаи се интересуваме именно от стойности, които се различават драстично от останалите, защото съдържат важна информация.

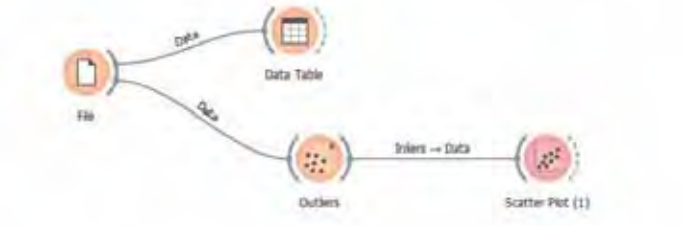
Анализ за такива данни се осъществява с инструмент *Z-score* [4]. Това е число, което показва колко се отдалечава дадена точка от средната, измерено в стандартни отклонения. Всяка стойност, която е по-ниска или по-висока от средната стойност плюс или минус три пъти стандартното отклонение може да се определи като отличителна стойност. Други подходи за откриване на отклонения са например линейни модели (линейна регресия, метод на главните компоненти и др.).

Но наличието на такива стойности може да се отрази негативно върху работата на моделите за машинно обучение и затова е необходимо те да се обработват по подходящ начин. Такива алгоритми са например линейни модели (регресионни и дискриминантен анализ) и SVM.

Отличителните стойности могат да бъдат възприети като *липсващи стойности* и да бъдат обработени като такива. Друг подход е *дискретизацията*. Това е процес на трансформация на непрекъснати променливи в дискретни, като стойностите се разпределят в определен брой интервали (bins). Дискретизацията е начин, чрез който мо-

жем да се справим с отличителните стойности, като те се поставят в най-ниския и най-високия интервал. След извършване на дискретизация тези променливи трябва да бъдат третирани като категорийни. Алтернативен подход е поставяне на горен и долен праг и *заместване с минимална и максимална допустима стойност*. Този подход позволява обработка на отличителните стойности без да се премахват записи от данните, но може да изкриви разпределението и връзката между отделните променливи.

За откриване на отличителните стойности на данните в системата *Orange* се използва инструмент *Outliers* (фигура 3).




*Фигура 3. Прилагане на инструмент „Outliers“.*

Идентификацията и обработката на отличителни стойности в данните е важен елемент при изграждане на модели за машинно обучение. Тя позволява получаването на по-високи стойности на метриките за оценка на работата на моделите, както и подобро цялостно представяне на самия модел.

**4. Проблем** – ***работа с балансирана извадка от данни***

В някой случаи, при решаване на задачи за класификация с методите на машинното обучение попадаме на извадка, при която представителите на класовете в целевата променлива не са равномерно разпределени, т.е. броят на обектите, принадлежащи към един клас, е значително по-различен от този на другите класове. Тогава казваме, че тя не е балансирана [5].

*Oversampling* е техника за балансиране на извадката с данни, при която се увеличава броя на обектите в по-малкото множество (minority), така че да се изравни с броя в по-голямото (majority).

*Undersampling* е техника, при която, се намаляват случаите в класа с повече представители до такава степен, че броят им да се изравни с този в класа с по-малко представители.

При прилагане на *oversampling* за балансиране на извадка се запазват всички оригинални записи, докато при *undersampling* може да се изгубят ценни данни, тъй като се премахват редове от множеството с повече представители. Може също така да се приложи и комбинация от двете техники. Изборът на техника зависи от данните, с които разполагаме, както и от решаваната задача.

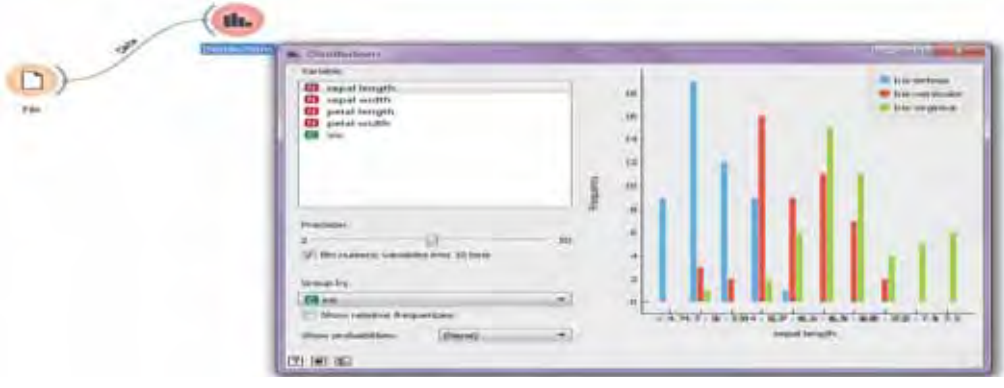
**5. Проблем** – ***коя е подходяща визуализация на данните***

При машинното обучение е много важен етапът на визуалния анализ, защото ни помага да добием представа за данните, като открием какви са зависимостите между отделните характеристики, какво е разпределението, дали има необичайно високи и ниски стойности и т.н. Визуализациите могат също да се прилагат и при представяне на крайните резултати от работата на моделите.

- ***Хистограми и графики на плътността на разпределението*** – използват се, когато трябва да се представи разпределението на данните и да се определи дали дадени измервания се различават съществено от други [4].

При хистограмата данните се разпределят в няколко интервала, като броя на обектите във всеки от тях се изобразява с височината на колоните. Графиката на плътността на разпределението е изгладен вариант на хистограмата и позволява полесно да се определи каква е формата на разпределението, защото не зависи от броя колони. В системата Orange те се представят чрез инструментът наречен *Distributions*. Той показва разпределението на стойностите на атрибути с дискретни или непрекъснати стойности.

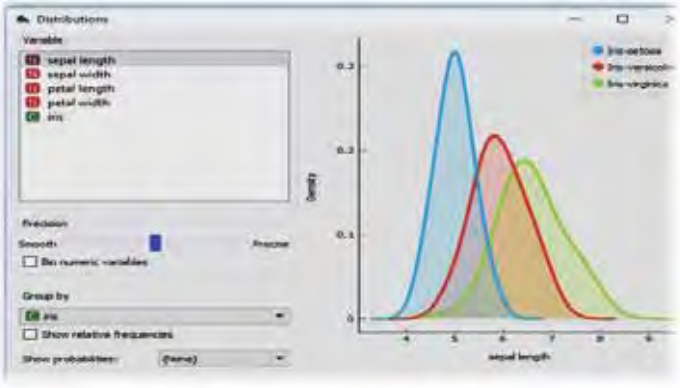
За *дискретни стойности на атрибутите*, графиката отразява колко пъти всяка стойност на атрибута се появява в данните. На фигура 4 е показана графика, като са използвани данните от „Iris.tab“ (файл от галерията на Orange, който съдържа информация за видовете на цветето ирис), сравнявайки стойностите на дължините на чашелистчетата за трите класа ириси.



*Фигура 4. Визуализация на данни с инструмент Distributions на „Orange“.*

Инструментът *Distributions* показва разпределенията на стойностите за указани променливи. Ако се маркират променливи с непрекъснати стойности, *Bin* инструментът ще дискретизира променливите, като ги зададе на интервали. Броят на интервалите се определя със зададена точност. Друга възможност е да зададем гладкост на кривите на разпределение на непрекъснатите променливи. От инструмента *Distributions* може да бъде поискано да показва раздели със стойности само за случаи от определен клас, използвайки *Group by*. Също могат да бъдат показани и изчислените вероятности чрез *Show probabilities*.

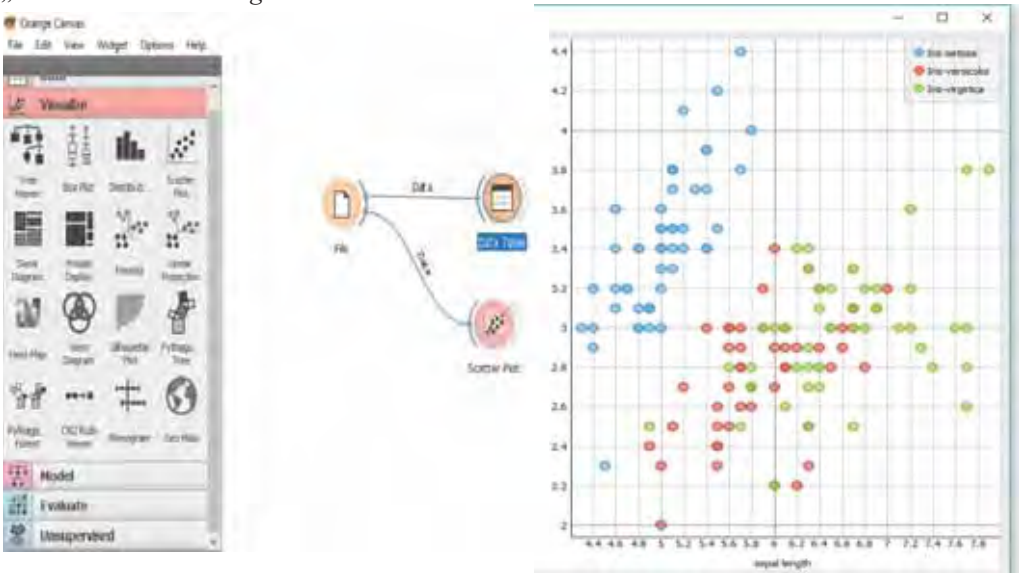
При *непрекъснати атрибути,* стойностите се показват като функционална графика. Класовите вероятности за непрекъснати атрибути се получават с оценка на плътността на гаусовото ядро, а появата на кривата се определя от лентата с инструменти – *Precision* (*smooth* or *precise*). В следващия пример, показан на фигура 5, отново се използва набора от данни за ирисите (Iris.tab), като е създадена графика с инструмент *Distributions* в системата „*Orange*“.




*Фигура 5. Функционална графика на данни с инструмент Distributions в „Orange“*

- ***Диаграми на разсейване*** (scatter plot) – полезни са, когато проверяваме каква е зависимостта между две различни величини.

Инструментът *Scatter Plot* осигурява двуизмерна визуализация, както за непрекъснати, така и за дискретни величини. Данните се показват като набор от точки, всеки, от които има ***x-axis*** стойност, определяща позицията по хоризонталната ос и ***y-axis*** стойност на атрибута, определящ позицията по вертикалната ос. На фигура 6 е представена визуализация на данните от файла за ирисите „Iris.tab“ чрез инструмент *„Scatter Plot*“ на *Orange*.




*Фигура 6. Визуализация на данните за ирисите с инструмент „Scatter Plot“ на Orange.*

Различните свойства на графиката, като цвят, размер и форма на точките, заглавията на осите, максималния размер на точките и трептенията, могат да се коригират чрез диалогов прозорец на инструмента. Системата може да използва визуализация на данните с опция *Find Informative Projections* (Start Evaluation). Функцията ще вър-

не списък с двойки атрибути за проекции в *scatter plot*, където случаите са добре разделени.

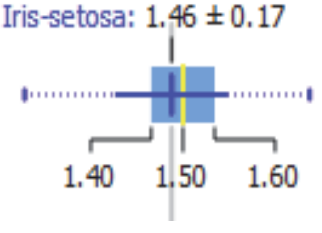
*Scatter Plot*, както и другите инструменти на *Orange*, поддържа възможности за приближаване и изключване на част от графиката, както и ръчен избор на данни. Тези функции са достъпни в долния ляв ъгъл на прозореца на инструмента. Инструментът по подразбиране е *Select*, който избира данни в област. *Pan* позволява да се премества *scatter plot* около избрания прозорец. С мащабиране може да се увеличава мащаба, докато *Reset zoom* нулира визуализацията до оптимален размер.

- ***Диаграми тип „кутия“*** (box plot)

Инструментът *Box Plot* на системата *Orange* показва разпределението чрез описателни статистики на извадката от данни и дава цялостна представа за данните:

- Средната (тъмносиня) вертикална линия представя модата.
- Светлосиния правоъгълник представя стандартното отклонение от модата (стандартно отклонение на средната стойност).
- Медианата е представена с жълтата вертикална линия.
- Непрекъснатата хоризонтална синя линия показва разликата между областта на първата част (от 25%) и останалата част (от 75%) от данните.
- Тънката пунктирана линия представя целия диапазон от стойности (от най-ниската до най-високата стойност в набора от данни за избрания параметър).

На фигура 7 е дадена примерна визуализация с инструмента, за ирисите от вид – *setosa.*




*Фигура 7. Визуализация на данни с инструмент Box Plot в системата Orange.*

Инструментът на Orange *Sieve Diagram,* построява диаграма с правоъгълници като площта на всеки правоъгълник е пропорционална на очакваната честота за избраната величина, докато наблюдаваната честота е илюстрирана чрез броя на квадратите в него. Разликата между наблюдаваната и очакваната честота се проявява в честотата на защриховане, използвайки цвят за индикация. Ако отклонението има положителна стойност се използва син цвят, ако е отрицателно – червен цвят.

- ***Стълбовидни диаграми*** (bar chart) – показват категории по едната ос и стойности по другата. Този вид диаграма е лесна за разбиране и се използва за сравнения между отделните променливи.

За атрибути с дискретни стойности отново използваме инструмента *Box Plot* в *Orange*. Но сега инструментът визуализира ленти (легнали стълбове), които представляват броя на случаите с всяка отделна стойност на атрибута. Нека заредим файла „*Zoo.tab*“ от галерията на *Orange*, в този случай *Box Plot* – инструментът (фигура 8) показва броя на видове животни в набора от данни за зоологическа градина.

*Фигура 8. Визуализация на данни с инструмента Box Plot в Orange.*

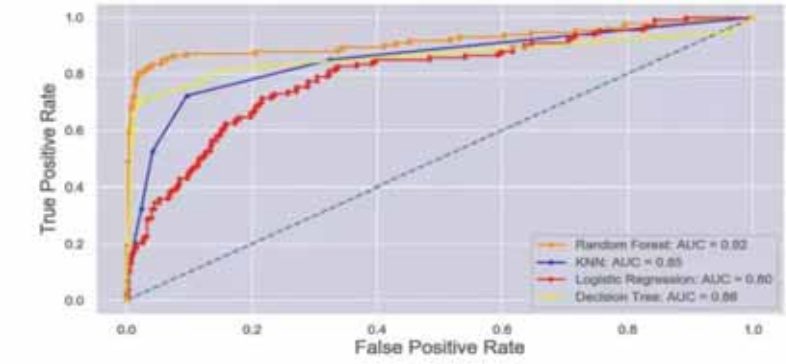
- ***Линейни диаграми*** – подходящи са, когато искаме графично да представим числова характеристика с непрекъснати стойности. Могат да бъдат полезни както и за правене на сравнения чрез визуализиране на повече от една линия.

В машинното обучение след създаване на модел на данните може да се оцени точността на предвиждане на модела, чрез инструментите *Confusion Matrix* и *ROC Analysis*. Работният поток е показан на фигура 9.


*Фигура 9. Работен поток за създаване на модели на данните и оценка чрез инструментите Confusion Matrix и ROC Analysis в Orange.*

*Receiver Operating Charactersitic* (ROC) кривата съпоставя TPR (True Positive Rate) и FPR (False Positive Rate), което дава информация за това до каква степен моделът правилно разпознава съответните класове. На фигура 10 е даден резултат от съпоставяне на четири различни класификатори. Колкото по-близо до горния ляв ъгъл е ROC кривата, толкова по-високо е качеството на класификатора. На графиката се вижда, че в конкретния случай за предвиждане, моделът, използващ алгоритъм *Random Forest* се справя най-добре.

Визуализацията на данни е изключително важна стъпка в машинното обучение. Чрез нея може да се даде представа за данните, която би била необходима в следващите етапи от изграждане на модела. Препоръчва се диаграмата да бъде възможно най-опростена и компактна. В някои случаи е по-добре да се представи множество от диаграми едновременно.




*Фигура 10. Визуализация с инструмента ROC Analysis в Orange.*

**Заключение**

Подготовката на данните е важен процес, който осигурява тяхната коректност, последователност и използваемост [6, 9]. Процесът на подготовка на данни, включва идентифициране на грешки, коригиране, изтриване, подредба или друга обработка, като се запазят само потенциално полезните данни, ръчно или с използване на софтуерни инструменти. Някои инструменти използват AI или машинно обучение, за да тестват по-добре точността на данните, идентифицират дубликати и спестяват време при анализа на данни. *Orange Data Mining* е базирана на компоненти, визуална среда за програмиране, която предоставя широк спектър от възможности за подготовка и обработка на данните.

**Литература**

[1] Bernard M.Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. John Wiley & Sons Ltd, New York, 2015.

[2] Fox G. Big Data HPC Convergence and a bunch of other things: http://www.slideshare.net/Foxsden/big-data-hpc-convergence-and-a-bunch-of-other-things

[3] A. Chapman, „Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data.“, Report for the Global Biodiversity Information Facility, Copenhagen, 2005.

[4] Six steps for data cleaning and why it matters, Published on November 20, 2020 in Maintenance by Leo Gimenez: www.geotab.com/blog/data-cleaning/

[5] Machine Learning: https://expert-bg.org/machine-learning-metriki-za-oczenka-na-klasifikaczionni-modeli/, последно достъпвано 20.08.2021.

[6] Orozova, D. Appropriate e-test system selection model, Comptes rendus de lAcademie bulgare des Sciences, Vol 72, No. 6, 811-820.

[7] Russell, S and P. Norvig. Artificial Intelligence: A Modern Approach, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[8] Venkatram K., Geetha Mary A., Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications, Cybernetics and Information Technologies, Vol. 17(2), 2017, 3-27, ISSN: 1311-9702;

[9] Popchev I., D. Orozova, Towards Big Data Analytics in the E-learning Space, Cybernetics and Information Technologies, Vol. 19(3), 2019, 16-24, ISSN: 1311-9702;

[10] Орозова, Д., С. Стоянов и И. Попчев, „Виртуално образователно пространство“ в Научна конференция с международно участие „Знанието – източник на иновация“, БСУ, Бургас,, 2013, pp. 153-159, ISBN 978-954-9370-99-7.

[11] https://orange.biolab.si. [Online]. https://orange.biolab.si/training/introduction-to-data-mining/