M. Kazionov, Bachelor student
O.S. Pasichnyk, PhD. in Education., As. Prof., language advisor
Khmelnytsky National University

DATA SCIENCE AND ITS APPLICATION IN HEALTHCARE

A well-organized healthcare system has become an essential part of our modern life. Without doctors, treatments, and disease prevention we'll return to the Stone Age. It is becoming obvious that only by a fundamental rethinking of our healthcare systems we can successfully address the serious challenges we are facing globally.

The tools for big data analytics and data science in medicine may vary, but the need drives technologies to evolve. An intricate net of different databases covers every aspect of the industry - from logistics to the genome structure. Each such database contributes to medical services in its way, and each of them requires data science tools to make the most of its contents [1].

For ages humanity was generating data. It is such an enormous amount that we can't even handle it and visualize properly. Our call history data, our movements, our internet behavior – all these things have become available and can have some value.

Data Science, or science how to operate data doesn't become a new fancy word in the IT world. It has become something bigger, something that can change the programming ecosystem, a business one, and even consumers. These changes can be compared to transformations caused by the invention of the personal computer or the steam engine. Data Science is already changing our world. As proof, we can see a lot of startups in the Big Data sphere or that of artificial intelligence.

Let's examine data scientist's activity for sustained analysis of required skills and knowledge. Data science involves a series of steps, one of them is – getting data. This process requires systemizing all the data, which must be analyzed. This can be taken from databases or different sources, for example extracting price value from one or more internet web-sites (this process is called web-scraping). Data can be provided in different formats, not only as numbers or text. This can be also images, videos, and sounds.[5]

Then the data cleaning stage comes. This process is needed to make data more human-friendly. This stage can be performed with such techniques as filling in empty values, creating new data groups, or getting rid of unnecessary data. In this process, you need to convert the data from one format to another to get one standardized format across all data.

The next stage is the data transformation. This stage can be used to make new useful groups of data, and to generate descriptive statistics. For example, if your data is stored in multiple CSV files, then you will consolidate these CSV data into a single repository so that you can process and analyze it [2].

After that, we have a very important step which is called data exploration. Primarily, we need to inspect the data and its properties. Different data types like

numerical data, categorical data, ordinal and nominal data, etc. require different treatments [2].

The next step is to compute descriptive statistics to extract valuable features and test significant variables. Testing significant variables often is done with correlation (heatmap plot). For example, exploring the risk of someone getting high blood pressure in relation to their height, weight, and other physical parameters. Do note that some variables are correlated, but they do not always imply causation.

The term "*Feature*" used in Machine Learning or Modelling, is the data features that help us to identify the characteristics that represent the data. For example, "*Name*", "*Age*", "*Gender*" are typical features of members or employees' dataset [3].

Lastly, we will utilize data visualization to help us identify significant patterns and trends in our data. We can gain a better understanding through simple charts like line charts or bar charts to help us understand the importance of the data.

The final stage is called modeling. In this stage, we are trying to answer the hypothesis questions we had got in the data exploration part. One of the first things you need to do in modeling data is to reduce the dimensionality of your data set. Not all your features or values are essential to describing your model and predicting its further development trends. What you need to do is to select the relevant ones that contribute to the prediction of results. After this process, you will need to be able to calculate evaluation scores (or measures) such as precision, recall and F1 score for classification. For regressions, you need to be familiar with R^2 to measure goodness-of-fit, and using error scores like MAE (Mean Average Error), or RMSE (Root Mean Square Error) to measure the distance between the predicted and observed data points [5].

Data Science technologies have already made some impact in fields related to healthcare: medical diagnosis from imaging data in medicine, quantifying lifestyle data in the fitness industry, just to mention a few.

Nevertheless, healthcare has been lagging behind in taking data analytics approaches, which is a paradoxical situation, since it was already estimated by the Ponemon Institute in 2012 that 30% of all the electronic data storage in the world was occupied by the healthcare industry [4].

It is evident that within the existing amounts of big data, there is hidden knowledge that could change the life of a patient or, to a very large extent, change the world itself. Extracting this knowledge is the fastest, least costly and most effective path to improving people's health [3].

Data Science technologies will definitely open new opportunities and enable breakthroughs related to, among others, healthcare data analytics addressing different perspectives: (1) descriptive, to answer what happened; (2) diagnostic, to answer the reason why it happened; (3) predictive, to understand what will happen next; and (4) prescriptive, to detect how we can make it happen [3].

Data analytics technologies could help provide more effective tools for behavioral change. Especially mobile health (mHealth) has the potential to personalize interventions, taking advantage of lifestyle data (nutrition, physical activity, sleep) and coaching style effectiveness data from large reference population

groups. Besides providing information to people, mHealth technologies exploit contextual information.

To prove the impact of Data Science and AI technologies on the healthcare sector, it is essential to apply these recommendations in large-scale pilots. The pilots are meant to serve as the best practice examples. Their objective is to demonstrate how the healthcare sector can be transformed with the aim to increase its quality, decrease costs and improve accessibility. This can be done by putting Data Science technologies at their core with the goal that their results can be scaled up and adopted by the whole healthcare sector.

REFERENCES

1. https://theappsolutions.com/blog/development/data-science-healthcare/
2. https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492
3. Sergio Consoli, Diego Reforgiato Recupero, Milan Petković, Data Science for Healthcare: Methodologies and Applications, Milan, 2019
4. https://www.nextech.com/blog/healthcare-data-growth-an-exponential-problem
5. http://www.deeplearningbook.org/
6. http://www.oecd.org/economy/growth/46508904.pdf