

Artificial Intelligence in Health Care: Predictive Analysis on Diabetes Using Machine Learning Algorithms

Shruti Wadhwa[1] and Karuna Babber[2(✉)]

[1] Nidus Technologies Pvt. Ltd., Chandigarh, India
[2] Post Graduate Government College, Chandigarh, India
karunababber.kb@gmail.com

Abstract. Background: The healthcare organizations are producing heaps of data at alarming rate. This data comprises of medical records, genome-omics data, image scan or wearable medico device data that presents immense advantages and challenges at the same time. These ever growing challenges can be surpassed by applying effective artificial intelligence tools.

Methods: This paper uses the large volume of multimodal patient data to perform correlations between Body Mass Index, Blood Pressure, Glucose levels, Diabetes Pedigree Function and Skin Thickness of people in different age groups with diabetes. Python and data analytic packages are used to predict diabetes among people.

Results: The blood pressure count of diabetic people comes around 75–85 mmHg and sometimes even higher whereas it is in the range of 60–75 mmHg for non-diabetic people. The people with high body mass index and glucose levels of 120–200 mg/dl and more are found to be diabetic as against the lower body mass index with glucose levels of 85–105 mg/dl of normal people. The Diabetes Pedigree Function count of diabetic people has a peak at 0.25 whereas it is 0.125 in case of non-diabetic people. A similar slight difference in values of Age and Skin Thickness has been found for both diabetic and non-diabetic people.

Conclusion: Above results indicate a strong relationship between Blood Pressure, BMI and Glucose levels of people with diabetes whereas a moderate correlation has been found between Age, Skin Thickness and Diabetes Pedigree Function count of people with diabetes. Although present analysis attested many of the previous research findings but getting these inferences matched through analytical tools is a sole purpose of this paper.

Keywords: Blood pressure · Body mass index · Diabetes Pedigree Function · Glucose · Skin thickness · Diabetes · Artificial intelligence

1 Introduction

Intelligent information can be a key for a healthier world. The more data we have, more optimally we can generate health specific information. These days we are surrounded with tons of data from almost every aspect of our lives be it societal, work, science or

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12250, pp. 354–366, 2020.
https://doi.org/10.1007/978-3-030-58802-1_26

health. In the technological world the term 'Big data' has been coined to describe such huge volumes of data. Douglas Laney [1] described big data by three V's that is Volume, Velocity and Variety wherein 'Big' part of big data indicates its large volume, velocity represents the rate and speed of data collection and variety stands for different types of organized and unorganized data. Over the time few other authors [2] added two more V's – Veracity and Variability into the big data definition. In the recent years almost every sector of research is analyzing the big data for various purposes. The most challenging task is to manage the heaps of structured and un-structured data [3] and to transform it into subject-oriented information. It is almost impossible to process big data with conventional software; we need technically advanced softwares with high-end computational power to make sense of such huge data. The implementation of Artificial Intelligence tools [4, 5] and algorithms can help us to generate decision-making information.

1.1 Big Data Repository for Healthcare

Healthcare is required at every stage of human life. Professionals come at the first place for primary care consultation, skilled professionals for secondary care, extra or intensive medical treatment comprises tertiary care and highly diagnostic and surgical procedures comes under the ambit of quaternary care [6, 7]. In the recent years multi dimensional model [8] of healthcare for diagnosis, prevention and treatment of health-related issues in human beings has been suggested. At all these levels the health specific accurate information is required at a single stop. So far heaps of health data has been collected from various sources and now it has come to brim where it becomes almost unmanageable with the current available technologies [9, 10]. Big data especially Artificial Intelligence can be viewed as a potential analyzer for improved health services. Figure 1 provides generalized analytic workflow for health care systems.

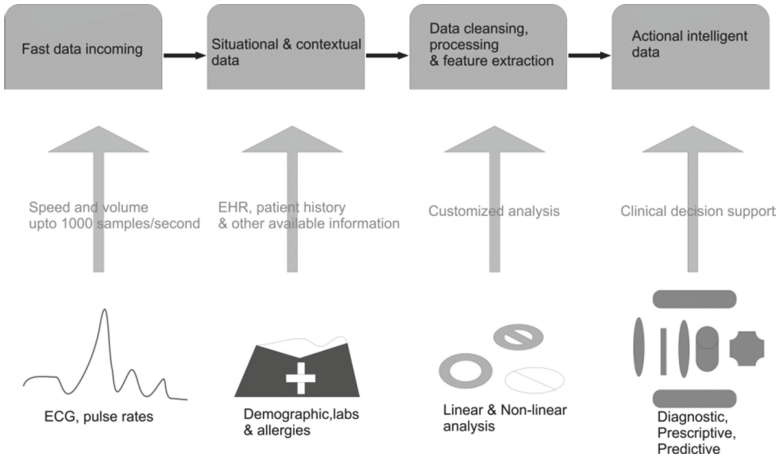

Fig. 1. Generalized analytic workflow for health care systems

1.2 Machine Learning and Healthcare

The machine learning in healthcare has opened plethora of applications [11] to improve patient life but errors in such applications may prove critical at times. A visual approach for 'Case based reasoning' mixed with quantitative and qualitative aspects have been proposed in [12]. The event sequencing of Electronic Health Record (EHR) of the patients [13] using deep learning algorithms are used to detect sepsis. The National Institutes of Health (NIH) of United States has recently started [14, 15] the 'All of Us' initiative (https://allofus.nih.gov) to collect more than one million of patients' data such as EHRs, medical images, environmental and socio-behavioural data. Data mining and machine learning is helping medical professionals make diagnosis easier by bridging the gap between huge data sets and human knowledge [16, 17]. A referential study in artificial intelligence in healthcare and medicine is provided in [18]. Intelligent machines may improve surgical procedures in the times to come [19, 20]. EHRs promise improvement in public health surveillance by timely reporting of disease outbreaks and can facilitate faster data retrieval for health insurance programs [21]. Similar to EHRs, Electronic Medical Records (EMR), Personal Health Records (PHR) and Medical Practice Management (MPM) software [22–24] and other related healthcare components collectively generate heaps of data that can be analyzed to provide real-time clinical care to patients in need. Figure 2 gives the fair idea about the framework for integrating multi-level information in order to provide personalized and cost-effective treatment to patients.

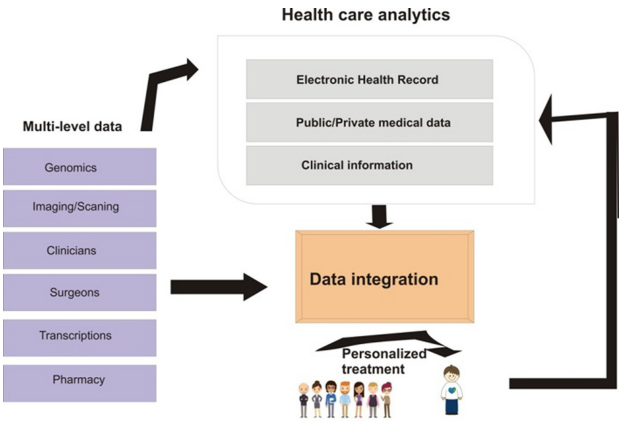

Fig. 2. Framework of healthcare data analytics

2 Data

The observed data has been part of the Pima Indians Diabetes Database collected by National Institute of Diabetes and Digestive Kidney Diseases [25]. The observational data contains different parameters like blood pressure, skin thickness, glucose, Diabetes

Pedigree Function (DPF) and Body Mass Index (BMI) of patients of all ages. Python and its data science related packages are used to carry on the analysis.

Firstly 'Pandas' [26] library file has been imported to read our data from a 'csv' file, then with the help of Extract, Transform, Load (ETL) tools [27] and 'Numpy' library file [28] data is transformed into a usable format to further use it on classification models. Figure 3 shows our dataset with 9 columns and 392 rows. The 'Matplotlib' library functions [29] have been used to draw graphs and other visualizations. The machine learning algorithms available in 'sklearn' library file [30] are used for final predictive analysis.

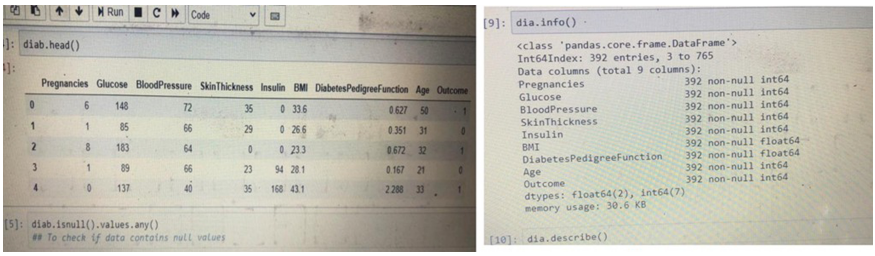

Fig. 3. Dataset with indexed parameters

To start with, we normalize the dataset to split it into two outcomes that is an outcome of zero (0) received for non-diabetic people and an outcome of one (1) received for diabetic people. The following bar graph (Fig. 4) gives us the fair idea.

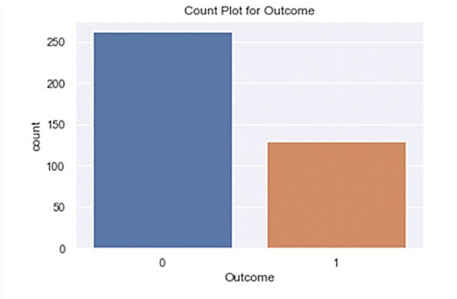

Fig. 4. Bar graph visualization of non-diabetic (0) and diabetic (1) people

From the above figure, it is clear that more than 250 people with '0' outcomes are non-diabetic whereas around 125 people with '1' outcome are diabetic. Now to make inferences of different parameters on diabetes, we have taken number of features into consideration, the details of which are provided in the results section.

3 Results

3.1 Blood Pressure (BP)

The measure of pressure of the blood in the circulatory system is defined as Blood Pressure [31]. In [32] authors explained effect of hypertension on diabetes mellitus. To find a correlation between blood pressure and diabetes, the bar graph (Fig. 5) visualization is provided below:

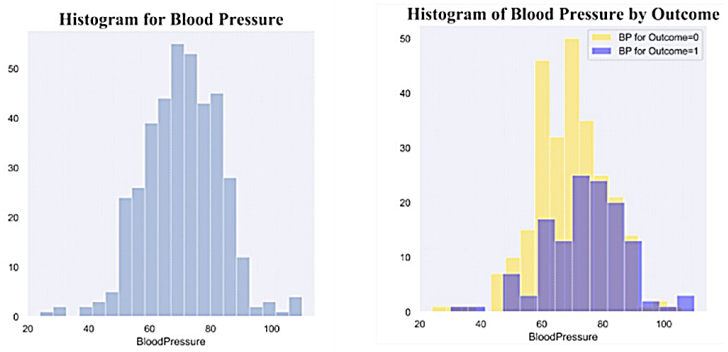

Fig. 5. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t. blood pressure

In graph I, the BP value is about normal with mean around 65–85 mmHg as against the ideal value of 80 mmHg. In graph II, for diabetic patients the BP value plot is little skewed towards right i.e. it is around 75–85 mmHg whereas for non-diabetic people it is around 60–75 mmHg.

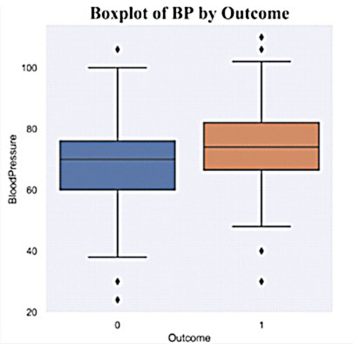

Fig. 6. Box-plots of non-diabetic (0) and diabetic (1) people w.r.t. blood pressure

The box-plot of BP in Fig. 6 shows 'whiskers' that is the maximum value of BP for non-diabetic people at 100 but it is at 110 with few outliers in case of diabetic people. Secondly around 75% diabetic people have BP in the range of 80–85 mmHg whereas in case of non-diabetic people around 75% people have a BP range of 70–75 mmHg. Both the plots indicate strong association between the two parameters.

3.2 Age of People

A person's age may play role for Type 2 Diabetes Mellitus. The risk of type 2 diabetes increases with the rising age [33, 34]. In graph I (Fig. 7), people within the age group of 20 or in their mid 20 s are non-diabetic whereas after the age of 30 and above people are prone to diabetes. In graph II, the median is at 25 years for non-diabetic people but the median for diabetic people is at 32 years. Both the graphs indicate a moderate correlation between the two features.

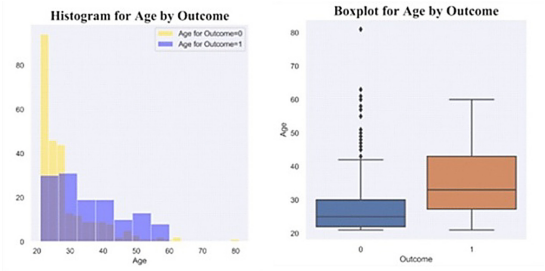

Fig. 7. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t. age

3.3 Body Mass Index (BMI)

The body mass index [35] is a person's weight in kilograms (kg) divided by his/her height in meters (m). The National Institutes of Health (NIH) has adopted BMI as a parameter to define normal weight, overweight and obesity of people rather than the traditional height/weight charts. The graph I (Fig. 8), shows the BMI level of all the people within the different age groups. In graph II, the BMI level peak is from 28–32 for non-diabetic people whereas it is skewed and tilted towards right at 36 for diabetic people.

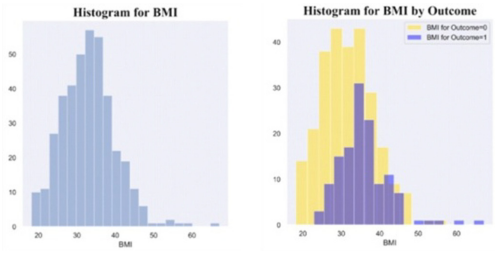

Fig. 8. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t BMI

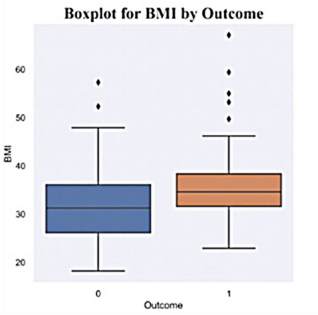

Fig. 9. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t BMI

In the box-plot (Fig. 9) more than 50% diabetic people have high BMI level as against the non-diabetic people. Secondly for diabetic people some outliers are beyond maxima whereas it is not the case in non-diabetic people. The graph visualizations indicate strong correlation between the two parameters.

3.4 Diabetes Pedigree Function (DPF)

It [36] is a function which scores likelihood of diabetes based on family history. Two histograms for DPF of both the outcomes (0 and 1) are presented in Fig. 10. Graph I shows the DPF count of all the people within the different age groups. In graph II the peak of DPF count comes at 0.125 for non-diabetic people whereas it is at 0.25 for diabetic people. The graph indicates nominal relation between two parameters.

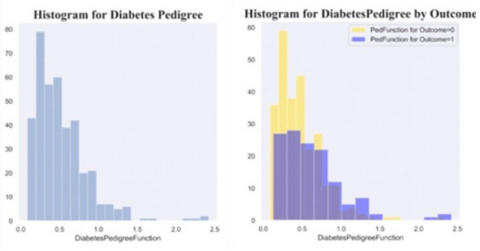

Fig. 10. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t DPF

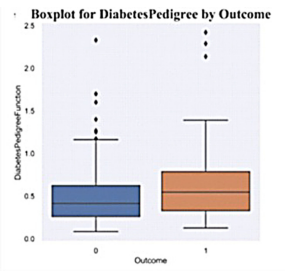

Fig. 11. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t DPF

The box-plot in Fig. 11 shows more than 50% non-diabetic people have a DPF count of 0.4 and it is slightly more with 0.525 DPF count for diabetic people. Secondly both the plots have outliers beyond maxima. Therefore, all the graphs indicate a weak correlation between the two features.

3.5 Skin Thickness

The skin thickness [37] is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus. The graph I of Fig. 12 provides the skin thickness of both the outcomes (0 and 1). In graph II the peak is at 30 for non-diabetic people whereas it is at 40 for diabetic people. The peak values indicate a moderate relation between the two parameters.

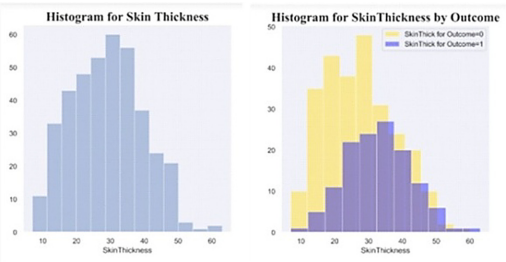

Fig. 12. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t skin thickness

In the box-plot (Fig. 13) more than 50% non-diabetic people have a skin thickness count of 28 whereas the skin thickness count is 34 for more than 50% diabetic people. Secondly very few outliers are beyond the maxima in both the cases and the difference between the maxima value for both diabetic and non-diabetic people is only 3 points. The graphs indicate nominal relation between the two features.

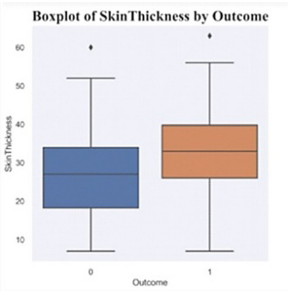

Fig. 13. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t skin thickness

3.6 Glucose Level

The global mean fasting plasma blood glucose level [38] in humans is about 100 mg/dl. The graph I of two histograms in Fig. 14 provides the glucose level of both the outcomes (0 and 1). The graph II shows the glucose level range of non-diabetic people between 85–105 mg/dl with little spike at 100–120 mg/dl whereas the range for diabetic people comes in between 120–200 mg/dl.

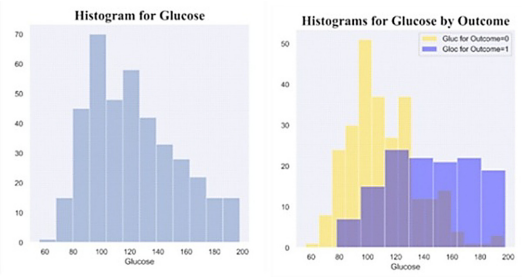

Fig. 14. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t glucose

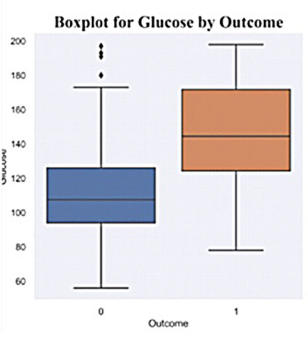

Fig. 15. Bar graph visualization of non-diabetic (0) and diabetic (1) people w.r.t glucose

In the box-plot (Fig. 15) more than 50% of non-diabetic people have glucose level 100–110 mg/dl. But more than 50% diabetic people have a range above 140 mg/dl. Secondly the maxima value for non-diabetic people is 170 mg/dl whereas it is 200 mg/dl in case of diabetic people. All the above graphs indicate a strong correlation between these two parameters.

4 Result Analysis and Discussion

The above provided results were analyzed on few machine learning algorithms and it is found that Logistic Regression, Support Vector Classifier (SVC) and Linear Support Vector Classifiers (LSVC) have performed well with higher mean accuracy rate as compared to Decision Tree, Gaussian Naive Bayes and Random Forest Classifiers [39–41]. The box-plot (Fig. 16) provides the Inter Quartile Range (IQR) of the mean accuracy rate of machine learning algorithms and the Table 1 has detailed readings of the actual mean of accuracy along with standard deviation of machine learning algorithms.

Table 1. Machine learning algorithms with mean of accuracy and standard deviation readings

Machine learning algorithm	Mean of accuracy	Standard deviation
Logistic Regression (LR)	0.789415	0.060090
K Neighbors Classifier (KNN)	0.766935	0.058651
Gaussian NB (NB)	0.767238	0.049213
SVC	0.782964	0.077649
Linear SVC (LSVC)	0.783065	0.062417
Random Forest Classifier (RFC)	0.750706	0.065619
Decision Tree Regressor (DTR)	0.706250	0.094293

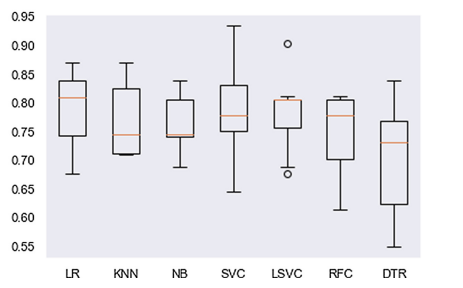

Fig. 16. Performance of machine learning algorithms w.r.t different parameters on outcome variable (1)

By applying the thumb rule of correlation and inferences received by applying machine learning algorithms, we can infer that BP, BMI and Glucose levels of people showed strong correlation with diabetes whereas DPF count, Age and Skin thickness have signalled moderate correlation with diabetes.

5 Conclusion

With the privilege of machine learning algorithms, big data health analytics has opened doors for new predictive systems. On the similar lines, this paper presents a predictive analysis on diabetes. We found strong correlation of diabetes with Blood Pressure, BMI and Glucose levels of people. But Age, Skin Thickness and Diabetes Pedigree Function count have shown moderate relationship with diabetes. Though we may have conventional clinical reports on correlation of these parameters with diabetes but getting these inferences matched through analytical tools is a sole purpose of this paper. Through this paper we successfully predict and visualize the correlations between different parameters on diabetes.

References

1. Laney, D.: 3D Data Management: controlling data volume, velocity and variety, Application delivery strategies. META Group Inc., Stanford (2001)
2. Mauro, A.D., Greco, M., Grimaldi, M.: A formal definition of big data based on its essential features. Libr. Rev. **65**(3), 122–135 (2016)
3. Jacobs, A.: The pathologies of big data. Commun. ACM **52**(8), 36–44 (2009)
4. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
5. Belle, A.: Big data analytics in healthcare. Biomed. Res. Int, E **2015**, 370194 (2015)
6. Viceconti, M., Hunter, P., Hose, R.: Big data, big knowledge: big data for personalized healthcare. IEEE J. Biomed. Health Inf. **19**, 1209–1215 (2015)
7. Nasi, G., Cucciniello, M., Guerrazzi, C.: The role of mobile technologies in health care processes: the case of cancer supportive care. J. Med. Internet Res. **17**(2), e26 (2015)
8. Bali, J., Garg, R., Bali, R.T.: Artificial intelligence in healthcare and biomedical research: why a strong computational bioethics framework is required. Indian J. Ophthalmol. **67**(1), 3–6 (2019). https://doi.org/10.4103/ijo.ijo_1292_18
9. Doyle-Lindrud, S.: The evolution of electronic health record. Clin. J. Nurs. **19**(2), 153–154 (2015)
10. Yin, Y.: The internet of things in healthcare an overview. J. Inf. Health Care **1**, 3–13 (2016)
11. Khare, A., Jeon, M., Sethi, I.K., Xu, B.: Machine learning theory and applications for healthcare. J. Healthcare Eng. (2017). https://doi.org/10.1155/2017/5263570
12. Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Seroussi, B.: Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. J. Artif. Intell. Med. **94**, 42–53 (2019). https://doi.org/10.1016/j.artmed.2019.01.001
13. Lauritsen, S.M., et al.: Early detection of sepsis utilizing deep learning on electronic health record event sequences. J. Artif. Intell. **104** (2020). https://doi.org/10.1016/j.artmed.2020.101820

14. Shameer, K.: Traditional bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief. Bioinform. **18**(1), 105–124 (2017)
15. Reform. Thinking on its own: AI in the NHS (2018)
16. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthcare J. **6**(2), 94–98 (2019). https://doi.org/10.7861/futurehosp.6-2-94
17. Jiang, F., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. **0**, e000101 (2017). https://doi.org/10.1136/svn-2017-000101
18. Tran, B.X., et al.: Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J. Clin. Med. **8**, 360–369 (2019). https://doi.org/10.3390/jcm8030360
19. Moore, S.F.: Harnessing the power of intelligent machines to enhance primary care. J. General Pract. **68**, 6–7 (2018)
20. Nesta. Confronting Dr. Robot creating a people-powered future for AI in health (2018)
21. Reisman, M.: EHRs: the challenge of making electronic data usable and interoperable. J. Pharma Theor. **42**(9), 572–575 (2017)
22. De Moor, G., et al.: Using electronic health records for clinical research: the case of the EHR4CR project. J. Biomed. Inf. **53**, 162–173 (2015). https://doi.org/10.1016/j.jbi.2014.10.006
23. Kaelber, D.C., Jha, A.K., Johnston, D., Middleton, B., Bates, D.W.: A research agenda for Personal Health Records (PHR). J. Am. Med. Inf. Assoc. **15**(6), 729–736 (2008). https://doi.org/10.1197/jamia.m2547
24. Roehrs, A., da Costa, C.A.: Personal health records: a systematic literature review. Med. Internet Res. Nat. Libr. Med. **19**(1), e13 (2017). https://doi.org/10.2196/jmir.5876
25. https://www.kaggle.com/uciml/pima-indians-diabetes-database/
26. https://pandas.pydata.org/
27. Parker, E.: Python & ETL 2020: A list and comparison of the top python ETL tools (2020)
28. https://numpy.org/
29. https://matplotlib.org/
30. https://scikit-learn.org/stable/
31. William, C., Shiel Jr.: Medical definition of blood pressure. J. Medterms Med. Dic. (1998)
32. Tsimihodimos, V., Clicerio, G.-V., Meigs, J.B., Ferrannini, E.: Hypertension and diabetes mellitus. J. Hypertens. **71**, 422–428 (2018). https://doi.org/10.1161/HYPERTENSIONAHA.117.10546
33. Suastika, K., Dwipayana, P., Semandi, M.S., Kuswardhani, R.A.T.: Age is an important risk factor for type 2 Diabetes Mellitus and Cardiovascular Diseases. IntechOpen (2012). https://doi.org/10.5772/52397
34. Kharroubi, A.T., Darwish, H.M.: Diabetes mellitus: the epidemic of the century. World J. Diabetes. **6**(6), 850–867 (2015). https://doi.org/10.4239/wjd.v6.i6.850
35. Nuttall, F.Q.: Body mass index: obesity and health – a critical review. J. Nutr. Today **50**(3), 117–128 (2015)
36. Mercaldo, F., Nardone, V., Santone, A.: Diabetes Mellitus affected patients classification and diagnosis through machine learning techniques. In: International Conference on Knowledge Based and Intelligent Information and Engineering Systems (2017). Procedia Computer Science, vol. 112, pp. 2519–2528
37. Jose Derraik, G.B., Rademaker, M., Cutfield, S.W., Pinto, E.T.: Effects of age, gender and anatomical site on skin thickness in children and adults with diabetes. PLoS One **9**(1) (2014). https://doi.org/10.1371/journal.pone.0086637
38. Frankum, S., Ogden, J.: Estimation of blood glucose levels by people with diabetes: a cross-sectional study. Br. J. Gen. Pract. **55**(521), 944–948 (2005)
39. Biau, G.: Analysis of random forests model. J. Mach. Learn. Res. **13**, 1063–1095 (2012)

40. Marijana, Z.-S., Sarlija, N., Has, A., Bilandzic, A.: Predicting company growth using logistic regression and neural networks, Cortian Oper. Res. Rev. (CRORR), **7**, 229–248 (2016)
41. Wu, D., Jennings, C., Terpenny, J., Gao, R.X., Kumara, S.: A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using Random Forests. J. Manuf. Sci. Eng. **139**(7) (2017). https://doi.org/10.1115/1.4036350