DataOps Lifecycle with a Case Study in Healthcare

Shaimaa Bahaa[1*], Atef Z.Ghalwash[2], Hany Harb[3]
Computer Science Dept., Misr University for Science & Technology (MUST)
Giza, Egypt [1, 3]
Computer Science Dept., Helwan University
Cairo, Egypt [2]
https://orcid.org/0000-0002-3458-0395

***Abstract*—The DataOps methodology has become a solution to many of the difficulties faced by data science and analytics projects. This research introduces a novel DataOps lifecycle along with a detailed description of each phase. The proposed cycle enhances the implementation of data science and analytics projects for achieving business value. As a proof of concept, the new cycle phases are applied in a healthcare case study using the UCI Heart Disease dataset. Two goals are achieved. First, a dataset reduction by features analytic in which the four most effective features are selected. Second, different machine learning algorithms are applied to the dataset. The recorded results show that using the four most effective features is comparable with using the full features (thirteen features), and both approaches show high accuracy and sensitivity. The average accuracy of the highest four features is 82.32%, and the thirteen features is 84.28%. That means that the selected four features affect the applications with 97.67% accuracy. Besides, the average sensitivity of the highest four features is 87.94%, while the thirteen features are 87.12%. The study shows an interesting and significant result that data modeling needn't be done for all data science projects which reduced the dataset.**

***Keywords*—DataOps lifecycle; DataOps in machine learning; DataOps in healthcare; DataOps in data science; feature extraction; feature selection**

I. INTRODUCTION

The fields of data science, analytics, and machine learning are expanding at an incredible rate. Businesses are now searching for experts who can sort through the data goldmine and assist them in making quick, informed business decisions. Although today, organizations have a great opportunity to access data-driven tools and business intelligence software. Most organizations fail to make business value from their investments in data [1]. Thus, resulting from the lack of maturity in data science projects, most implementations are laptop-based research projects that never impact customers. In addition to, local applications that are not built to scale for production workflows, or high-cost IT projects. Therefore, selecting the method for implementing a data-driven project must be done carefully to help when maintaining or even adding a new feature(s).

The legacy architecture and tools that require special skills to use by data scientists have become bottlenecks for business stakeholders. These tools are costly resources, especially when producing unplanned data analysis. Machine Learning and Artificial Intelligence algorithms are just tips of the iceberg [2] for getting business and customer value from data. Therefore, the operation of affecting data is the most crucial aim.

Data analytics is used in business to help organizations make better business decisions to meet and increase customer value. Data draws beneficial conclusions by collecting and organizing it; a data-driven process covering everything from data collection to analysis. DataOps has emerged to meet such requirements. *DataOps* is an emerging set of practices, processes, and technologies for building and enhancing data and analytics pipelines [3]. The term DataOps is a merge of data and operations which was first introduced by Lenny Liebmann in a 2014 blog post titled "3 reasons why DataOps are essential for big data success". The term wasn't popularized until Andy Palmer's 2015 blog post "From DevOps to DataOps". Since then, interest has grown when the term DataOps was included in Gartner's "Hype Cycle" for data management in 2018 [4].As Agile has a manifesto [5] for it4s principles. DataOps has its own manifesto [6] too, which consists of 18 principles, unlike Agile 12 principles. The DataOps manifesto has been published by Christopher Bergh, Gil Bengiat, and Eran Strod [4]. The DataOps manifesto principles have complemented the initiative that came out in 2018 called "The DataOps Philosophy".

The problem DataOps has come up for solving and minimizing analytics "cycle time" between the proposal of a new idea and the deployment of finished analytics. For example, many organizations require months of cycle time to deploy 20 lines of SQL. The long cycle times are the primary reason analytics projects fail [7]. This has led to discouraged and disappointed users and disturbing creativity. The factors that lengthen cycle time are *Poor Teamwork, Lack of Group Cooperation, Waiting for Systems, Waiting for Data Access, Over-Caution, Requiring Approvals, Inflexible Data Architecture, Process Bottlenecks, Technical Debt and Poor Quality*, which were mentioned in [8]. These obstacles pushed data experts to find an effective solution; therefore, DataOps came up. DataOps's goal for data science is to turn unprocessed data into a useful data science product. DataOps has provided utility to customers through a rapid, scalable, and repeatable process.

Data experts have given DataOps many definitions depending on their points of view. As a result, there have been several attempts to define the concept of DataOps. For example, Gartner [9] defined DataOps as a collaborative data management practice focused on improving the communication, integration, and automation of data flows

between data managers and data consumers across an organization. While Eckerson [10] Group defined DataOps as an engineering methodology and set of practices designed for the rapid, reliable, and repeatable delivery of production-ready data, operations-ready analytics, and data science models. DataKitchen [11] said that DataOps is a collection of technical practices, workflows, cultural norms, architectural patterns, and much more. However, the most appropriate definition adopted here is that DataOps is a methodology that applies Agile development, DevOps, and lean manufacturing principles [12], all together to data analytics development and operations where they are the intellectual heritage for DataOps.

Agile is an application of the theory of constraints to software development, in which smaller lot sizes decrease work-in-progress and increase overall manufacturing system throughput. DevOps is a natural result of applying lean principles, for example, eliminating waste, continuous improvement, and broad focus on application development and delivery. Lean manufacturing also contributes a relentless focus on quality using tools such as statistical process control, to data analytics [11]. Due to different DataOps definitions, trying to evaluate different solutions and determine whether they will help to achieve DataOps goals or not is a confusing matter.The authors in [13] introduced a DataOps Vendors Landscape which was organized by the six key capabilities required for DataOps success.

The major contributions of this paper are summarized in the following points:

- Introducing a novel approach for the DataOps lifecycle with a detailed description for each phase.
- The most effective features selection and extraction.
- Proving that data modeling is not necessary for all data science and analytics projects.
- Presenting a case study in healthcare as a proof of concept.
- Dataset reduction for the UCI Heart Disease dataset.
- A comparison between different machine learning algorithms that have applied to the dataset for both the highest four features and all (thirteen) features.

This paper is organized as it follows. Section II presents the related work. Section III introduces the proposed DataOps lifecycle. Experiments and recorded results for a case study in healthcare are shown in Section IV. Section V has the conclusion and future work.

II. Related Works

This section will investigate both DataOps related works and the UCI Heart Disease dataset related works. The number of DataOps related works are quite a few because of being a new research field. The author in [14] has illustrated the broad character of DataOps and shown that it is not a particular method or tool. However, a collection of principles and a way of doing things on a cultural, organizational, and technological level. He differentiated between the exploration of DataOps as a discipline, which includes methods, technologies, and concrete implementations, and the investigation of the business value of DataOps. While authors in [15] have defined DataOps as an application of DevOps to data, which means how effective data operations can be when DevOps concepts are applied to data for managing and deriving analytics. They also outlined the DataOps process and platform as well as the data challenges in the manufacturing and utilities industries. In [16], the authors said that DataOps is a new approach that aims to improve the quality and responsiveness of the data analytics lifecycle. In addition, they broadly organized dataOps into three steps: build, execute, and operate.

The lifecycle of a DataOps process has been illustrated in [17]. Besides, it illustrates the main collaborators in the DataOps process in charge of generating business value. In addition, they have gathered and highlighted good practices in DataOps reported in the academic literature, which serves as a starting point. While [18] defines DataOps as a method for accelerating the delivery of high-quality results through automation and orchestration of data life cycle phases. Furthermore, a case study in collaboration with Ericsson was conducted and introduced. They used the key phases of the data analysis methods to explore the key phases of the data besides checking their similarities to the popular DataOps approach. The common limitations of the above related works were either ambiguity of the DataOps lifecycle or the shortage of applications.

In [19], the model they proposed has four phases: first, data gathering that was the UCI Machine Learning dataset. Second, they used two methods for the features selection: Pearson's Correlation Heatmap where they selected 9 features and Chi Squared Test that selected 6 features. The third stage consists of K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Decision Tree (DT). After applying those algorithms, they have further used Stacking and Voting ensemble techniques for better results. Although, their model performed better when they have used Pearson's Correlation Heatmap selected features. Their model has some limitations that it has taken more time to generate outcomes.

While [20], they used Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Xtreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM), and K-Nearest Neighbors (KNN) for the prediction of the UCI Machine Learning dataset individually. Then multi-model ensembles were created (Ensemble 1 and Ensemble 2), which have far higher accuracies than individual models. The models with the best values of the evaluated parameters were gathered. In order to train and test the models on five distinct folds and to determine the optimal values for the hyperparameters in each of the implemented classification algorithms, fivefold cross validation and GridSearchCV were employed. They used all the UCI Machine Learning dataset features in addition to, they grouped models to reach their accuracy and that were their limitations.

Also [21], chose the well-known Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest

Neighbor (K-NN) algorithms for the UCI Machine Learning dataset prediction; also, for Hungarian cardiovascular disease dataset. The proposed system consisted of data acquisition, pre-processing, feature/attribute selection, classifications, and performance evaluation. The FCBF and mRMR were the feature selection algorithms. They used the info gain function selection method that's available on Weka for actual feature ranking. The top 10 features were ranked, which was a limitation.

Authors in [22] present three approaches. First Approach was without doing feature selection and outliers detection. The second approach was with doing feature selection and no outliers detection. The third approach was doing feature selection and also outliers detection. In all approaches they used Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and XGBoost ML algorithms for the UCI Machine Learning dataset prediction. Their results they have been reached needn't to do three approaches. In addition, the feature they chose to be either important or not for heart disease predication were a limitation.

III. Proposed DataOps Lifecycle

Data analytics projects are failing despite qualified people, powerful tools, and huge investments. Data scientists spend 75% [23] of their time massaging data and executing manual steps. Slow and error-prone development disappointed and frustrated data team members and stakeholders.

According to the adopted definition, to manage data in accordance with corporate objectives, DataOps combines DevOps and Agile approaches. For instance, DataOps would position data to make recommendations for better product marketing, converting more leads, if the goal was to increase lead conversion rate. While DevOps procedures are utilized for code optimization, product builds, and delivery, Agile techniques are employed for data governance and analytics development. DataOps uses statistical process control (SPC) to continuously monitor and verify the data analytics pipeline, much like lean manufacturing does. SPC increases data processing efficiency, improves data quality, and ensures that statistics are kept within reasonable bounds. SPC helps to notify data analysts right away in the event of an anomaly or error so they can respond.

This study proposes a novel DataOps lifecycle as shown in Fig. 1, along with a detailed description for each phase. The significance of this cycle is to investigate and highlight that data modeling is not necessary for all data science and analytics projects that reduce the dataset. The following subsections will illustrate each phase in detail.

A. Define Data Domain

A common mistake in data science projects is the confusion between defining a data area and a data domain. A data domain is a specific area in a large area. For example, if we have a data science project in healthcare, the area is healthcare, while the domain is heart disease. This is called the first data domain definition. The second data domain definition is what disease in the heart disease area we're going to work with.

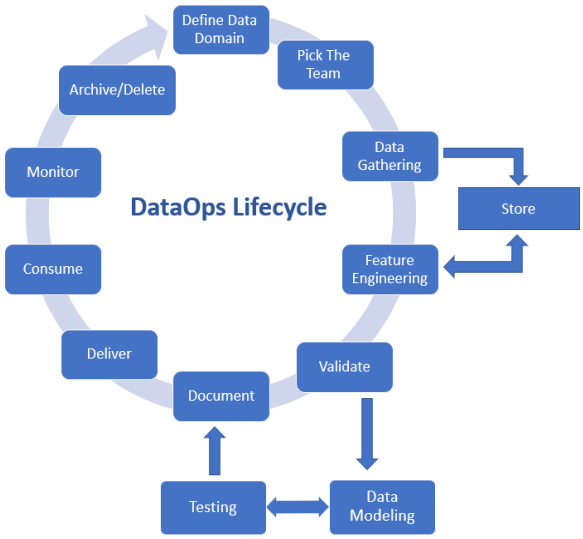

Fig. 1. Proposed DataOps lifecycle

B. Pick the Team

From this study point of view, this phase is the most critical phase that will determine the success of the coming phases or not. Data scientists have been working on different tools and outputting the results in different forms. This common problem adds another step to reorganize the output and merge all the results. Table I defines the criteria that must be met to approve each team member and the team as a whole.

Picking the team may differ from task to task. So, a smaller team would be selected by the unicorn for simpler problems. There are two main types of tasks, and each of them will be handled differently: First, system creation from scratch including data flow, data architecture, data schema, integration, etc. Second, solve smaller tasks from working with existing data or data that can be obtained. Each will require different handling methods and will be taken into consideration during the selection process.

C. Data Gathering

After everything was agreed upon, the team should start data gathering related to the domain. Data may be ready-made or performed from scratch. This phase should be done carefully as it may need to include any additional data, like performing surveys or scraping some data from the web. Then, the collected or performed data has to be stored.

D. Feature Engineering

Because not all data is important, this phase is concerned with obtaining and extracting the important data features. So, based on the data domain expert and, of course, the unicorn's opinion, the team tries to extract only the important features.

TABLE I. TEAM CRITERIA

Criteria	Criteria Details	
Members and Tasks	1. Data Unicorn	Data scientist who has the knowledge for all aspects of data science projects as, data engineering, statistical analysis, business analysis, ML, programming, or visualization. He must approve all the other members and he manages the whole project.
	2. An expert of domain knowledge	Any data project must have an expert in domain knowledge (with no experience in SWD) to explain all domain details and tricks the team want to know, i.e., educational expert in education project and medical expert in medical project.
	3. Data Analyst	Analyzes data (i.e., Visualizations: Charts, Graphs, Dashboards, Tables, Reports).
	4. Data Engineer	Develop, constructs, tests & maintain complete data architecture (i.e., Schema design, Data lakes).
	5. Data Scientist	Analyzes and interprets complex data in addition, he is a data wrangler who organizes (big) data.
	6. DataOps Engineer	He creates the mechanisms for workflow, manages cycle time and optimizes the quality.
	7. AI Engineer	Create end-to-end applications that include the data model(s).
	8. Operations Engineer	Deploying the applications into production environments and support service-level agreement.
Common Knowledge	They all must know source control, containerization, clean code, design pattern, security (to some level).	
Tools	All members must work with the same tool, or different tools that their output format is the same (provided that they will unify this later).	
Agreement	They must respect the data security and environment ethics.	

The next phase will be stepped to get the insights. The featured data may need to be stored or to store the featured criteria that have been adopted for similar data.

E. Store (Optional)

After collecting the data, it must be stored initially. Then, it might be restored or stored again in some other format after the next phase is done. For the sake of memory, the team might choose to store the data again or override the stored data based on its importance and size.

F. Validate

In this phase, the team seeks to gather all the collected insights. The task goal will be tried to reach based on the provided information. If they did, they wouldn't have to do the data modelling phase as results might be easily concluded only from the insights relying on the experience of the team. Data modelling leads to a lot of time, money, and resources; therefore, avoiding it on small tasks increases the speed. So, based on the problem type and the data, the team may also select a collection of algorithms and combine them together to obtain the best result for reaching the task goal. Thus, the team must successfully manage this phase in order to avoid the next if possible.

G. Data Modeling

If this phase is applied in the proposed cycle, it means that the team has tried every possible scenario to understand the data. Nevertheless, it was rather too complicated or critical for data modelling to be avoided. The team then must select the most perfect and appropriate algorithm for data modeling.

H. Testing

After reaching the goal, which of course, may differ from one task to another, in which data tasks need to be analyzed, visualized, modeled, processed, stored, or some of them, or all of them at the same time. The team must test whether the results match the original goal or not. It must be checked step by step, there is no overfitting, underfitting, or any other data-related problems. The testing phase also must be reproducible. As the data modelling phase may be required to resolve a tested issue(s).

I. Document

All the work that has been done must be well described and documented. The documentation must be self-explanatory so that any member of any department can easily understand what has been done. Documentation may also involve technical writing for developers to complete from where the team stopped. Besides, if new data is created, a new data schema must be written to fairly describe the new changes. In addition to, the reasons why they needed to change the original data must be defined.

J. Deliver

The task delivery could be a tricky process. If the task contains new data, it must be put in the right place without any contradiction with the original data or any other data. If it contains reports, it must be very readable with visualizations to make the image clearer. Finally, if it has a new model delivered, it must be uploaded to the place where it will be used. So, an API could be created or maybe a model with a specific format. Therefore, the delivery must be teamwork to get the results in the right format to be used and easily deployed in production.

K. Consume

At an enterprise level, the published model(s) can be reused to derive various analytics required for business. “Recommend” solutions for understanding consumer preferences can be applied to a host of product lines. The tested and deployed solutions can be used with similar data sets to solve similar problems. This way not only does the enterprise save time by quickly applying proven

methodologies, but it also ensures the building of robust solutions through the continuous evolution process.

L. Monitor

Tasks delivery and doing investigations is not the end of the work. All the tasks delivered must be monitored to see whether the team succeeded in doing reliable work in the long term or not. For example, the team may have been assigned to work on a part that is rarely used. Besides, the team may have focused on the wrong features at first, and it was so obvious. Furthermore, monitoring is essential to predict and avoid an immediate system failure or even a small failure. Expect any changes before they happen. Detecting any possibility of performance reduction and being prepared for what changes may come may be enough. Thus, monitoring is essential for further improvement.

M. Archive/Delete

By the end, the organization might decide to cull out unwanted data or archive it to optimize resources and size management as well. Periodic data audits should be carried out to ensure production systems use fewer resources, running more efficiently and reducing storage costs overall. Data archiving plans have to be made for easy retrieval and more cost-effective information storage. Furthermore, irrelevant data needs to be purged.

VI. Experiments and Results

This section introduces a healthcare case study for the proposed DataOps lifecycle. Healthcare was chosen as it is the most sensitive data in which the results affect human life. According to WHO [24], cardiovascular diseases are the leading cause of death globally. It takes an estimated 17.9 million human lives each year. So, the chosen dataset was the UCI Machine Learning Heart Disease dataset [25]. This dataset contains 76 attributes, but all published experiments refer to using a subset of 14 (13 feature and target column) of them.

The "goal" which is the target, refers to the presence of heart disease in the patient. It is an integer valued between 0 (no presence) and 1 (presence). The other 13 features are:

1) **age,** that stores the age of the patient in years.

2) **sex,** where 1 is for males and 0 for females.

3) **cp,** chest pain type in which 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, and 4 = asymptomatic.

4) **trestbps,** which is resting blood pressure that was measured in mm Hg on admission to the hospital.

5) **chol,** serum cholestoral measured in mg/dl.

6) **fbs,** is fasting blood sugar > 120 mg/dl if 1 then true, while 0 = false.

7) **restecg,** resting electrocardiographic results that has three values 0 indicates normal, while 1 indicates having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) and 2 showing probable or definite left ventricular hypertrophy by Estes' criteria.

8) **thalach,** includes the maximum heart rate value achieved.

9) **exang,** exercise induced angina where 1 means yes and 0 means no.

10) **oldpeak,** that equals ST depression induced by exercise relative to rest.

11) **slope,** which is the slope of the peak exercise ST segment that has 3 values, if 1: upsloping, 2: flat, and 3: downsloping.

12) **ca,** includes a number of major vessels (0-3) coloured by flourosopy.

13) **thal,** where 3 = normal, 6 = fixed defect, and 7 = reversible defect.

The proposed DataOps lifecycle shown in Fig. 1 was applied as follows:

A. Define Data Domain

As illustrated in Section 3.1, the data area in this case study is healthcare. But the data domain is cardiovascular diseases.

B. Pick the Team

For this stage, as explained in Section 3.2. Choosing the team may differ from task to task. Therefore, in this case study, two main members from Table I must be present, in which they are an unicorn and a domain knowledge expert at least.

C. Data Gathering

After starting data gathering related to the domain. The UCI Machine Learning Heart Disease dataset has been chosen.

D. Feature Engineering

This phase has been done using Python 3 and the Jupyter Notebook IDE. Feature engineering has been done as follows:

- **Step 1:** Read/Load dataset.
- **Step 2:** Get the dataset information. The dataset information has displayed features(columns) name, each feature datatype, datatype, and total number of rows and columns.
- **Step 3:** Checking for null values.
- **Step 4:** Checking for duplicate.
- **Step 5:** Remove duplicate.
- **Step 6:** Generate Correlation heatmap. Where correlation heatmap is an essential step for data analysis, exploratory information in a visually appealing way. The value of the coefficient of correlation can take any value from-1 to 1 [26]. When the value is 1, it's a direct correlation between the two variables. That means when one variable increases, the opposite variable also increases. While if the value is -1, it's an indirect correlation between two variables, in which when one variable increases, the opposite variable decreases. Therefore, when 0 value, there's no correlation between two variables as the variables

change in a random manner with reference to one another.

E. Store

This stage may be visited many times. In this case, it has been visited twice. One for storing the gathered data and one after finishing the feature engineering stage.

F. Validate

In this phase, more data analysis has been applied to decide whether the data modelling phase is needed or not. Also, validate phase has been done using Python 3 and the Jupyter Notebook IDE. It may be done using anthor tools as Microsoft Excel or Power BI.

- **Step 1:** Display dataset description. A dataset description provides the following information for each feature: five number summery (minimun value, 25%, 50%, 75%, maximum value) in addition to mean, standard deviation, and count. For example, *age* feature five number summery (29, 47, 55, 61, 77 respectively). This means that minimum age in the dataset is 29, first quarter of age is 47, second quarter of age is 55, third quarter of age is 61, and maximum is 77 years. In addition to mean equals 54, with standard deviation 9, and count equals 303.
- **Step 2:** Generate correlation between target column and each feature in descending order. After tagging the absolute correlation values. This step gives the most effective features that correlate with target column. The most four effective features were exang, cp, oldpeak, and thalach as shown in Fig. 2.
- **Step 3:** Exploratory Data Analysis for each feature. This step gives more information about each feature. In addition, it represents heart disease root cause value for each feature. Table II illustrates each feature root cause value for having heart disease. The features have been arranged in descending order as same as in Fig. 2.

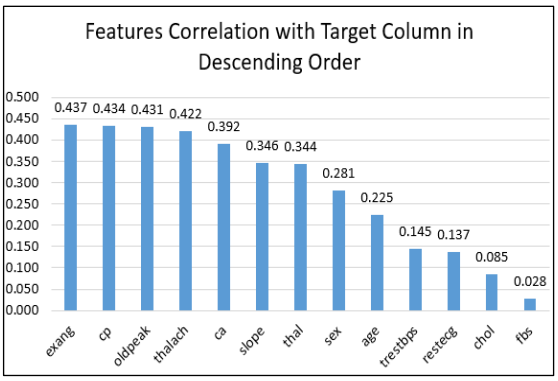

Fig. 2. Features correlation in descending order

Thus, in order to determine the most common causes of heart disease and what each cause truly affects, the validation step is enough for doing that in addition to defining a set of values for heart prediction. While, if the task is the prediction of heart disease or not, data modelling needs to be done.

G. Data Modeling

For the heart disease prediction. This step has done to prove and support the proposed cycle. Two models have been developed. The same python code has been applied for both. One with all the features and the other with only the highest four features (exang, cp, oldpeak, and thalach) that highly affect the target. The two models were simply implemented. Using the following sklearn classifiers: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). In addition to that, XGBoost and Neural Network models with only one hidden layer and just 300 epochs.

H. Testing

Two test cases have been done. Where accuracy, f1 score, sensitivity, and specificity have been used in evaluation. According to [35,36], accuracy is the number of correctly classified data samples over the total number of data samples.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}. \quad (1)$$

An F1 score is a measure of a test's accuracy.

$$F1\ Score = \frac{2\,TP}{2\,TP\,+FP\,+FN}. \quad (2)$$

Sensitivity is the accuracy of a test to properly identify patients with a disease. In other words, it is the number of true positives divided by the number of actual positives.

$$Sensitivity = \frac{TP}{TP+FN}. \quad (3)$$

Specificity is the accuracy of a test to properly identify people without the disease. This means that is the number of true negatives divided by all actual negatives. The test is positive if the person has the disease and, therefore, the test is positive. While true negative means the person doesn't have the disease and therefore the test is negative. A false positive is when the person doesn't have the disease and therefore the test is positive. A false negative means the person has the disease and therefore the test is negative.

$$Specificity = \frac{TN}{TN+FP}. \quad (4)$$

In Case 1, the results for both developed models, the models with the highest four features and the models with all features, have been recorded in Tables III and IV, respectively. In Table III, the highest four features model has achieved great accuracies especially, for the neural network that was developed only with one hidden layer and 300 epochs. The accuracy recorded was 87.32% against 86.89 for the developed model with all features.

TABLE II. EACH FEATURE ROOT CAUSE VALUE

Feature	Value	Comment
exang	0	Samples without exercise induced angina are much likely to have heart disease. The same result was mentioned in [27].
cp	3	Samples with non-anginal pain are much likely to have heart disease. The same result was mentioned in [28].
oldpeak	[0, 1]	ST depression induced by exercise relative to rest with values from 0 to 1 are much likely to have heart disease.
thalach	[140, 170]	Samples with heart rate value from 140 to 170 are much likely to have heart disease.
ca	0	Number of major vessels equals 0 is much likely to have heart disease.
slope	3	The down sloping samples of the peak exercise ST segment are much likely to have heart disease.
thal	6	Fixed defect thalassemia samples are much likely to have heart disease.
sex	1	Males are much likely to have heart disease which was also mentioned in [29].
age	>= 40	Age grater or equal than 40 years old samples are much likely to have heart disease as mentioned in [30].
trestbps	[140, 200]	Resting blood pressure value from 140 to 200 samples are much likely to have heart disease as mentioned in [31,32].
restecg	1	Resting electrocardiographic of value 1 indicates having ST-T wave abnormality is much likely to have heart disease.
chol	[200, 300]	Serum cholesterol samples value from 200 to 300 are much likely to have heart disease as also mentioned in [33].
fbs	0	Samples with fasting blood sugar less than 120 mg/dl are much likely to have heart disease as mentioned in [34].

TABLE III. MODELS WITH HIGHEST 4 FEATURES COMPARATIVE ANALYSIS

Classifier	Accuracy	F1 Score	Sensitivity	Specificity
LR	85.96	86.21	89.29	82.76
NB	82.46	83.87	92.86	72.41
SVM	84.21	85.25	92.86	75.86
K-NN	80	79.52	84.62	76.09
DT	77.19	77.97	82.14	72.4
RF	82.46	83.33	89.29	75.86
XGBoost	78.95	79.31	82.14	75.86
Neural Network	87.32	86.15	90.32	85.0

In addition, the accuracies of LR, SVM and XGBoost were also greater than the same classifiers accuracies in Table IV. This proves that highest four features (exang, cp, oldpeak, and thalach) really affected the target column. Moreover, the sensitivities of the highest four features model were either higher than (as LR, NB, SVM and XGBoost) or comparable with the sensitivities of all features model. Thus means, the highest four features model greatly classify samples with heart diseases.

In Case 2, the recorded results for machine learning classifiers in the model with the highest four features were comparable to the results of both P. Gupta et al.'s model (with 13 features) [20] and Bharti et al.'s model (with 13 features) [22]. Fig. 3 to 5 illustrates the accuracy, sensitivity, and specificity comparisons respectively. The highest four features model classifiers (LR, SVM, RF and XGBoost) achieved higher accuracies than the same classifiers in Bharti et al.'s model. In consideration, their results were recorded after applying three approaches. Also, KNN, RF and XGBoost accuracies in highest four features were higher than P. Gupta et al.'s model. Moreover, the average accuracy of the highest four features model recorded 81.6% against 82.15% in P. Gupta et al.'s model and 80.88% in Bharti et al.'s model.

TABLE IV. MODELS WITH ALL (13) FEATURES COMPARATIVE ANALYSIS

Classifier	Accuracy	F1 Score	Sensitivity	Specificity
LR	85.25	86.96	88.23	81.48
NB	83.61	85.71	88.23	77.78
SVM	83.61	85.71	88.23	77.78
K-NN	84.21	86.67	90.69	75.76
DT	82.89	85.06	86.04	78.79
RF	90.16	91.18	91.18	88.89
XGBoost	77.63	80.0	79.07	75.76
Neural Network	86.89	88.0	85.29	88.88

The sensitivities, shown in Fig. 4, of all classifiers in the highest four features model were higher than both the sensitivities in both P. Gupta et al.'s model and in Bharti et al.'s model (except for KNN was 84.64% against 85%). This means that the highest four features model has classified the samples with heart diseases greatly better than both comparable models. While the specificities of P. Gupta et al.'s model classifiers were higher than the highest four features model which mean it has classified samples without heart diseases better.

I. Document

All the previous steps have to be documented step by step.

J. Deliver

After writing the documentation, it is time to deliver all that has been done to the operations team for being deployed in production.

K. Consume

In this study, the introduced models were decided to be consumed.

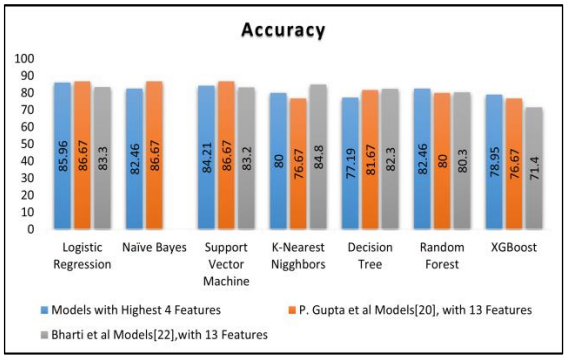

Fig. 3. Accuracy comparative analysis

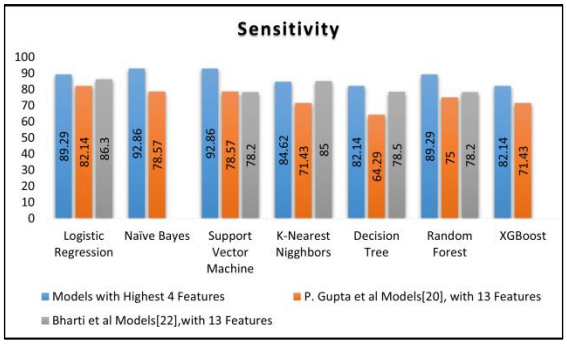

Fig. 4. Sensitivity comparative analysis

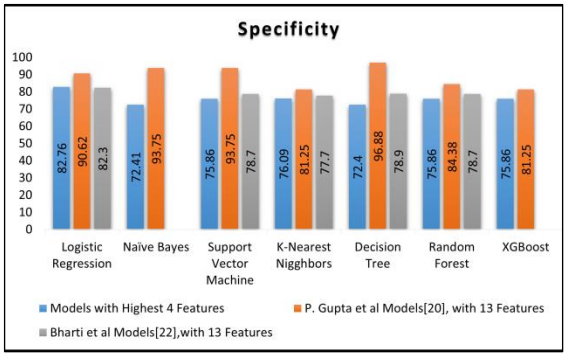

Fig. 5. Specificity comparative analysis

L. Monitor

In real production, this phase must be done to observe models' behaviors.

M. Archive/Delete

Choosing to do this phase will be an unicorn decision. In this study all the work has been archived.

VII. Conclusion and Future Work

This study proposed a novel DataOps lifecycle along with a detail for each phase that was applied to a healthcare case study. For this case, the UCI Machine Learning Heart Disease dataset has been used which has 13 features in addition to target column. The dataset has been analyzed without modeling to find the best most effective features. This analysis came up with highest four effective features (exang, cp, oldpeak, and thalach) that affected the target column, indicating that data modeling is not necessary for all data science project which led to dataset reduction. Then, two models, with the same python code, have been developed for this dataset. The first included 13 features. The second one was with only the highest four (exang, cp, oldpeak, and thalach) features after removing duplicates in rows (20 rows). Two comparisons using accuracy, f1score, sensitivity, and specificity have been done.

Case 1 is the results of the developed model with 13 features and the model with only four features. The comparison showed that the highest four feature model produced higher accuracy and sensitivity, especially for neural networks, with 87.32% and 90.32%. Considering that the neural network model has been developed with only one hidden layer and just 300 epochs; also, the average accuracy and sensitivity of the highest four feature model was 82.32% and 87.94%, respectively.

Case 2, The results of machine learning classifiers in the four features model were comparable to both P. Gupta et al.'s model [20] and Bharti et al.'s [22] results. The comparison showed that DataOps gives great impact results when applied to machine learning model(s). The accuracy of classifiers in the highest four features model, with an average of 81.6%, was greater than the accuracy of Bharti et al., with an average of 80.88% and comparable with P. Gupta et al. with an average of 82.15%. The sensitivity of all classifiers in the highest 4 feature model, with an average of 87.6%, was greater than the sensitivity of both P. Gupta et al. with an average of 74.49% and Bharti et al., with an average of 81.24% while the specificities were comparable. In addition to that, Bharti et al.'s research mentioned that CP and thalach features [Fig. 5(c) & (e)] [22] were not important for heart disease unlike the proof from the proposed DataOps lifecycle.

For future work, the proposed DataOps lifecycle may apply to other fields such as the economy, education, or industry aside from applying it to deep learning model(s).

Notes

For any additional information or more explanation about either the proposed cycle or the code, feel free to contact co-author.

References

[1] R. Bean, "Why Is It So Hard to Become a Data-Driven Company?" Harvard Business Review, Feb 2021. [Online]. Available: https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company. [Accessed November 2021].

[2] B. Chadha and S. Juwe, Agile Machine Learning with DataRobot, Packet Publishing, 2021, pp. 4-7.

[3] W. Eckerson and J. Earth, "DataOps: Industrializing Data and Analytics," Eckerson Group, 2019.

[4] H. Atwal, Practical DataOps: Delivering Agile Data Science at Scale, Isleworth, UK: Springer, 2020, p. 20.

[5] R. C. M. e. a. Kent Beck, 2001. [Online]. Available: https://agilemanifesto.org/principles.html. [Accessed November 2021].

[6] G. B. E. S. Christopher Bergh, "DataOps Principles," 2017. [Online]. Available: https://dataopsmanifesto.org/en/. [Accessed December 2021].

[7] DataKitchen, "Minimizing Analytics Cycle Time with DataOps," 2017. [Online]. Available: https://medium.com/data-ops/minimizing-analytics-cycle-time-with-dataops-b1a1b6e5ab22. [Accessed December 2021].

[8] "Enabling Design Thinking in," in The DataOps Cookbook, DataKitchen, 2019, pp. 116-117.

[9] K. Graziano, G. Adams, W. Eckerson and M. Ferguson, "what is DataOps?," 2020. [Online]. Available: https://www.truedataops.org/. [Accessed December 2021].

[10] D. Wells, "DataOps: More Than DevOps for Data Pipelines," 2019. [Online]. Available: https://www.eckerson.com/articles/dataops-more-than-devops-for-data-pipelines. [Accessed December 2021].

[11] "What Is DataOps?," 2019. [Online]. Available: https://datakitchen.io/what-is-dataops/. [Accessed January 2022].

[12] A. M. Francés, "Problems with your data? You need DataOps," 2021. [Online]. Available: https://anjanadata.com/en/problems-with-your-data-you-need-dataops/. [Accessed January 2022].

[13] B. Pfefferle, "The DataOps Vendor Landscape, 2021," 2021. [Online]. Available: https://dataops.datakitchen.io/pf-eckerson-everything-you-need-to-know-about-dataops-solutions/pf-blog-dataops-vendor-landscape?utm_source=datakitchen&utm_medium=referral&utm_campaign=webinar_eckerson_dataops_solutions. [Accessed January 2022].

[14] J. Ereth, "DataOps – Towards a Definition," LWDA, pp. 104-112, 2018.

[15] P. R. Sahoo and A. Premchand, "DataOps in Manufacturing and Utilities Industries," International Journal of Applied Information Systems (IJAIS), 2019.

[16] A. Capizzi, S. Distefano and M. Mazzara, "From DevOps to DevDataOps: Data Management in DevOps processes," arXiv[cs.SE], 2019.

[17] M. Rodriguez, L. Jonat˘ and M. Mazzara, "Good practices for the adoption of DataOps in the software industry," Information Technologies, Telecommunications and Control Systems (ITTCS), 2020.

[18] A. Raj, D. I. Mattos, J. Bosch, H. H. Olsson and A. Dakkak, "From Ad-Hoc Data Analytics to DataOps," in International Conference on Software and Systems Process, South Korea, 2020.

[19] F. Rahman and M. A. Mahmood, "A Comprehensive Analysis of Most Relevant Features Causes Heart Diseasec Using Machine Learning Algorithms," in Proceedings of the International Conference on Big Data,IoT, and Machine Learning, 2022.

[20] P. Gupta, S. Mala, A. Shankar and V. S. Asirvadam, "Heart Disease Detection Scheme Using a New Ensemble Classifier," in Advances in Data and Information Sciences, 2022.

[21] Z. Alom, M. A. Azim, Z. Aung, M. Khushi, J. Car and M. A. Moni, "Early Stage Detection of Heart Failure Using Machine Learning Techniques," in Proceedings of the International Conference on Big Data, IoT, and Machine Learning, 2022.

[22] R. Bharti, A. Khamparia, M. Shabaz and G. Dhiman, "Prediction of Heart Disease Using a Combination of Machine," Hindawi, 2021.

[23] "What Data Scientists Really Need," in The DataOps Cookbook, 2019, pp. 126-127.

[24] "cardiovascular diseases," 2021. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. [Accessed February 2022].

[25] "Heart Disease Data Set," 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease. [Accessed December 2021].

[26] B. Williams and S. Cremaschi, "Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction," in Computer Aided Chemical Engineering, Elsevier, 2020, pp. 1639-1644.

[27] L. Anderson, A. M. Dewhirst, J. He, M. Gandhi, R. S. Taylor and L. Long, "Exercise-based cardiac rehabilitation for patients with stable angina," Cochrane Database Syst Rev, 2017.

[28] J. Constant, "The Diagnosis of Nonanginal Chest Pain," The Keio Journal of Medicine, vol. 39, no. 3, 1990.

[29] S. H. Bots, S. A. E. Peters and M. Woodward, "Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010," BMJ Global Health, 2017.

[30] J. L. Rodgers, J. Jones, S. I. Bolleddu, S. Vanthenapalli, L. E. Rodgers, K. Shah, K. Karia and S. K. Panguluri, "Cardiovascular Risks Associated with Gender and Aging," J Cardiovasc Dev Dis, vol. 6, 2019.

[31] F. D. Fuchs and P. K. Whelton, "High Blood Pressure and Cardiovascular Disease," Hypertension, 2019.

[32] E. Rapsomaniki, A. Timmis, J. George, M. Pujades-Rodriguez, A. D. Shah, S. Denaxas, I. R. White, M. J. Caulfield, J. E. Deanfield, L. Smeeth, B. Williams, A. Hingorani and H. Hemingway, "Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people," Elsevier Sponsored Documents, 2014.

[33] J. Roland, "What Is Serum Cholesterol and Why Is It Important?," healthline, 26 January 2017. [Online]. Available: https://www.healthline.com/health/serum-cholesterol#prevention. [Accessed 1 September 2022].

[34] C. Park, E. Guallar, J. A. Linton, D.-C. Lee, Y. Jang, D. K. Son, E.-J. Han, S. J. Baek, Y. D. Yun, S. H. Jee and J. M. Samet, "Fasting Glucose Level and the Risk of Incident Atherosclerotic Cardiovascular Diseases," Diabetes Care , 2013.

[35] S. Amelia, H. Roberta and T. Alison, "What are sensitivity and specificity?" BMJ, 2019.

[36] H. N. B, "Confusion Matrix, Accuracy, Precision, Recall, F1 Score," Dec 2019. [Online]. Available: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd. [Accessed Dec 2021].