

Chapter 47

Performance Analysis of Machine Learning Algorithms by Using WEKA and Scikit-Learn



K. Sai Madhuri, B. Prathusha, G. Manish Reddy, Gayatri Deshmukh, K. Bhuvan Sai Kumar, and Aklilu Markos

Abstract This study compares the performance of WEKA and Scikit-learn in predicting whether a person has diabetes or not based on different machine learning algorithms. The dataset contains information on the patient's age, insulin level, blood pressure, and other relevant features. We compare the performance of algorithms such as KNN, decision trees, random forests, and logistic regression dataset, and evaluate their accuracy using various performance metrics. The results of this study can provide insights into which algorithm is most suitable for predictive analysis of the diabetes dataset and can help the user make informed decisions to improve their performance. The findings of this study can aid in selecting the appropriate tool for similar data analysis tasks. Through this research, we hope to highlight the effectiveness of WEKA and Scikit-learn in data analysis and how these tools can be leveraged to uncover insightful information that may guide individual choices.

Keywords WEKA · Machine learning · KNN · Decision tree regression · Scikit-learn

K. Sai Madhuri · B. Prathusha · G. Manish Reddy · G. Deshmukh (✉) · K. Bhuvan Sai Kumar
CSE Department, Vallurupalli Nageswarao Rao Vignana Jyothi Institute of Engineering and
Technology, Hyderabad, India
e-mail: 9704582222gd@gmail.com

K. Sai Madhuri
e-mail: saimadhuri_k@vnrvjiet.in

B. Prathusha
e-mail: prathusha_b@vnrvjiet.in

A. Markos
Hawassa University, Hawassa, Ethiopia
e-mail: Akimark@hu.edu.et

1 Introduction

Millions of individuals throughout the world suffer with diabetes, a chronic illness. The early detection and prevention of diabetes can be aided by predicting the risk of the disease. Based on different patient characteristics including age, insulin level, and blood pressure, machine learning algorithms have shown to be successful in predicting diabetes [1]. WEKA and Scikit-learn are well-known tools for data analysis and machine learning in this context, giving a variety of algorithms to analyze datasets.

The University of Waikato in New Zealand developed the open-source data mining program Waikato Environment for Knowledge Analysis (WEKA). The analysis of datasets is made easier by the graphical user interface (GUI) that is provided. A wide variety of techniques for preprocessing data, classification, regression, clustering, and visualization are supported by WEKA. Additionally, a complete set of evaluation techniques are offered to assess the effectiveness of the models produced by various algorithms. Although WEKA is primarily used for academic study and instruction, it can also be used for business-related purposes [2].

Scikit-learn is a popular cloud-based platform that provides an easy-to-use environment for data analysis and machine learning. It allows users to write and execute code in a Jupyter Notebook format, with access to powerful computing resources such as GPUs and TPUs [3]. Jupyter Notebooks, the programming language Python, and a number of libraries for scientific computing, data analysis, and visualization, including NumPy, Pandas, Scikit-learn, and Matplotlib are all included. Windows, macOS, and Linux are just a few of the operating systems that Scikit-learn supports. Additionally, it offers Scikit-Learn Navigator, an integrated development environment (IDE) with a user-friendly interface for managing the libraries and environments used for data analysis. Both tools offer a range of algorithms to analyze datasets, including classification, regression, and clustering [4].

This study focuses on comparing the performance of WEKA and Scikit-learn in predicting the risk of diabetes using a diabetes dataset. The dataset contains information on various patient attributes, such as age, gender, BMI, and blood pressure, along with a class label indicating whether the patient has tested positively or not for diabetes. The analysis aims to evaluate the accuracy of various machine learning algorithms, such as KNN, logistic regression, decision trees, and random forests in predicting the risk of diabetes.

WEKA and Scikit-learn offer similar capabilities in terms of algorithm selection and performance evaluation. The findings of this study can help in selecting the appropriate tool for predicting diabetes using machine learning algorithms and aid in healthcare and medical research [5].

Data-Flow Diagram for Data Mining Process

See Fig. 1.

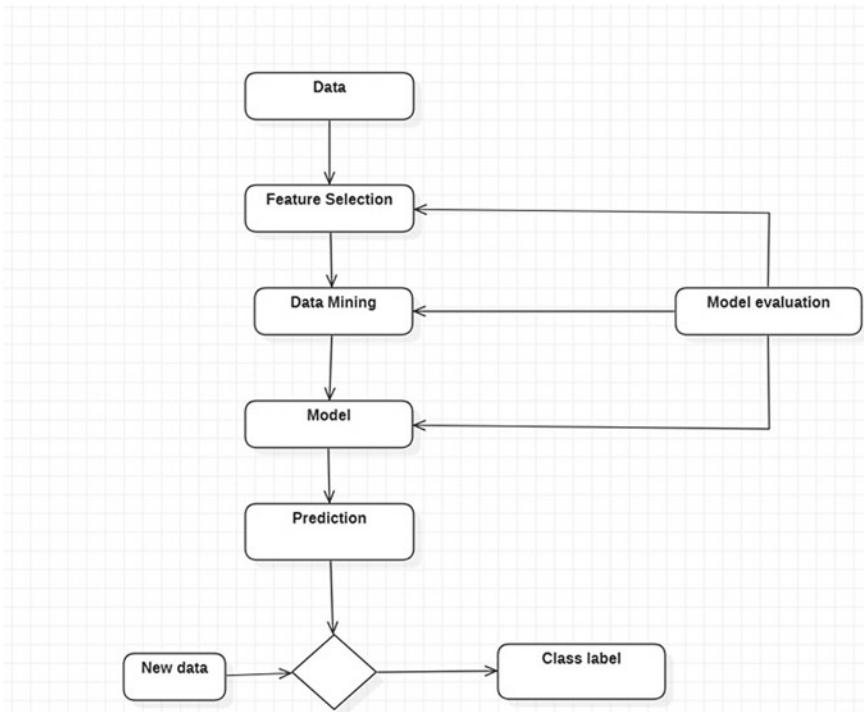


Fig. 1 Outline of the data mining process. It involves the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision-making, and other data requirements to eventually cost-cutting and generating revenue.

2 Literature Review

S. No.	Title	Methods/Approach	Pros/Cons	Year
1 1	WEKA powerful tool in data mining	<ul style="list-style-type: none"> Preparing the data Classify, apply algorithm Generate trees Analysis 	<ul style="list-style-type: none"> Accessible free under the general public license (GNU) Portability Easy to use 	2016
2	Performance analysis of classification techniques in data mining using WEKA	<ul style="list-style-type: none"> Data preprocessing Classification Clustering Data visualization 	<ul style="list-style-type: none"> Accurate Less time taken 	2021

(continued)

(continued)

S. No.	Title	Methods/Approach	Pros/Cons	Year
3	WEKA tool with different data mining techniques	<ul style="list-style-type: none"> • Association • Prediction 	<ul style="list-style-type: none"> • Less documentation • Clear graphical outputs 	2020
4	A survey on machine learning: Concept, algorithms and applications	Research and statistics	<ul style="list-style-type: none"> • Quality predictions • Higher response rates <p>Cons:</p> <ul style="list-style-type: none"> • No visual aids • Difficult to develop rapport 	2017
5	A survey on data collection for machine learning	<ul style="list-style-type: none"> • Data searching • Data augmentation • Crowd sourcing 	<ul style="list-style-type: none"> • Higher response rates • Allows clarification • Larger target 	2019
6	Review of data mining with WEKA tool	• KDD	<ul style="list-style-type: none"> • High performance • Minimum time 	2016

2.1 Methodologies

This section discusses the methods and materials used in the experimentation and analysis of the proposed dataset. On the common diabetes dataset, the analysis is conducted.

- Data analysis: Here, one will learn how the data analysis phase of a data science life cycle is carried out.
- Exploratory data analysis: EDA is one of the most crucial phases of a data science project, and at this stage, one must understand how to draw conclusions from visualizations and data analysis.
- Model building: Here we will be using four standard machine learning models and then we will choose the best-performing tool for those respective models.

2.2 Dataset

In machine learning and data science, the diabetes dataset is well-known and frequently used for classification and predictive modeling applications. This dataset includes information on people who have undergone a diabetes test, including whether or not they have received a diagnosis. The dataset has a number of characteristics, such as demographic data, medical history, and test results from a lab.

The UCI Machine Learning Repository contains the original diabetes dataset, which was created by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset has eight characteristics and 768 observations, including:

- Pregnancies: The number of times the individual has been pregnant
- Glucose: Plasma glucose concentration a 2 h in an oral glucose tolerance test
- Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: Triceps skin fold thickness (mm)
- Insulin: 2-h serum insulin (mu U/ml)
- BMI: Body mass index [weight in kg/(height in m)²]
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age in years.

The target variable is a variable indicating whether or not the individual has been tested positive with diabetes.

This dataset has been used extensively for machine learning applications including prediction and classification. Predicting a person's diabetes based on their lab examination results and medical history is a typical undertaking, as is figuring out which characteristics are most indicative of diabetes. The dataset is frequently used to evaluate and contrast the effectiveness of several machine learning techniques, including random forest, decision trees, and logistic regression.

2.3 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a crucial step in any data analysis project, including working with the diabetes dataset. EDA involves examining and summarizing the data to gain insights and identify patterns, relationships, and potential issues. Here are some steps you could take when conducting EDA on the diabetes dataset (Fig. 2):

- a. Load and explore the dataset: Start by inserting the dataset into your favorite R or Python data analysis program. To gain a notion of what the data includes, take a short glance at it. To view the top few rows or some fundamental summary statistics of the data, you might utilize methods such as head () or summary ().
- b. Check for missing values: Before beginning the analysis, it is crucial to locate and deal with any missing values in the dataset. You may check for missing values and determine how many observations have missing data using methods like is.na () and summarize ().
- c. Examine distributions and relationships: Consider utilizing histograms, boxplots, or density plots to examine the distributions of the variables in the dataset. To investigate connections between variables, you might alternatively utilize scatterplots or correlation matrices [6].

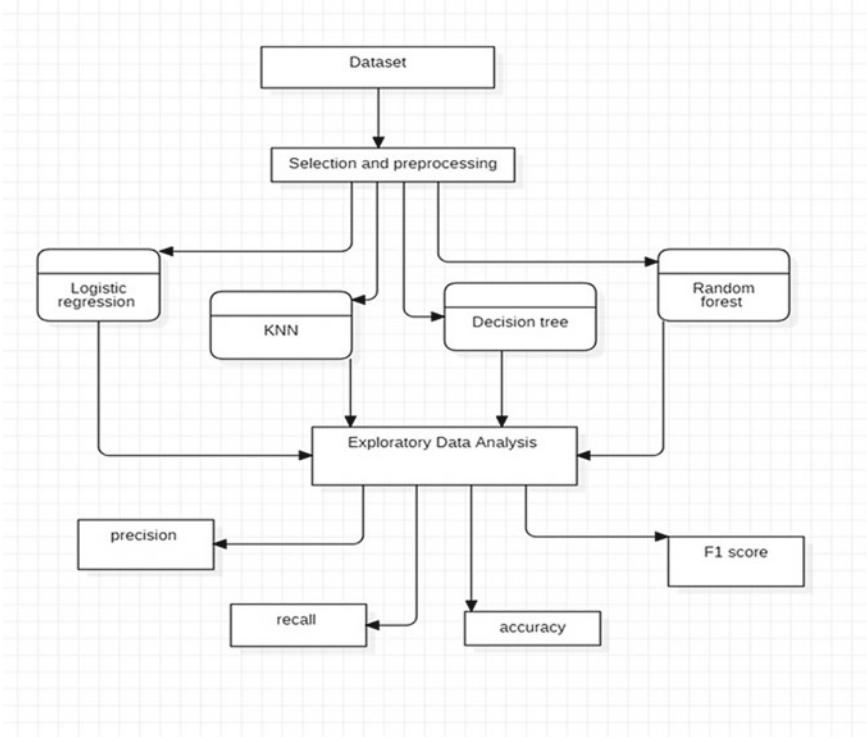


Fig. 2 Well-known algorithms supported by WEKA tool and how to save and load good algorithm configurations. WEKA supports several standard data mining tasks, specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Exploratory data analysis (EDA) is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods

- d. Identify outliers and extreme values: Verify the dataset for any unusual values or outliers that could have an impact on your study. To find these values, you might use scatterplots or boxplots.
- e. Explore variable interactions: Consider the potential interactions between the various dataset variables. To investigate these interactions, you may design scatterplots or other visualizations.
- f. Consider transformations: Depending on the way the variables in the diabetes dataset are distributed, I might consider converting them to enhance the effectiveness of my analysis. For example, I could scale the variables to have a mean of 0 and a standard deviation of 1, or I might use log transformations [13].

2.4 Models

The models that have been built for performing this study are based on four different ML algorithms. Predictive models for diabetes diagnosis have been developed and extensively tested using the diabetes dataset. Numerous machine learning techniques, such as logistic regression, decision trees, KNN, and random forest may be used on this dataset [7].

To prevent overfitting while developing models for the diabetes dataset, it is crucial to divide the data into training and testing sets effectively. When a dataset is divided into k subsets, one subset is used for testing, while the other subsets are utilized for training. This method is known as k-fold cross-validation.

A common approach for binary classification problems, such as determining whether or not a person has diabetes is logistic regression. It is simple to understand in terms of odds ratios and models the likelihood of the target variable as a function of the predictor factors [8].

Another well-liked approach for classification jobs is decision trees, which are frequently employed due to their usability. They operate by repeatedly dividing the data into subsets according to the predictor variables, and they may be seen as a tree structure.

A nonparametric approach called k-nearest neighbors (KNN) may be applied to classification and regression problems. A new instance is categorized using KNN based on the training sets, k closest neighbors dominant class. Prior to training the model, the hyper parameter k must be selected. Although KNN is quite straightforward and simple to comprehend, it can be computationally costly for big datasets and high values of k.

As part of an ensemble learning process called random forest, several decision trees are combined to increase prediction accuracy. Each decision tree in random forest is trained using a random subset of the training data, and each node of the tree takes into account a random subset of the predictor variables. After that, the average of all the forecasts made by the forest trees yields the final prediction. The strong algorithm random forest can handle high-dimensional data and nonlinear correlations, but if there are too many trees, it may be prone to overfitting [9].

Ultimately, my choice of algorithm for building a predictive model on the diabetes dataset will depend on the specific goals of my analysis and the characteristics of the dataset. It is important for me to evaluate the performance of different models using appropriate metrics, such as accuracy, precision, recall, and F1 score, and to interpret the results in the context of the problem I am trying to address. By carefully selecting and tuning the appropriate algorithm, I can achieve accurate and reliable predictions that can help inform decisions related to diabetes diagnosis and management [10].

3 Implementation and Results

To implement algorithms in WEKA on the diabetes dataset, firstly it is required to load the dataset into the software. Then, select the desired algorithms from the list of available classifiers and apply them to the data. We also adjust the hyperparameters of each algorithm and use cross-validation to evaluate the performance of the models. For simplification, to apply the KNN algorithm to the diabetes dataset in WEKA, we selected the “IBk” classifier from the list of classifiers and specified the value of the k parameter [11]. Then, we could run the algorithm on the data and evaluate its performance using metrics such as accuracy and $F1$ score. The results of implementing these algorithms on the diabetes dataset in WEKA can provide valuable insights into the patterns and relationships in the data, as well as the accuracy of the predictive models. By comparing the performance of different algorithms and adjusting their hyperparameters as necessary, we can identify the best algorithm for the task at hand and achieve accurate and reliable predictions [12].

3.1 Experiment Results

This section discusses the results of the analysis being done with the help of a few formatting tables. Experimental results for the diabetes dataset can be reported in terms of precision, recall, and $F1$ score. These metrics are commonly used to evaluate the performance of binary classification models on imbalanced datasets, where one class is much more prevalent than the other. Precision is defined as the number of true positive predictions divided by the total number of positive predictions. It measures the proportion of predicted positives that are actually positive.

Recall, on the other hand, is defined as the number of true positive predictions divided by the total number of actual positives. It measures the proportion of actual positives that are correctly predicted as positive.

The $F1$ score is the harmonic mean of precision and recall and provides a single metric that balances the trade-off between precision and recall (Tables 1 and 2).

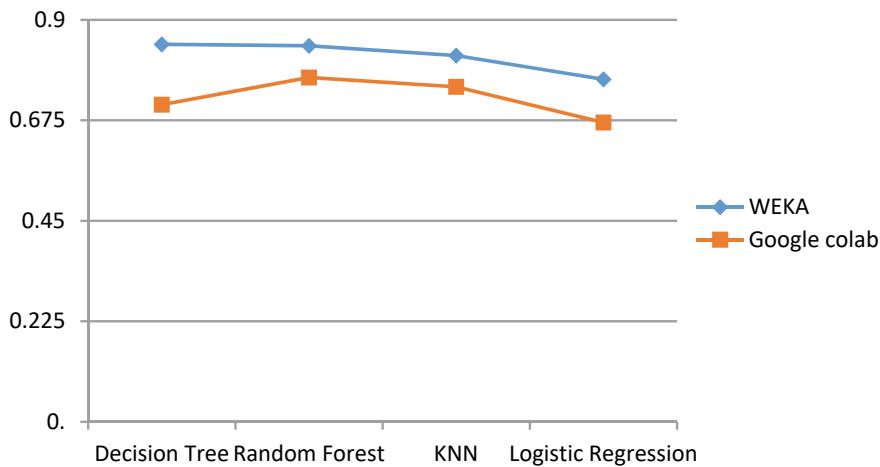
After implementing the KNN algorithm on the diabetes dataset, we find that the precision of the model is 0.82, the recall is 0.81, and the $F1$ score is 0.86. This

Table 1 Precision, recall, and $F1$ score of the proposed models on the two selected tools

Algorithm	WEKA			Scikit-learn		
	Precision	Recall	$F1$ score	Precision	Recall	$F1$ score
Decision tree	0.845	0.843	0.84	0.71	0.74	0.71
Random forest	0.842	0.84	0.836	0.771	0.768	0.762
KNN	0.82	0.81	0.86	0.75	0.76	0.75
Logistic regression	0.767	0.772	0.765	0.67	0.69	0.66

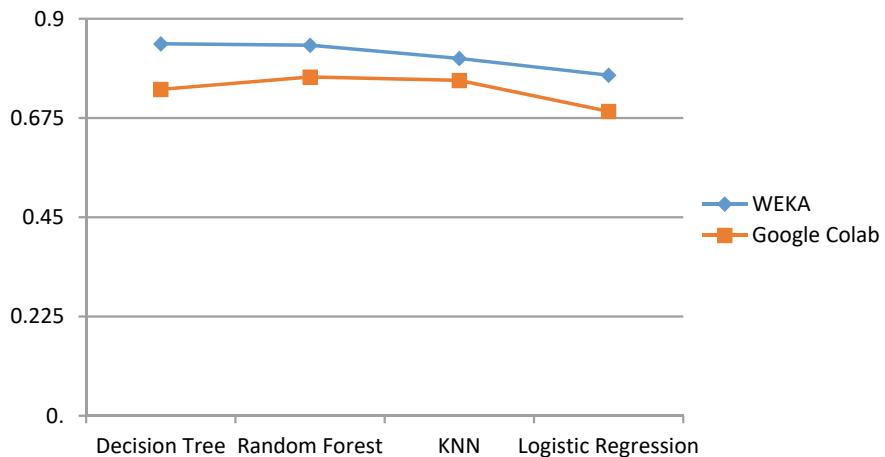
Table 2 Accuracy for the two selected tools on the two proposed models

Algorithm	Accuracy (%)	
	WEKA	Scikit-learn
Decision tree	73.82	71.25
Random forest	78.12	77.50
KNN	79.22	76.31
Logistic regression	77.20	68.50

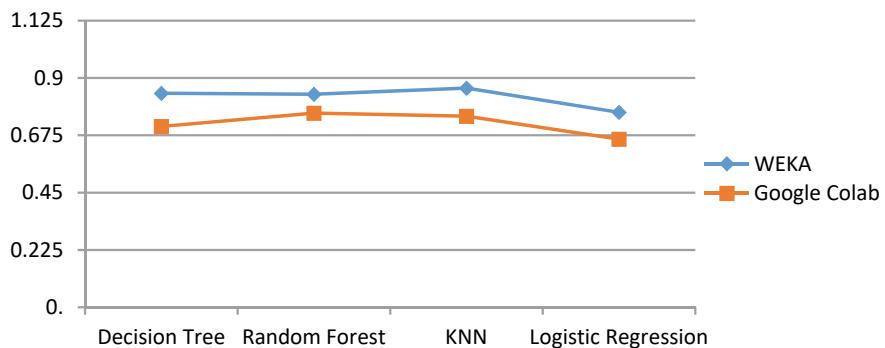


Graph 1 Precision for the two selected tools on the two proposed models

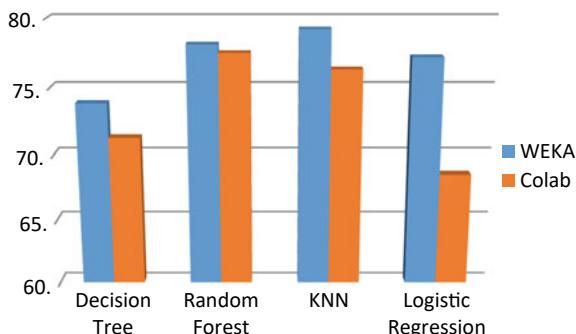
indicates that the model correctly identifies 82% of positive cases, and of the cases it identifies as positive, 81% are actually positive. The F_1 score of 0.86 indicates that the model achieves a balance between precision and recall (Graphs 1, 2, 3 and 4).



Graph 2 Recall for the two selected tools on the two proposed models



Graph 3 *F*1 score for the two selected tools on the two proposed models



Graph 4 Overall accuracy values comparing both the two selected tools

4 Conclusion

This paper originally represents the concepts of data mining along with data preprocessing in WEKA. WEKA is a dominant tool in data mining and analytics, which is used to test and train different datasets with different classifications and clustering algorithms such as k-means, random forest, decision tree, etc. [14]. The leading objective of machine learning researchers is to design effectively in terms of both time and space to improvise the performance over an extensive domain. In the perspective of machine learning environment, the proficiency with which a method utilizes data resources also plays a vital role to improve the accuracy along with the time and space complexity. This paper mainly focuses on comparing the WEKA tool and Sk-learn module. In Weka algorithms are uses majority vote which predicts the class of the instance as the class predicted by majority. The class probability of the instance is computed as fraction of that predicts. The Sk-learn module basically predicts the class of instance as follows. The predicted class of an input instance is the class with the highest mean probability. The predicted class probabilities of an input instance are computed as the mean predicted class probabilities. So, this work finally represents different accuracy levels by comparing WEKA tool and Sk-learn module.

References

1. <https://www.cs.waikato.ac.nz/~ml/weka/index.html>
2. <https://developer.ibm.com/technologies/analytics/>
3. MF bin Othman, TMS Yau (2007) Comparison of different classification techniques using WEKA for Breast Cancer. In: 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, vol 15(04), pp 520–523
4. Gera M, Goel S (2015) Data mining—techniques, methods and algorithms: a review on tools and their validity. *Intl J Comput Appl* 113(19):22–29
5. Karla JB, Nikola B (2014) An overview of free software tools for general data mining. In: IEEE 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp 1112–1117
6. Amin N (2015) Comparison of different classification techniques using WEKA for hematological data. *Am J Eng Res (AJER)* 4(3):55–61
7. Siddiqui, Ausaf A (2018), Data mining tools and techniques for mining software repositories: a systematic review. In: BIG data analytics. Springer, Singapore, pp 717–726
8. Sharma A, Kaur B (2017) A research review on comparative analysis of data mining tools, techniques and parameters. *Intl J Adv Res Comput Sci* 8(7):523–529
9. Kaur K, Dhiman S (2016) Review of data mining with Weka tool. *JCSE Intl J Comput Sci Eng* 4(8):41–44
10. Soucy P, Mineau GW (2001) A simple KNN algorithm for text categorization. In: Proceedings, 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, pp 647–648. <https://doi.org/10.1109/ICDM.2001.989592>
11. Mitchell TM (1997) Machine learning. McGraw-Hill International, Sydney

12. Cour T, Sapp B, Taskar B (2012) Learning from partial labels. *J Mach Learn Res* 12:1501–1536
13. Bach SH, He BD, Ratner A, Re C (2017) Learning the structure of generative models without labeled data. *ICML* 2017:273–282
14. Polyzotis N, Roy S, Whang SE, Zinkevich M (2018) Data lifecycle challenges in production machine learning: a survey. *Sigmod Rec* 47(2):17–28