Final Project CE 888
Learning from Imbalanced Datasets (Supervised and Unsupervised Learning

VIRAJ KUMAR DEWANGAN
REGISTRATION NUMBER - 1901181
APRIL 23, 2020

***Abstract*—Supervised machine learning is that task in which target labels are known prior and the motive is to train the data using those labels and build a predictive model and then predict the class labels of the test data which does not have targets labelled yet.Whereas, Unsupervised Machine Learning involves such datasets in which there are no labels so first they need to be grouped into various similar clusters based upon smallest distance calculation from the centroid algorithms and identifying closely spaced points having similar characteristics. In this project, a new approach has been tested to deal with imbalanced datasets, based on a mixture of supervised and unsupervised learning. Classifiers like Decision Tree and Random Forest have been deployed for supervised learning whereas Elbow method, Silhouette Method, K-Means Clustering have been taken into use for Unsupervised Learning phase of the assignment after performing Stratified Cross Validation and bin sampling.Permutation testing has been carried out for every resulting combination to present the results and then they have been compared with the baseline supervised learning methods and the finding of the experiments have been summarised.**
The objective of this assignment is not merely making the datasets balanced but to come up with and devise the best methods and practices to deal with imbalanced data sets so that the information presented by such data is preserved and harnessed for better decision making and accurate and efficient analysis which would not be possible if we just carried out balancing operation. [1]

Keywords - Imbalanced Datasets, Machine Learning, Supervised Machine Learning, Unsupervised Machine Learning, Decision Tree, Random Forest,Stratification, Cross-Validation, Elbow Method, Silhouette Method, K-Means Clustering, Permutation Testing

I. Introduction

Imbalanced datasets are those datasets in which the representation of different classes in target columns is not equal. Imbalanced datasets are naturally present in majority of the encountered tasks or situation as there is very less probability of all classes or targets to be represented equally or homogeneously due to various complexities inherent in that system. [2] Such imbalanced datasets are encountered both in supervised and unsupervised machine learning scenarios. However, since the proportion of unsupervised learning task is itself greater than supervised ones owing to non labelled targets or data to analyse, imbalanced datasets are more likely to be dealt while carrying out unsupervised machine learning.The main **motivation** of carrying out this project or assignment is to compare the performance of Supervised and Unsupervised Machine Learning when we deal with imbalanced datasets so to establish a concrete approach to deal with imbalanced datasets.

II. Background

There are various ways to make an imbalanced dataset a balanced one - DOWNSAMPLING [3] and UPSAMPLING. Downsampling is done by removing the imbalancing class by multiplying it with the percentage of imbalance present in the data and the reverse techinque is applied to do upsampling [4] where the count of less represented class is increased to achieve the balance. These methods of Downsampling and Upsampling are more relevant and used in the context of Digital Signal Processing (DSP) where the intention is to either increase the sample signal rate in up sampling and decrease the rate in down sampling. [5] However this results in loss of actual data and the hence the analysis thereafter might not be correct due to data loss.

According to the book by Alpaydin on **Introduction to Machine Learning**, Machine Learning is a computational technique to do human tasks using computers without involving human intervention.It basically aims to mimmick human brain working in real world applications and scenarios. Machine Learning is a very broad concept as it involves amalgamation of various entities like Data Mining and Artificial Intelligence to make informed decisions from the available data and also keep on updating and improving the decision making skills with time and repeated training which is then stored in the database and is referred to as **Prior Learning** [6]

Data Mining takes it names from the analogy of earth mining and means that amidst availability of so much data both useful and scrap, extract the useful and relevant data chunks and then process and analyse them for information and insights. Artificial Intelligence is making a bot think and learn like a human by replicating the working of neuron cells

and synapses in brain using Neural Networks. [6]

The following diagram shows the basic schema of Machine Learning Model which has been adapted from the lecture slides of *Dr. Liviu Ciortuz, Department of CS, University of Iasi, Romania* on **MACHINE LEARNING**

TYPES OF MACHINE LEARNING

Machine Learning is mainly classified into following 4 types-

1. Supervised Learning
2. Unsupervised Learning
3. Semi-Supervised Learning
4. Reinforcement Learning

1. In **Supervised Learning**, the labelled training data is used to create a function or model which then maps an input of unlabelled data to different classes or labels. [7]

2. In **Unsupervised Learning**, we try to detect patterns and relations between the unlabelled points. It helps in mapping probability density functions to a dataset when proper relations are found. Unsupervised Machine Learning is mainly done through Principal Component Analysis (PCA) and Cluster Analysis. [6] In PCA the main objective is to exploit correlation between different features and points to capture maximum variance and hence maximum information. This is also known as Dimensionality Reduction and is a part of data pre-processing or data wrangling phase of Machine Learning
Cluster Analysis is one of the focus areas of this assignment and deals with creating clusters of similar points within the feature space.

2(a) **Supervised V/S Unsupervised Learning**- The main difference between supervised and unsupervised techniques lies in the data they handle and the type of application they are put into use. While Supervised Learning deals with labelled data to build predictive models, Unsupervised deals with unlabelled data and is mainly used to estimate probability density functions in statistics.

3. **Semi-Supervised Learning** involves combining supervised and unsupervised methods where the available data is very meagre in amount and the unlabelled data is high. It is mainly used in application like audio classifications where the sample labelled data is limited and the data to analyse is very large. It is mainly done to reduce costs as labelling of data is an expensive affair. [8]

4. **Reinforcement Learning** is that concept which is mostly used in games and involves the concepts of rewards and penalty to arrive at the best possible action given a state of the system. Markov Decision Process are inherent part of this and assume that all information is captured in a given state and that the optimal course of action is take take is that which gives maximum reward or minimum penalty. It is an additive learning model which keeps on updating itself to take better decision from the prior learning acquired. [6]

Cross Validation is a method used in predictive validation tasks to make sure that the results obtained are correct. In it, the data set is partitioned into K-Folds and then the same classification task is run again and again with different combinations of training and testing data to get an average result of the experimental setup. An example of Cross Validation is LEAVE ONE OUT where one K data fold is permuted as test data for all combinations and results are then averaged. [9]

Stratification is a kind of data sampling technique in which the proportion and characteristics of the data are maintained intact by taking into consideration disjoint groups within the data.It is widely used while doing cross validation.Stratification is particularly easy when there are binary labels and there is great scope and work going on multi label data stratification at the moment. [6]

Now, the next interest of discussion is the two classification methods used in the project.

Decision Tree Classification is used to construct classification trees where the nodes or leaves represent the class labels and the the branches associates the relevant features linked to that label. It is a highly useful tool for the predictive classification tasks in machine learning. K Nearest Neighbors, Naive Bayes, Support Vector Machines are some of the other similar classification methods. [10]

In **Random Forest Classification**, we construct a multitude of decision trees using subsets of feature space and then take the mode or median or mean of all the decision trees to best classify the dataset. Random forest can be seen as a bagging concept or ensemble concept in classification which improves on Decision tree by using mode/mean/median of various trees. Hence the performance is better. [11]

Moving on to the literature for unsupervised and clustering techniques employed in the project-

Clustering is a technique to group unclassified data points in a dataset into distinct groups or clusters having similar properties and features. This is done by using algorithms which work on smallest distance computation as spatial points having least distance are likely to belong to same cluster or group. The group points are mapped to the cluster centroid or means in this algorithm.
[12]
In **K-Means Clustering** k refers to the number of clusters or groups that the experiment or model wants to generate.

Mean refers to the distance mapping of data points to the mean or centroid of the cluster, hence the nomenclature 'K-MEANS CLUSTERING'.

Elbow Method is a clustering method which uses heuristics to arrive at the optimal value of K to be taken for clustering of the dataset.This involves mapping the explained variations with the number of clusters using different k values and taking that K value from the curve which lies at elbow point, hence the name ELBOW METHOD. [13]

Silhouette Method is a validation tool for clustering used to explain consistency and how well has the data been classified. It explains how cohesive(similar to own cluster) or adhesive(similar to different cluster) a data point is. The value of Silhouette ranges from -1 to +1. Higher the value that is more positive or less negative the value implies high similarity to the cluster and more cohesiveness. Similarly negative values shows less linkage or negative correlation and incorrect grouping. [1]

Permutation Testing is one of the statistical significance test in whhich the null hypothesis is to test the distribution of the test statistics using various permutation of the dataset and to see whether the distribution remains invariant and same given different permutation. It also known as Randomisation Test and is relevant under the context of Resampling. [14]

III. Methodology

Three datasets were chosen for this assignment from the Kaggle. The datasets were such that each one of them had data imbalance in different proportions. The names of the datasets are mentioned as under-
1. DATASET 1 - Bank Marketing data for opening term deposits with imbalance of 65.23 %
2. DATASET 2 - Porto Seguro's Safe Driver Prediction data for determining insurance claims with imbalance of 96.34 %
3. DATASET 3 - Credit Card Fraud Detection data with highest imbalance of 99.69 %

The datasets were then loaded into the Google Colab Python environment, explored,inspected for their characteristics and preprocessed accordingly to make them ready for the tasks of Supervised Machine Learning of Decision Tree Classification and Random Forest Classification.

Lets us look into the characteristics and information about the datasets.

DATASET 1 - BANK MARKETING DATA

The first dataset is taken from the data science community Kaggle and is a dataset related to direct marketing campaigns of a Portuguese Bank. The link to download and access the data is https://www.kaggle.com/henriqueyamahata/bank-marketing. This dataset was a part of Kaggle competition in which the task was to predict whether a client would subscribe to a term loan or not given all these information about the client.There was 65.23 % imbalance in this dataset(lowest of three). The dataset has the following attributes summarised below using python platform-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Fig. 1. BANK MARKETING DATA

There are 15 features here and the target column is y which shows whether clients subscribe loan or not.The dataset had different datatypes of features and had to be transformed into numerical and categorical data types as a part of preprocessing to run classifiers on it as Classifiers only work with categorical and numeric data types. Also I had to deal with missing values in the data. Since the NaN values were only present in 2 or 3 rows, I simply removed them without causing any harm to the original data.Fig 1 shows information about bank marketing data.

DATASET 2 - PORTO SEGURO's SAFE DRIVER PREDICTION DATA

The second dataset is taken from the Kaggle competition of predicting the probability of an auto insurance policy holder to file a claim. https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data is the link to access the data. It has an imbalance of 96.34%. With 56 input features the label column here is named as 'target' which shows whether one

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 595212 entries, 0 to 595211
Data columns (total 59 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   id              595212 non-null  int64
 1   target          595212 non-null  int64
 2   ps_ind_01       595212 non-null  int64
 3   ps_ind_02_cat   595212 non-null  int64
 4   ps_ind_03       595212 non-null  int64
 5   ps_ind_04_cat   595212 non-null  int64
 6   ps_ind_05_cat   595212 non-null  int64
 7   ps_ind_06_bin   595212 non-null  int64
 8   ps_ind_07_bin   595212 non-null  int64
 9   ps_ind_08_bin   595212 non-null  int64
 10  ps_ind_09_bin   595212 non-null  int64
 11  ps_ind_10_bin   595212 non-null  int64
 12  ps_ind_11_bin   595212 non-null  int64
 13  ps_ind_12_bin   595212 non-null  int64
 14  ps_ind_13_bin   595212 non-null  int64
 15  ps_ind_14       595212 non-null  int64
 16  ps_ind_15       595212 non-null  int64
 17  ps_ind_16_bin   595212 non-null  int64
 18  ps_ind_17_bin   595212 non-null  int64
 19  ps_ind_18_bin   595212 non-null  int64
 20  ps_reg_01       595212 non-null  float64
 21  ps_reg_02       595212 non-null  float64
 22  ps_reg_03       595212 non-null  float64
 23  ps_car_01_cat   595212 non-null  int64
 24  ps_car_02_cat   595212 non-null  int64
 25  ps_car_03_cat   595212 non-null  int64
 26  ps_car_04_cat   595212 non-null  int64
 27  ps_car_05_cat   595212 non-null  int64
 28  ps_car_06_cat   595212 non-null  int64
 29  ps_car_07_cat   595212 non-null  int64
 30  ps_car_08_cat   595212 non-null  int64
 31  ps_car_09_cat   595212 non-null  int64
 32  ps_car_10_cat   595212 non-null  int64
 33  ps_car_11_cat   595212 non-null  int64
 34  ps_car_11       595212 non-null  int64
 35  ps_car_12       595212 non-null  float64
 36  ps_car_13       595212 non-null  float64
 37  ps_car_14       595212 non-null  float64
 38  ps_car_15       595212 non-null  float64
 39  ps_calc_01      595212 non-null  float64
 40  ps_calc_02      595212 non-null  float64
 41  ps_calc_03      595212 non-null  float64
 42  ps_calc_04      595212 non-null  int64
 43  ps_calc_05      595212 non-null  int64
 44  ps_calc_06      595212 non-null  int64
 45  ps_calc_07      595212 non-null  int64
 46  ps_calc_08      595212 non-null  int64
 47  ps_calc_09      595212 non-null  int64
 48  ps_calc_10      595212 non-null  int64
 49  ps_calc_11      595212 non-null  int64
 50  ps_calc_12      595212 non-null  int64
 51  ps_calc_13      595212 non-null  int64
 52  ps_calc_14      595212 non-null  int64
 53  ps_calc_15_bin  595212 non-null  int64
 54  ps_calc_16_bin  595212 non-null  int64
 55  ps_calc_17_bin  595212 non-null  int64
 56  ps_calc_18_bin  595212 non-null  int64
 57  ps_calc_19_bin  595212 non-null  int64
 58  ps_calc_20_bin  595212 non-null  int64
dtypes: float64(10), int64(49)
memory usage: 267.9 MB
```

Fig. 2. PORTO SEGURO's SAFE DRIVER PREDICTION DATA

would file an insurance claim or not.The data types involved here were all numeric so no futher preprocessing was required before carrying out the classification task of Decision Tree and Random Forest baselines. Fig 2 shows the information of Porto Seguro's Data.

DATASET 3 - CREDIT CARD FRAUD DETECTION DATA

The last dataset having the largest imbalance of 99.69% is the Credit Card Fraud Detection data taken from Kaggle. it contains transactions made by European Credit Cardholders in September 2013.Here the target or label is the 'Class' column which shows whether the transaction is fraud or not Attributes present here have already undergone the process of Principal Component Analysis(PCA) and the resulting 28 features are

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 206058 entries, 0 to 206057
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    206058 non-null  float64
 1   V1      206058 non-null  float64
 2   V2      206058 non-null  float64
 3   V3      206058 non-null  float64
 4   V4      206058 non-null  float64
 5   V5      206058 non-null  float64
 6   V6      206058 non-null  float64
 7   V7      206058 non-null  float64
 8   V8      206058 non-null  float64
 9   V9      206058 non-null  float64
 10  V10     206058 non-null  float64
 11  V11     206058 non-null  float64
 12  V12     206058 non-null  float64
 13  V13     206058 non-null  float64
 14  V14     206058 non-null  float64
 15  V15     206058 non-null  float64
 16  V16     206058 non-null  float64
 17  V17     206058 non-null  float64
 18  V18     206058 non-null  float64
 19  V19     206058 non-null  float64
 20  V20     206058 non-null  float64
 21  V21     206057 non-null  float64
 22  V22     206057 non-null  float64
 23  V23     206057 non-null  float64
 24  V24     206057 non-null  float64
 25  V25     206057 non-null  float64
 26  V26     206057 non-null  float64
 27  V27     206057 non-null  float64
 28  V28     206057 non-null  float64
 29  Amount  206057 non-null  float64
 30  Class   206057 non-null  float64
dtypes: float64(31)
memory usage: 48.7 MB
```

Fig. 3. CREDIT CARD FRAUD DETECTION DATA

all of numeric data types with only one NaN or missing value in rows.That NaN row was removed using dropna() built in function. Hence our dataset is fit to carry out the Supervised Learning Classification task. Fig 3 shows the summary of the dataset.

In the second step, two different baseline methods were employed for supervised machine learning phase and cross-validation was performed on the datasets using decision tree and random forest classifier. The documentation available in scikit learn was referred to implement these baselines.

The confusion matrix and other performance metrics like Kappa Statistic, Accuracy,Precision, Recall, F1 Score, Support were used as evaluation parameters for the classification stage. In general, higher the value of all these parameters better is the performance of the classifier.

Till so far, we completed the supervised learning required for the project and assignment on the imbalanced datasets.

UNSUPERVISED LEARNING PHASE-

Now we needed to carry out the unsupervised learning on the datasets.

To start with, the datasets were partitioned into 10 bins, keeping intact the original imbalance ratio of the datasets.To achieve this, I carried out K-Fold Stratified Cross-Validation with K=10.

Moving further,the Elbow method and the Silhouette method were used to identify the number of clusters in the dataset which are heuristics method to determine the optimal k value.The results showed agreement between the two methods. Then the optimal K value that was given by these methods was used to run K Means Clustering and the cluster with the lowest output criteria was chosen as the final cluster. Centroid were identified and the samples of the minority class in them were noted.

The next step performed was permutation test and the average and standard deviation of the results were summarised.

Finally, the results of baseline methods of Decision Tree and Random Forest Classification were compared with the results of Unsupervised learning tasks. Box plots of the cross-validation results were used to arrive at the conclusion which method is best under which conditions and how does the data imbalance affects the results and experiments in hand.

IV. Experiments

This sections details the experimental setup used for the assignment. I had the choice of implementing this project using various Python platforms like Jupyter, Spyder, Google Colaboratory, Visual Studio Code. Out of all these 4 options I decided to choose Google Colab which offers efficient, fast and easy to use online cloud Python platform and is linked to GITHUB which is not offered by other platforms.

The data was uploaded into the active session once the runtime is configured to GPU(Graphics Processing Unit) or TPU(Tensor Processing Unit). Since our task was quite light as we were dealing with non image data and deep learning was not required, NONE hardware accelerator was used runtime environment type and used for the experimental setup.

The different algorithms deployed for the experiments were Decision Tree Classifier, Random Forest Classifier. The train and test data were split in the ratio 80:20 to perform the experiments.

V. Discussion of Results & Conclusion

Appropriate Evaluation metrics were imported to assess the performance of the system and used for comparison. The metrics like Confusion Matrix, Recall, Precision, Accuracy , Cohen Kappa statistics were used to analyse and compare the experiment results.

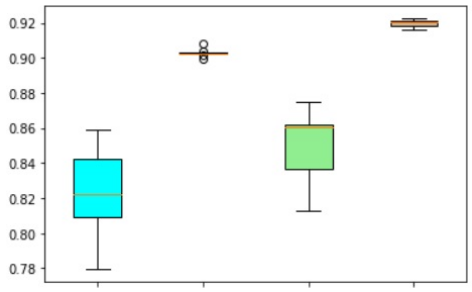

Fig. 4. RESULTS of BANK DATA

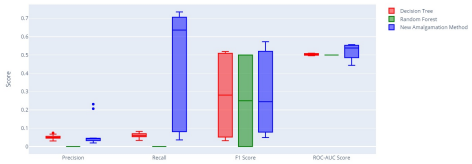

Fig. 5. RESULTS OF PORTO DATA

In the next part, we experimented to assign a sample x from the unseen fold to its closest cluster.A rule was formed that if the cluster had only one class instance, x was comfortably assigned that label. Otherwise, the new model(Permutation test model plus the supervised) trained with data from that cluster was used to assign a label to x.

Another kind of experiment was carried out by training a random forest for each of the clusters if those clusters contained samples from more than one class told us that this approach was indeed performing better given high values of evaulation metrics

DATASET 1

Fig 4 shows the results for the bank data which had 65% imbalance.During the supervised learning phase, when Decision Tree and Random Forest classification were carried out, Random Forest performed better than decision tree as expected since it uses bagging and ensemble techniques which improves its score. The higher Recall, Precision , F1 Score and Cohen Kappa values in case of Random Forest testify this conclusion. While in unsupervised learning phase , The Elbow method and Silhouette Method agreed in producing the optimal K value equal to 4 in low imbalance. When permutation test was done in conjuncton with supervised learning We got to know that for datasets with low imbalance the mixture of supervised and unsupervised gives the best results as evident from the figure.The performance is improved significantly when we use a mixture of both techniques.

DATASET 2

The Porto dataset also yields k = 4 in elbow and silhouette heuristics but there is performance drop when the

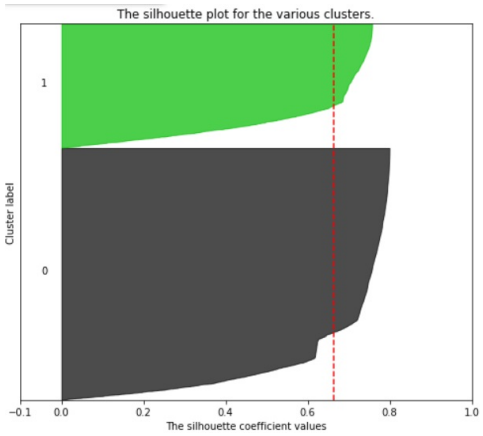

Fig. 6. Silhouette Plot of Credit card data

supervised tasks are combined with unsupervised than in single supervised learning cases which is evident from their lower scores of Recall , Precision and accuracy. Hence a good strategy in case of more than 90 % imbalance datasets would be to either go for supervised classification all alone or predictive analysis or performing unsuoervised clustering to the dataset to have better performance. This is because of overfitting the data and hence the data is biased and variance is more. This problem can be solved by using feature selection adn dimensonality reduction which would reduce bias and hence tune the data.

DATASET 3

Figure 6 displays the results of silhouette method on credit card fraud data with k = 2. Here also elbow and silhouette method gave optimal k = 4 and the same results were reproduced such as random forest is better than decision tree due to ensemble approach it follows and performance drops significantly when dealing with the amalgamation of k Means CrossValidation and single supervised techiques.

We can conclude many significant results from these experiments and project 2 -

1. Given binary label class , Low imbalance datasets perform better when supervised and unsupervised techniques are used together.

2. Given binary label calss, imbalanced datasets with more than 90 % imbalance shows performance drop in mixture of the methods due to overfitting.

3. Given high imbalance, one must do undersampling first and then apply ensemble methods for high performance.

4. Principal Component Analysis and other feature engineering yield better results in unsupervised methods.

5. Other classification ways like Support Vector Machines need to be explored because they can be used both for supervised and unsupervised machine learning.Also, SVMs are better in performance than Decision trees and Random Forests.

The above analyses summarise the depth and power of the entire experimental setup deployed and hence one can be sure of the results produced and draw meaningful and practical conclusions on methods to deal with imbalanced datasets under different environments, fields, time complexity and most important of all factors- PROBLEM STATEMENT in hand.

REFERENCES

[1] Wikipedia, "Silhouette (clustering) — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Silhouette\%20(clustering)&oldid=931344504, 2020, [Online; accessed 23-April-2020].

[2] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011.

[3] Wikipedia, "Downsampling (signal processing) — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Downsampling\%20(signal\%20processing)&oldid=943354289, 2020, [Online; accessed 23-April-2020].

[4] ——, "Upsampling — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Upsampling&oldid=943355736, 2020, [Online; accessed 23-April-2020].

[5] ——, "Data conversion — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Data\%20conversion&oldid=952427915, 2020, [Online; accessed 23-April-2020].

[6] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[7] Wikipedia, "Supervised learning — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Supervised\%20learning&oldid=949790121, 2020, [Online; accessed 23-April-2020].

[8] ——, "Semi-supervised learning — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Semi-supervised\%20learning&oldid=951900850, 2020, [Online; accessed 23-April-2020].

[9] ——, "Cross-validation (statistics) — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Cross-validation\%20(statistics)&oldid=952066808, 2020, [Online; accessed 23-April-2020].

[10] ——, "Decision tree learning — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Decision\%20tree\%20learning&oldid=952099189, 2020, [Online; accessed 23-April-2020].

[11] ——, "Random forest — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Random\%20forest&oldid=951450315, 2020, [Online; accessed 23-April-2020].

[12] ——, "Cluster analysis — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Cluster\%20analysis&oldid=952277367, 2020, [Online; accessed 23-April-2020].

[13] ——, "Elbow method (clustering) — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Elbow\%20method\%20(clustering)&oldid=952286034, 2020, [Online; accessed 23-April-2020].

[14] ——, "Resampling (statistics) — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Resampling\%20(statistics)&oldid=951039037, 2020, [Online; accessed 23-April-2020].