

**Machine learning model to predict asthma exacerbation using pollen concentration**

Course: ALY 6980 Capstone

By,

Gitesh Poudel

Pranshu Kumar Premi

Shashan Mukesh Pande

Shruti Avinash Pawar

Instructor: Valerie A. Atherley

27 June 2021

## Introduction

Asthma is a condition that causes breathing difficulties by factors that cause airway inflammation and swelling**.**Asthma symptoms include wheezing, shortness of breath and heavy chest. 12 million people experience exacerbations and 25% of them require to be hospitalized. Asthma can be fatal and thus early prediction of an exacerbation will be beneficial to the patient. There are various factors risk factors that can cause asthma exacerbations (Fergeson, Patel and Lockey, 2016). The goal of this project is to research about the factors that can cause these exacerbations, collate the required data, and build a Machine learning model with a good accuracy, precision, and recall. The tools that we will be using for this project are  excel and python notebook.

## Challenges and workaround

Our project objective is to predict whether a particular patient will face asthma exacerbation in certain environmental condition, especially pollen concentration in air. The first challenge we faced was data availability with the exact date of when the patient met asthma exacerbation. The second challenge was getting historical data from API for pollen concentration. The website <https://www.getambee.com/> has historical data only for up to 3 months and is a paid account. We were able to pull 100 records daily using free API calls. The API provided by <https://www.breezometer.com/> does not give historical data. Therefore, with whatever information we could draw, we simulated the pollen concentration looking based on aggregated pollen trend throughout the year in different part of USA. Since we are working on pitching a prototype, we thought of a workaround to create a dummy data set that is simulated for the variables we think are correlated to exacerbations found in literature review. One drawback of this workaround is that it does not completely replicate the real-world trend. The dataset in use does not represent a stratified sample of overall population but a random sample of synthetically generated instances. This prototype can be replicated by Keva health with real worlds observation and labeled instances collected from patient’s check in. To generated, collect, and utilize the data for purposed Machine learning model, Keva health needs to establish a data engineering pipeline. We have designed a framework of data and machine pipeline recommendation for KEVA health, the detail of recommended system is discussed in Appendix 1.

## Purpose Statement

Build a prototype Machine Learning classification model to predict whether individual patients will be impacted by asthma exacerbation in particular environmental condition represented by pollen concentration of trees, weed and grass, measured in range of 0 – 5, 0 being lowest and 5 means highest. The prototype model will be trained using synthetic patient’s data discussed in above section, which mimics the patient’s data points collected by Tele health service provider using dedicated telehealth service mobile application with location tracking capability.

## Research questions

1. What data points are required to predict asthma exacerbation in telehealth setting?
2. Which classification model is best to predict the exacerbation?
3. Will target class imbalance impact the model performance and accuracy?
4. Will the model be able to attain a good accuracy and F1 score?
5. Can the model be used for potential research by our sponsor Keva Health?
6. Do we need tuning? or can tuning help in increasing the accuracy of the model in general?

## Project Design

1. Curate a synthetic dataset, using random simulations and pollen concentration trend from aggregated pollen concentration trend through the year.
2. Perform exploratory data analysis to understand the patterns in the curated data.
3. Fit classification models to determine the best model to use and achieve a better predictive accuracy.

## Literature Review

**Advances in the clinical and mechanism research of pollen induced seasonal allergic asthma**

As stated in the article, the main reason to cause seasonal allergy-infused asthma is pollen. It causes allergic rhinitis that leads to increased risk of an exacerbation. There are more than 150 pollens that originate from trees, grasses, and weeds. One pollen that is said to affect patients worldwide during spring is Birch pollen. Humulus and Artemisia pollen are the types that are more prevalent in autumn. The concentration of pollen aeroallergens determines the severity of asthma. Allergic asthma is caused due to Ige dependent activation. Tests such as pollen provocation are conducted to check if a certain pollen can provoke Ige activation (Xie, Guan and Yin, 2019). Considering what we have summarized so far from this research article, pollen concentration will be a good predictor variable for asthma exacerbations.

**Racial Differences in Allergic Sensitization: Recent Findings and Future Directions**

Disparity in races is considered as one of the factors that can determine how sensitive a person is to allergic reactions. The level of sensitization determines the risk of asthma exacerbation. Socio economic and genetic factors of the races can be influential factors that cause the difference in sensitization. As per the study conducted by The National Health and Nutrition Examination Surveys (NHANES), in the age range of 6 – 59, at least one non-Hispanic black person was more likely to be positively tested in the skin prick test for allergy as compared to non-Hispanic white person. The allergens of the subjects include both food and pollen allergens. Another study conducted at Hartford Connecticut including 791 children facing asthma stated that African American were more likely to be sensitized to pollen from grass and trees. The paper summarizes various literature reviews and states that African Americans are more likely to be sensitized by aeroallergens as compared to White however, there were also studies or datasets available where African American children did not have a higher sensitization rate. (Wegienka et al., 2013)

**Sensitization patterns and association with age, gender, and clinical symptoms in children with allergic rhinitis in primary care: a cross-sectional study**

A nasal membrane disease that causes symptoms like congestion, itching and sneezing is called allergic rhinitis. It is caused by any of the aeroallergens by sensitization and is comorbid with asthma. Study indicated that both adults and children have atopic disease due to polysensitization. There were 699 patients that were a part of the study in this review out of which 89% were sensitized to pollen allergy and 69% out of them were children. As per the statistics we can see that sensitization is higher in children as compared to adults. The results also suggest that there is no significant difference in male and female sensitization to allergens.

**Detecting asthma exacerbations using daily home monitoring and machine learning, Journal of Asthma, DOI: 10.1080/02770903.2020.1802746**

In this journal, the authors attempted to leverage machine learning models to detect asthma exacerbation by using general health monitoring data. The motivation of this study came from recent respiratory disease and asthma studies, research found that the early detection of asthma exacerbation and patient awareness has helped reduce the acute result. Dataset used in this research comes from the SKURA study (NCT00839800), which contains information about the daily PEF (peak expiratory flow) and symptom scores recorded by participants. Data were records by each participant on their paper dairy, where they made notes of information such as PEF twice daily, morning symptom (scale 1 to 3, 1 being no symptom and 3 being severe), evening symptoms (same scale), no. of puffs of inhaler taken overnight and during the day and whether asthma disturbed their sleep previous night.

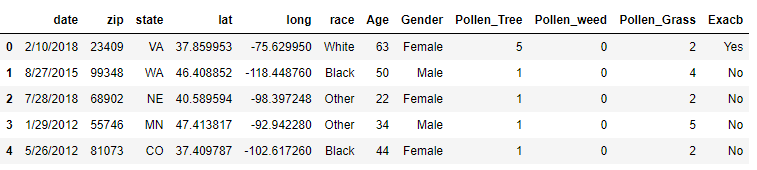
Dataset consisted of 728,535 records, recorded by patients themselves in 2010. Among the overall data, there were 576 cases of severe exacerbation events. The research used this data to train a model to predict the event of such a severe attack or any kind of asthma exacerbation happening on the same day or within 3 days. Before training the model, data were preprocessed using normalization, PCA, and standardization. Also, different classification algorithms like logistic regression, decision tree, naive Bayes, and perceptron algorithms were implemented and evaluated for model accuracy. They found out that logistic regression worked best, it had an area under the curve of 0.85.

In this research, we saw that health monitoring data with was collected manually in 2010 was a good enough predictor to predict the asthma exacerbation in advance, this implies the data generated based on patient’s check ins in telehealth app with conjunction of environmental data can be used to predict the asthma exacerbation.

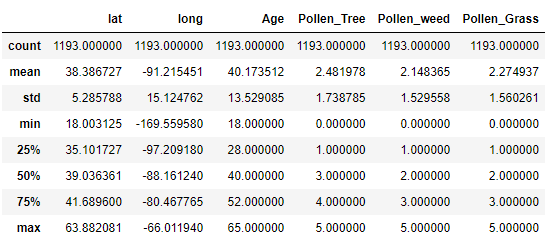
## Procedure

### Exploratory Data Analysis

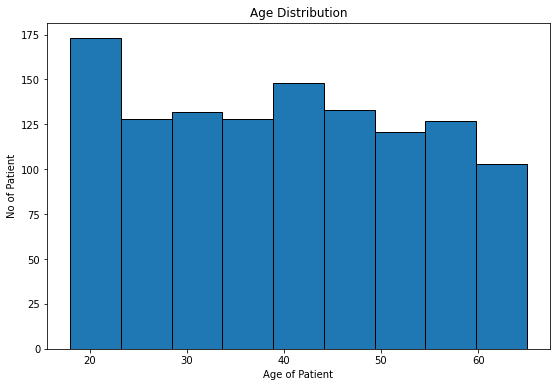
The data variables of our project dataset consist of demographic information of the patients who have faced exacerbations and the pollen concentration that we have curated as one of our predictor variables. Below is a snip of the dataset:

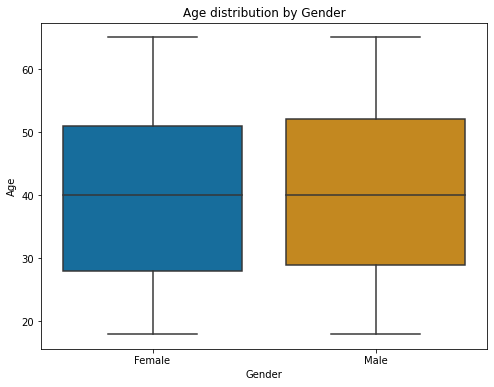


 The dataset has 12 variables and 1193 values. There are no null values in the data. Below is the descriptive summary of the data.

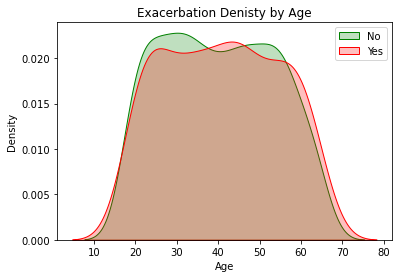


Pollen concentration is normalized in a range of 0 - 5, 5 being high and 0 being low. The average pollen concentration in tree weed and grass is around 2, which is considerably low as compared to 5. The standard deviation however is approximately 1.5 for pollen concentration with some areas present that have a pollen concentration of 5. The average age of patients in the dataset is 40 and it ranges from 18 to 65. This means that the data set is more focused on adult asthma patients. Below is the distribution of patient’s age:

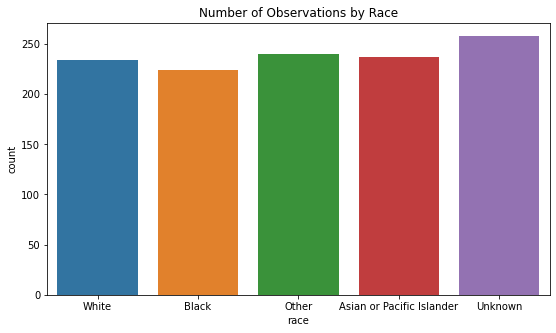




The gender wise age distribution is the same with average age of both male and female.



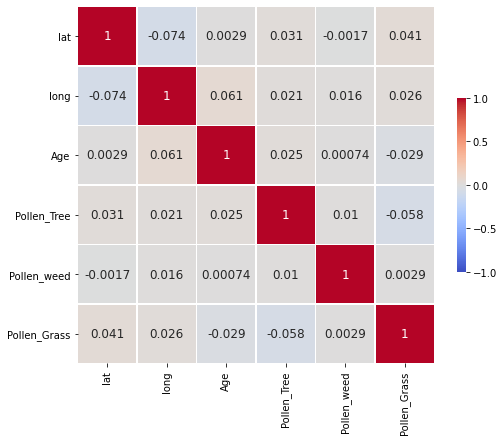
From the density plot of exacerbations by age, we can see that ‘No’ exacerbations are peaked at the age of 25 and 55 whereas the ‘Yes’ exacerbations are peaked at 45. The values are concentrated in almost all age groups.



The simulated data includes approximately equally distributed data entries for the respective races. To understand which state has the most concentration of asthma exacerbation, we have plotted heatmap.



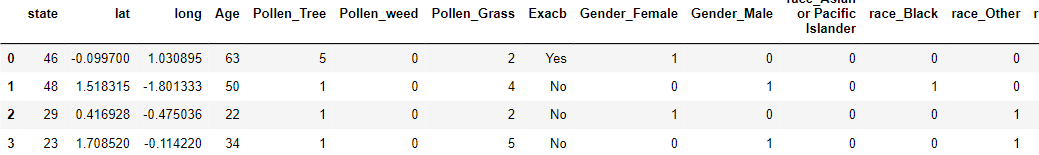
In the heatmap it is evident that most of the patient admissions for asthma are concentrated on the east coast. Also, the number of patients facing exacerbations is high on the east coast as compared to the west coast. We compared this with the asthma prevalence heatmap on the CDC website and found that the prevalence is high in Northeast area for adults. Thus, we can conclude that the simulated data is replicating the asthma related admissions trend to some extent.



From the correlation matrix it is evident that there is no strong positive or negative correlation between the predictor variables. This means there is no multicollinearity in the variables and will be good for the classification model.

### Data cleaning and preparation

There were no null values present in the data. Data variables such as gender, race and exacerbations were categorical variables. These variables were changed into numerical data for the classification model using StandardScaler in python. Below is the snip of the data prepared to feed in the model.



### Predictive Modeling

Our project objective is to classify whether a patient will get an asthma exacerbation or not based on the predictor variables. Out of the various supervised classification model that we tried; we came to a consensus to summarize below classification models for our curated data study:

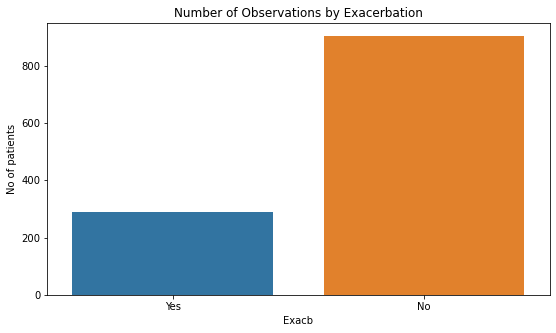
**Decision Tree**: A hierarchical tree-based model which uses parameters like dept of the tree that determines the limit till where the nodes of the tree will split (Seig, 2018). To train our decision tree model, we used DecisionTreeClassifier from Sklearn python library with the depth of node as ‘6’. The accuracy of the model is 83%.

**KNN:** KNN model runs on the assumption that similar data points exist close in proximity. The parameter of this algorithm is k which is the number of neighbors (Harisson, 2018). To train the KNN model, we used KNeighborsClassifier from SKlearn python library with k value as ‘5’. The accuracy of the model is 80%

**Support Vector Machine (SVM):** The reason this model was chosen is because it provides significantly good accuracy and the computation power required is less. The objective of SVM model is to find a hyperplane that divides data into classifications. This model works good even on multi-dimensional data (Gandhi, 2018). To train the SVM model, we used SVC from sklearn python library with kernel as linear. The model accuracy is 86% which is highest amongst the three models so far.

**Random Forest:** Decision tree is a building block of random forest model.It is an ensemble of multiple decision trees, and the parameter of this model is the number of trees. Random forest is a better model as compared to decision tree because splitting the data spreads the error across trees and the output is derived as an aggregate or maximum of the output of all trees. (Yie, 2019). To train the random forest model, RandomForestClassifier is used from Sklearn python library with tree size as 100. The accuracy is 86%.

To determine which model is a good fit for our curated data, using only the accuracy of the model will not be sufficient. It is evident from the below chart that there is a class imbalance in the exacerbations. The number of patients facing an exacerbation are comparatively lower than those not observing any exacerbations



Precision and Recall will be good indicators to determine which model is a better fit.

Precision = True Positive / (True Positive +False Positive)

Recall = True Positive / (True Positive + False Negative)

Precision is important the focus is to minimize the false positive and recall in important when the focus is on minimizing the false negatives. The goal of our project is to improve recall without diminishing the precision of the model (Brownlee, 2020). We have summarized the Precision recall and F1 scores along with accuracies in the below table:

**Table 1:** Comparison of model accuracy, precision, recall and F1 score before oversampling

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | | **Recall** | | **F1-Score** | |
| **Yes** | **No** | **Yes** | **No** | **Yes** | **No** |
| Decision Tree | 0.83 | 0.67 | 0.86 | 0.52 | 0.92 | 0.59 | 0.89 |
| KNN | 0.80 | 0.61 | 0.82 | 0.33 | 0.94 | 0.43 | 0.88 |
| SVM | 0.86 | 0.70 | 0.90 | 0.67 | 0.91 | 0.69 | 0.91 |
| Random Forest | 0.86 | 0.77 | 0.88 | 0.59 | 0.95 | 0.67 | 0.91 |

To reduce the imbalance in the data, we have performed oversampling of the data using the SMOTE library in python. Below are the outputs of the models based on the oversampled data.

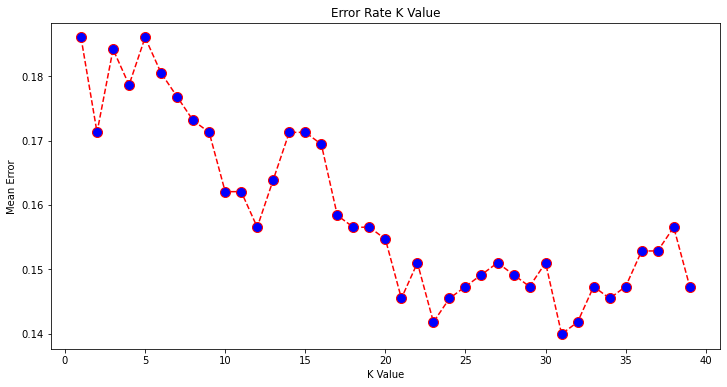
**Table 2:** Comparison of model accuracy, precision, recall and F1 score after oversampling

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | | Recall | | F1-Score | |
| Yes | No | Yes | No | Yes | No |
| KNN | 0.82 | 0.92 | 0.76 | 0.71 | 0.93 | 0.80 | 0.84 |
| SVM | 0.89 | 0.94 | 0.85 | 0.84 | 0.94 | 0.89 | 0.89 |
| Random Forest | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 |

From table 2 we can see that the recall has increased without hurting the precision.

As of now, Support Vector machine is the best model in terms of accuracy, precision, and recall.

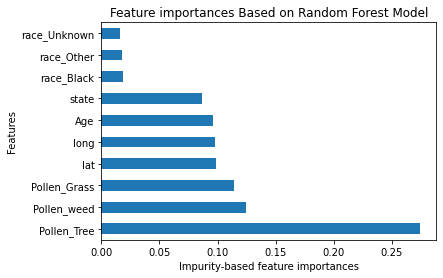
Our focus is now to optimize KNN and Random Forest model to see if there is any change in the accuracy to beat the SVM model.

**KNN tuning:** To check for an optimum value of k, we checked for the error generated by the KNN model for k values between 1 to 40 and plotted the below chart:  


From the chart it is evident that the error value is lowest when k value is approximately 31. We therefore used a revised k value as 31 and run the model. The accuracy improved to 85%.

Further, we performed a grid search to determine the tree size and max dept of the random forest model. The best parameters determined are tree size ‘500’ and node split dept of ‘8’. There was no change observed in the model accuracy even after tuning the parameters.

To understand which feature is impacting the tree-based model the most, we plotted a feature importance chart and found the below result:



As seen in the feature importance chart, race is the least important feature in terms of exacerbation prediction whereas pollen produced from tree is the highest impacting feature in the model.

**Table 3:** Comparison of model accuracy, precision, recall and F1 score after oversampling and parameter tuning.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | | Recall | | F1-Score | |
| Yes | No | Yes | No | Yes | No |
| KNN | 0.82 | 0.94 | 0.75 | 0.69 | 0.95 | 0.80 | 0.84 |
| SVM | 0.89 | 0.94 | 0.85 | 0.84 | 0.94 | 0.89 | 0.89 |
| Random Forest | 0.88 | 0.90 | 0.86 | 0.86 | 0.90 | 0.88 | 0.88 |

From table 3, it is evident that the tuned KNN and Random Forest model accuracies are still lower than the SVM model. In terms of precision and F1-score, SVM is performing the best and has a recall of 0.84.

## Conclusion

In this project, we demonstrated that it is possible to predict the event of asthma exacerbation of individual patients based on environment factor like pollen concentration from different plants, given that historical data is preserved and used as knowledge based on trained model. The successful working of such model implies that, the forecasted (3 days in advance) pollen concentration can be used to predict and warn patients about the possible difficulties beforehand and generated suggested based on specific case.

We also observed that, for this kind of data and data points, a tree-based ensemble model, random forest classifier is the best algorithm to be used with respect to model performance metrics and accuracy.

## Appendix

**Appendix 1:** Data Analysis infrastructures and pipeline recommendations

Major challenges we experienced in this project was lack of availability of case specific labeled data required for training. To be able to establish a knowledge base to predict the event of asthma exacerbation with respect to patient’s information and environmental factors, i.e., pollen concentration, we need labeled dataset with patient’s specific data. An ideal dataset for such model should includes patients’ demographic information (Age, sex, race), environmental factor (pollen concentration) and data label reflecting the magnitude of asthma exacerbation( i.e., severity or binary indicator whether it was manageable or not).

Below is the data pipeline and ML pipeline recommendation for Keva Health, which is intended to keep record of patients, manage data, and use it to develop a knowledge based capable of predicting future instances with the help of forecasted pollen concentration datapoint.

Diagram

Description automatically generated

Data Pipeline:

* Patient’s Demographics: Information such as Patient’s age, race, sex, and unique identifier id. This information can be accessed from patients account with Keva Health
* Patients’ Check ins: Information collected every time a patient check-in in Keva health‘s Mobile application
* GPS reading: In every check in event, GPS (latitude and longitude) readings should be recorded
* API data channel: BreezoMeter Pollen API uses GPS reading to retrieve Pollen concentration for day and location
* API data channel: Google Reverse Geo coding API can be used to transform latitude and longitude reading to state and zip code of patient, which will be useful for exploratory data analysis
* All the above datapoints from different channel needs to be combined to create a one instance records
* Similarly, multiple customer check ins record should be stored in data warehouse.

ML Pipeline:

* Data collected in data warehouse is extracted, transformed, and preprocessed for machine learning application.
* Raw data can also be used to conduct exploratory data analysis
* Preprocessed data can be used to model a supervised classifier and unsupervised clustering
* Cluster based on patient’s data can help us generate association between predictor and event of asthma exacerbation.
* A trained classification model will then be used with forecasted pollen concentration from API channels to predict the possible asthma difficulties. Based on prediction, suggestion to manage asthma should be communicated to patients.

## Reference

Fergeson JE, Patel SS and Lockey RF. 21 August 2016. Acute asthma, prognosis, and treatment. Retrieved on 24 June 2021 from <https://www.jacionline.org/article/S0091-6749(16)30800-4/fulltext>

Xie ZJ, Guan K and Yin J. 15 February 2019. Advances in the clinical and mechanism research of pollen induced seasonal allergic asthma. Retrieved on 24 June 2021 from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420698/>

Wegienka G, Johnson CC, Zoratti E and Havstad S. 13 June 2013. Racial Differences in Allergic Sensitization: Recent Findings and Future Directions. Retrieved on 24 June 2021 from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888051/>

MA de Bot C, Roder E, Pols DHJ, Bindels PJE, Wijk RG, van der Wouden JC, Moed H. 4 March 2013. Sensitization patterns and association with age, gender, and clinical symptoms in children with allergic rhinitis in primary care: a cross-sectional study. Retrieved on 25 June 2021 from <https://www.nature.com/articles/pcrj201315>

Seif, G. 30 November 2018. A Guide to Decision Tree for Machine Learning and Data Science. Retrieved on 26 June 2021 from <https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956>

Olivier Zhang, Leandro L. Minku & Sherif Gonem (2020) Detecting asthma exacerbations using daily home monitoring and machine learning, Journal of Asthma, DOI: 10.1080/02770903.2020.1802746

Harrison, O. 10 September 2018. Machine Learning Basics with the K-Nearest Neighbors Algorithm. Retrieved on 26 June 2021 from <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

Gandhi, R. 7 June 2018. Support Vector machine – Introduction to Machine Learning Algorithms. Retrieved on 26 June 2021 from <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

Yie, T. 12 June 2019. Understanding Random Forest. Retrieved on 27 June 2021. Retrieved on 26 June 2021 from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Brownlee, J. 3 January 2020. How to Calculate Precision, Recall and F1 measure for imbalanced Classification. Retrieved on 27 June 2021 from <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/>