## Predictive Analysis

Predictive analytics encompasses a variety of statistical techniques but we chose to use [machine learning](https://en.wikipedia.org/wiki/Machine_learning) to analyze current and historical facts, which would allow us to make predictions about future or otherwise unknown events.

Our choice regarding this project was between supervised and unsupervised learning, since this project is made for learning purpose we chose to implement both, to see what each of them are good at and why we should consider them.

**Supervised learning**

Overview

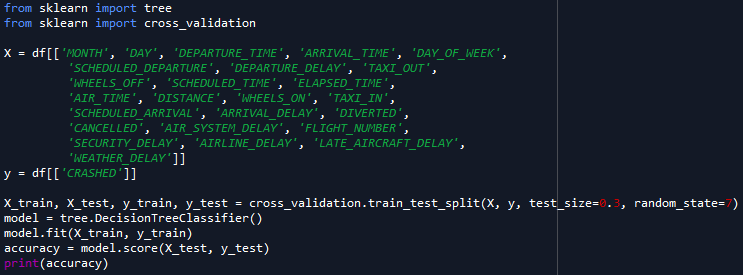
Supervised learning maps an input to an output based on example input-output pairs. It uses labeled data consisting of features (X- input object) and labels (y- a desired output). A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples. In best case scenario this will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

* Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Examples of Supervised Learning:

* Linear Regression
* Logistic Regression
* Decision Tree
* SVM (Support Vector Machine)
* Naive Bayes
* kNN (k- Nearest Neighbors)
* Random Forest

Our code:



As shown above we started by importing necessary libraries for job and then continued to separate data frame into features and labels. Than we used cross\_vaidation to split this for training purpose (70% training/30% testing). After this we trained the algorithm and then tested the results receiving the score of 99,53% accuracy. Although the resulting score may seem very accurate, we determined that it is this precise only because the possibility of animal collision is tiny, making the algorithm very biased towards plain not crashing.

We decided to use decision tree, because it is good solution to solve classification problems, which we are facing, as we want to know whether the plain will or will not crash on a specific flight.

Some other algorithms we investigated were:

**Naive Bayes**, which we opted not to use, as although is fast and highly performing algorithm for machine learning, it is really biased to often recurring results, which would mean the algorithm would just consider all flights to not crash, as it is most represented result.

**Logistic Regression,** it is used to estimate discrete values (like 0/1, yes/no, true/false) based on given set of independent variables. In simple words, it predicts the probability of occurrence of an event. Since, it predicts the probability, its output values lie between 0 and 1. We considered this but opted to use decision tree as we wanted to see clear answer yes or no.

**Unsupervised learning**

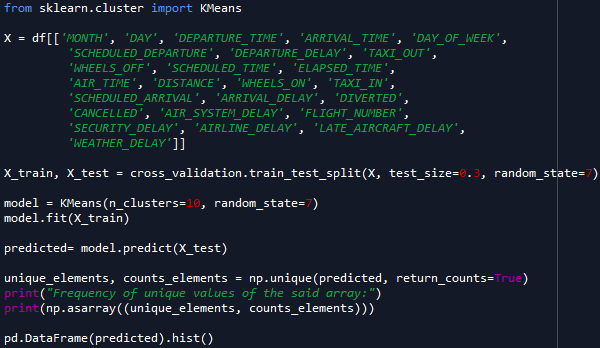
Overview

Unsupervised learningdescribes hidden structure of "unlabeled" data. It can be divided in following categories: [Clustering](https://en.wikipedia.org/wiki/Data_clustering), [Anomaly detection](https://en.wikipedia.org/wiki/Anomaly_detection) and [Neural Networks](https://en.wikipedia.org/wiki/Artificial_neural_network). Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the algorithm, it is used to predict pattern in data —which is one way of distinguishing from supervised.

* Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)
* Anomaly detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically, the anomalous items will translate to problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.
* Neural networks are computing systems inspired by the biological neural networks that constitute animal brains. Such systems "learn" (i.e. progressively improve performance on) tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the results to identify cats in other images. They do this without any a priori knowledge about cats, e.g., that they have fur, tails, whiskers and cat-like faces. Instead, they evolve their own set of relevant characteristics from the learning material that they process.

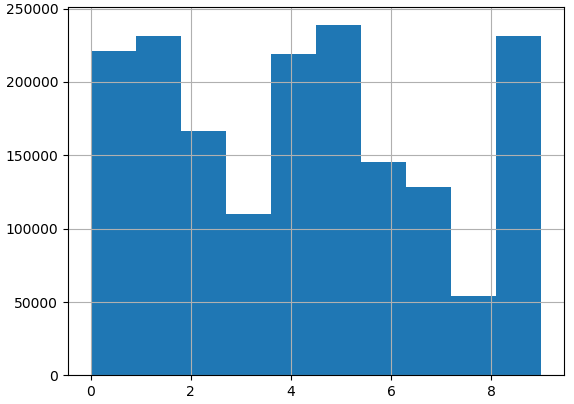
Examples of Unsupervised Learning:

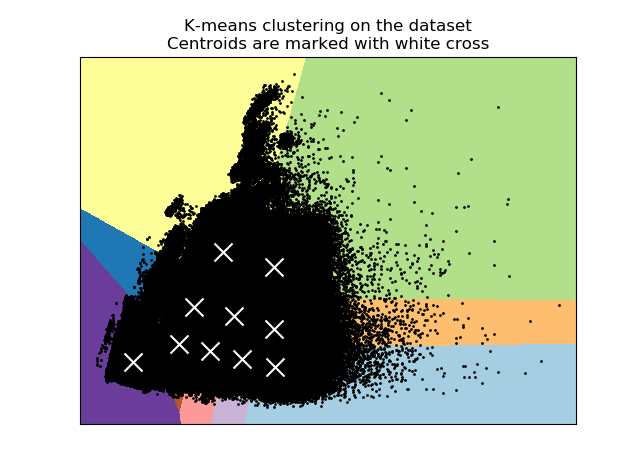
* K-Means

Our code:

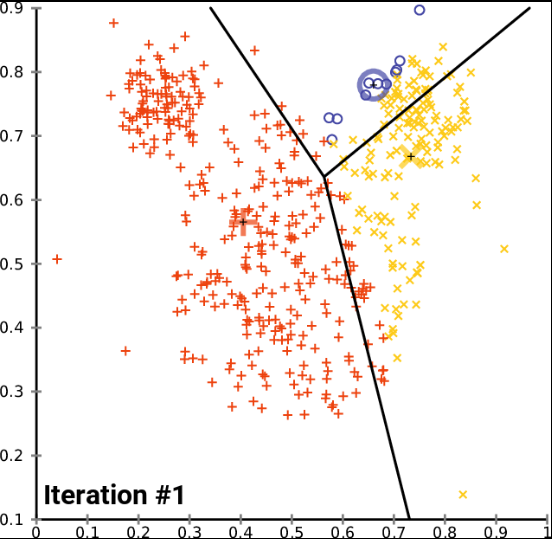
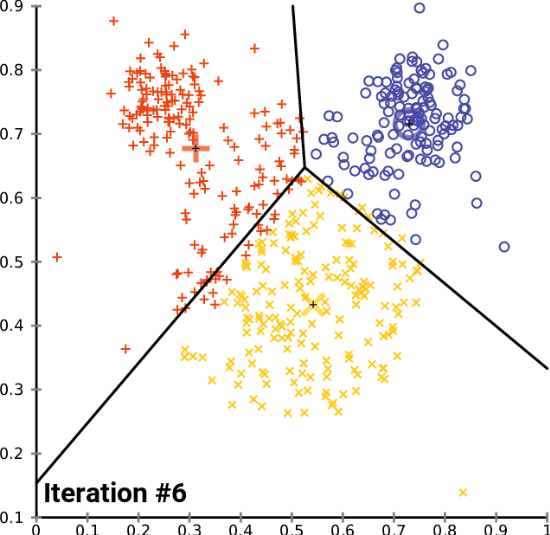
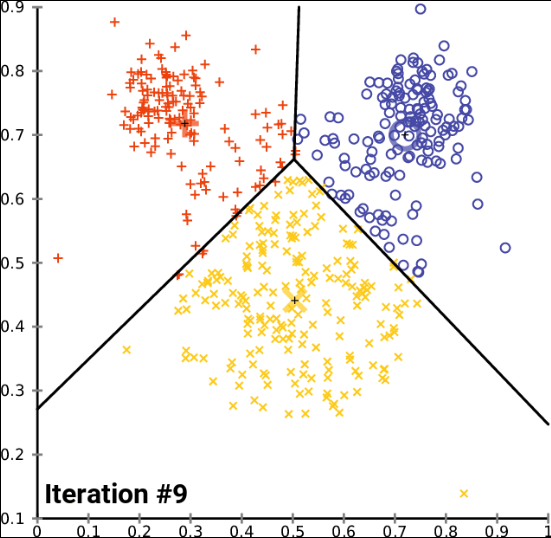
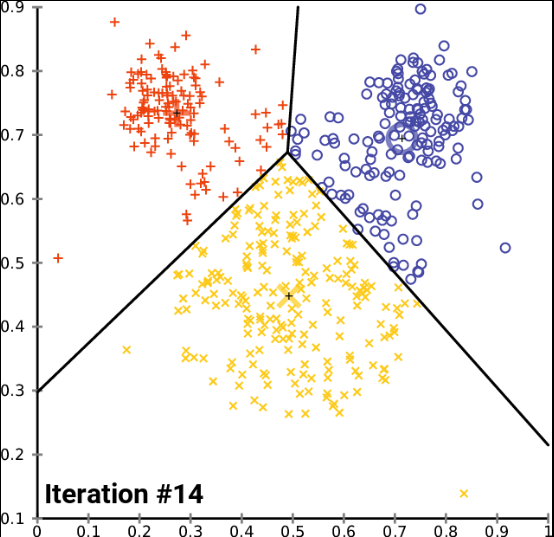
As for supervised learning, for unsupervised learning we also started by importing necessary libraries for job and then we used cross\_vaidation to split this for training purpose (70% training/30% testing). After this we trained the algorithm and then tested the results by predicting in which cluster it would fit. After executing this the results are following:





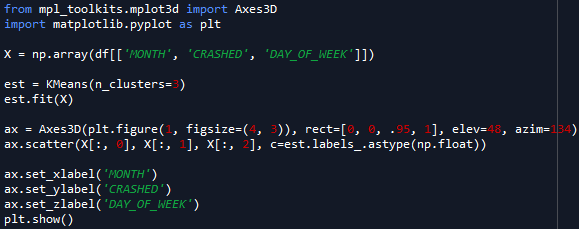


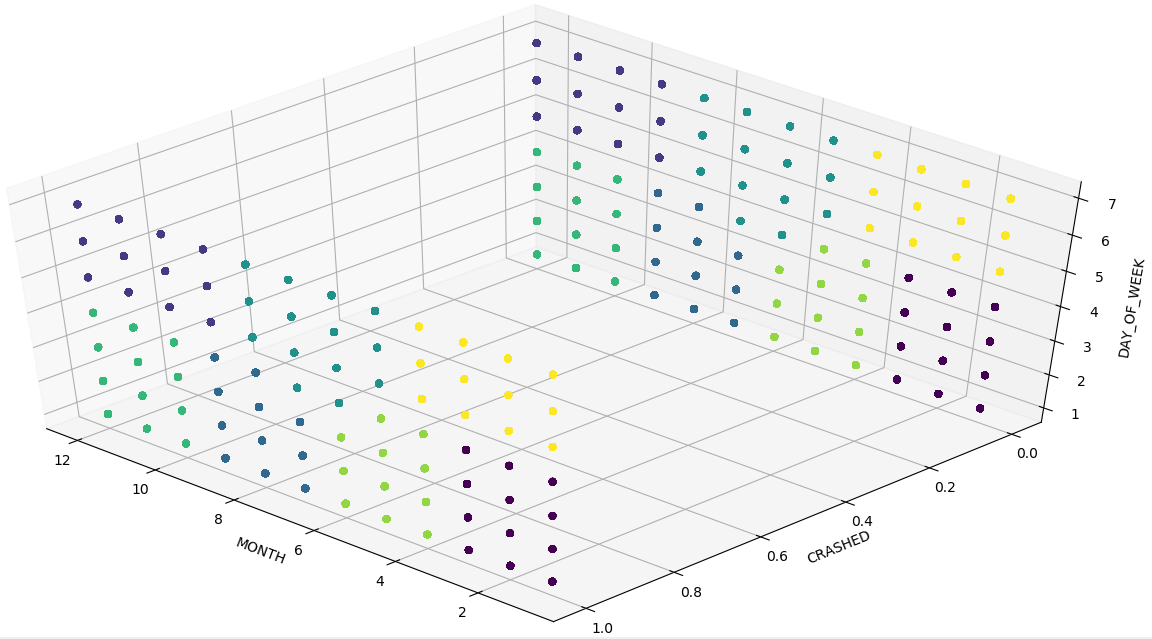
As can be seen from results shown above, some cluster are bigger than others, for example cluster nr. 9 has almost 5 times more values than cluster nr.8.

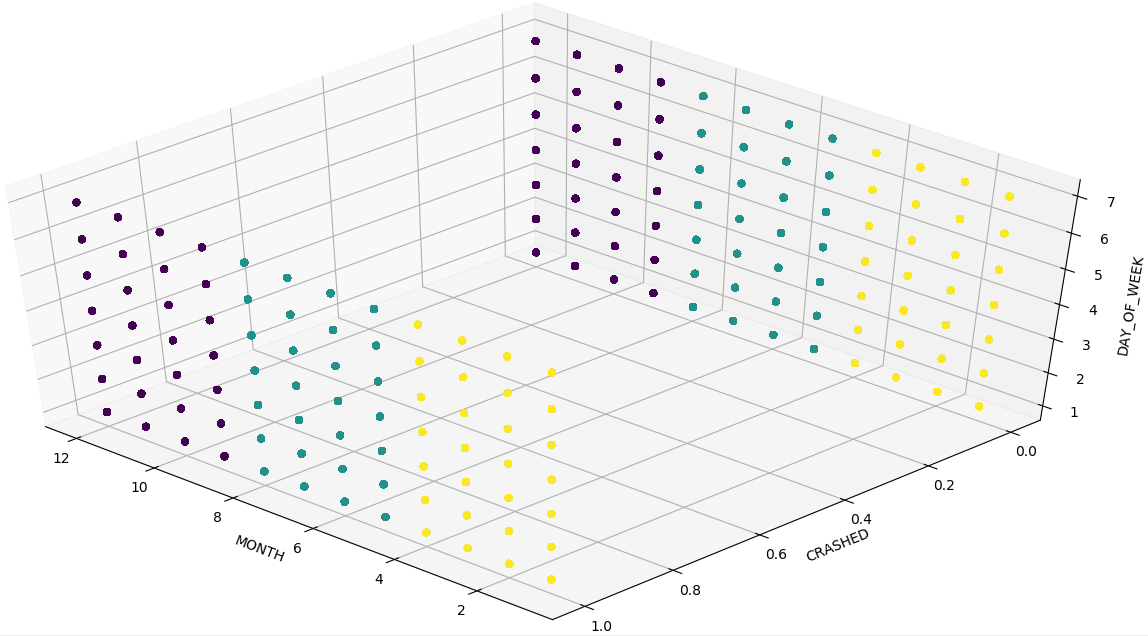
We decided to use K-Means, because it is good solution to solve problems related to unsupervised learning, as it clusters data to allow us to see correlations which otherwise we could have looked over.

As a bonus we also decided to visualize the work done by unsupervised learning algorithm to see how day of the week, month and crushed values match with each other.





10 clusters



3 clusters

**Final thoughts**

We overestimated our skills when choosing the projects idea, even though we achieved our goals for this project we should have done more research regarding our idea. The result of 99.5% accuracy may seam high but we determined that this is basically the same as ratio of flight crushing to not crushing, to make our prediction more accurate we would need more informative datasets and other data which we didn’t use in making of this project(some of the data we could have used to increase the efficiency of prediction could be: weather patterns, information about birds and they’re flight patterns and many more)