Data Analytics;

More prominent than ever

Christian Solomon

Computer Science, Minnesota State University Moorhead, 1104 7th Avenue South, Moorhead, MN 56563

Scope

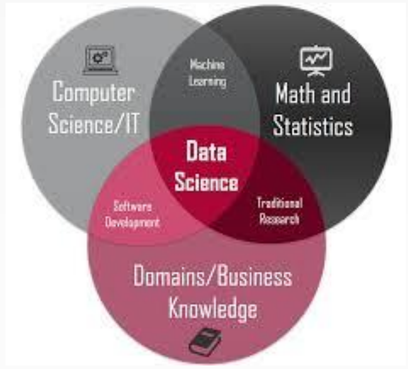
- How can we use different algorithms for different problems?
- What is the information we are trying to discover from the dataset?
- What kind of datasets will be analyzed?

What is Data Analysis?

- Data analysis is a process of inspecting, cleansing, modeling, and transforming data to discover useful information.
- Data is analyzed using algorithms that are made more efficient using different computer programs.

Why Research Data Science?

- Analyzing data results in numerous beneficial findings that play an important role in not only the technology industry but also society as a whole.
- By successfully finding the most efficient algorithms, analyzing specific datasets will be much simpler, which in turn results in better analysis.



The Dataset

Information

- The dataset used for this research was initially contributed to the UCI Machine Learning repository nearly 30 years ago.
- Mushroom hunting is a phenomenon that has reached major popularity in recent years

Content

- This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family.
- Each species is defined into three:
  - Definitely Edible
  - Definitely Poisonous
  - Of Unknown Edibility

Analysis

- After analyzing the data, we aim to answer two primary questions:
  - Which algorithm/model works best for this dataset?
  - What is the primary determinant in a mushrooms toxicity?

Results

- In order to choose the best model, three Machine Learning algorithms were used to test the set:
  - Naïve Bayes Classification
  - Logistic Regression Classification
  - Random Forest Classification
- Random Forest classifier was the best model, which achieved 100% accuracy with the test set.
- We were then able to deduce the most determinant features:

```
In [46]: z = zip(clf.feature_importances_,X_train.columns)
         z.sort(reverse=True)
         z[:10]

Out[46]: [(0.15844522158060251, 'odor_f'),
          (0.072093232716836098, 'gill-size_n'),
          (0.071449650799149014, 'ring-type_p'),
          (0.059524344656014208, 'stalk-surface-below-ring_k'),
          (0.054395896612936, 'gill-color_b'),
          (0.053292416415563093, 'odor_n'),
          (0.051462205469969005, 'stalk-root_e'),
          (0.037758414413626332, 'odor_p'),
          (0.037439645501368912, 'stalk-surface-above-ring_k'),
          (0.033770321762183406, 'odor_c')]
```

Conclusion

- In conclusion, Data science is a broad topic that is contributing a lot in society today.
- Many algorithms exist, therefore it is important to know which algorithm is most efficient for each data set. For instance:
  - For prediction (e.g., linear regression, logistic regression)
  - For classification (e.g., decision trees, random forest)
  - Time-series forecasting (e.g., regression-based)

FAQ's

- What is the best algorithm for data analytics?
  - There is currently no "one wins all" algorithm that works better than others for each dataset, different algorithms work best for different datasets.
- What is the best way to find the most efficient algorithms?
  - Firstly, we must understand the dataset and know what we kind of information we are trying to fin from the data. Only then will we be able to determine which models/algorithms we should test with.
- Is data analytics important? If so, why?
  - Yes!!! Data analytics is the future! Analyzing data is so important in today's world not only for the technology industry but for science as a whole. By analyzing data we can find out relevant information such as safe/deadly mushrooms, chances of individuals getting cancer, predict weather and much more