# RANKING SYSTEM FOR BIOFUEL PRODUCTION

**Prepared by**

Michele De Filippo, PhD
Data Scientist

michele.defilippo@alumni.ust.hk

**July 9, 2020**

## 1. Project Summary

One of the challenges we face in this century is to shift from fossil fuel energy to renewable energy. Biofuels are one of a few solutions to continuous rising of oil prices, greenhouse gas emissions reduction and depleting oil reserve. To produce an acceptable quality biofuel and to ensure that its production is sustainable, we are faced with a challenge of identifying the right kind of feedstock to use at a given site. The answer to this problem could be found in second generation biofuels produced from agroindustry residues, as its use does not compete with food resources. Furthermore, it allows exploitation of raw materials with low commercial value and offers an alternative to their disposal which again incurs cost.

Scientists have reported the potential use of several plant species for biofuel production. However, to date no one has evaluated these feedstocks for all the characteristics that are required to be tested, including fuel properties, engine performance, emission characteristics and production potential. Parameters to be used to evaluate biodiesel quality are established by Fuel Quality Standards (Biodiesel) Determination 2019 in Australia and the European Standard EN 14214 in Europe. Cetane number or the Flash Point are two examples of these important parameters. Other important parameters are the production efficiency such as conversion rate of oil to biodiesel, oil content or yield. The yield per hectare of the crop is also an important parameter for biofuel production as the perfect solution for one region is not the same for another region.

This proposal will explore an innovative and new approach for ranking feedstocks. It will use secondary data sources from the published literature. Online data mining is performed using Scopus and other sources and the collected data will be modelled using a Deep Neural Network (DNN) system to rank the feedstock.
The idea is to select feedstock properties (the most informative ones) and to use the selected feedstocks to design a Multi-criteria Decision Analysis (MCDA) ranking method to rank. The output of this step is the set of feedstocks in which one more column represents the rank of each feedstock. In this process, more than 20 biofuel parameters (e.g. Oil content, oil properties, biodiesel properties, including cetane number) for 100 feedstocks will be considered. We use this to train an DNN system that can be used to rank new (unseen) set of feedstocks based on the full set of properties.

The above ranking will first be made for a given location, such as Rockhampton based on its agroclimatic features (e.g. rainfall, temperature, light intensity etc). The proposed method/model can then be deployed to other sites (e.g. Brisbane, Cairns, Townsville) by replacing the agroclimatic parameter in the DNN algorithm.

Briefly, the MCDA process that is used to set priorities among different attributes, and this is based on three principles: structure of the model; the second, comparative judgment of the feedstocks and synthesis of priorities. Three techniques are used for creating the rankings, namely Weighted Sum Model (WSM), Weighted Product Model (WPM) and TOPSIS.

The ranking output needs to be predicted through usage of a suitable Machine Learning (ML) or Deep Learning (DL) approach, and the data is randomly selected for training, validation, and testing. Several techniques are used, including Multi-variate Regression (MVR), Deep Neural Networks (DNN) and Multi-layer Perceptron (MLP), to determine the best performance of the model. Open source frameworks in Python will be used to build the models, including Tensorflow, Keras and SCIKIT-LEARN packages.

This model can then be used to rank unknown feedstocks. Therefore, trial and error play a significant role in this process. An MCDA+DNN based system for prediction and identification of superior feedstocks is a novel technology that can be used to automatically rank feedstocks for their suitability for biofuel production at a given location.

Figure 1 gives a broad overview of the described scope of work.

*Proposed data model for ranking system*

*Using MCDA and Deep Neural Network to develop a ranking system*

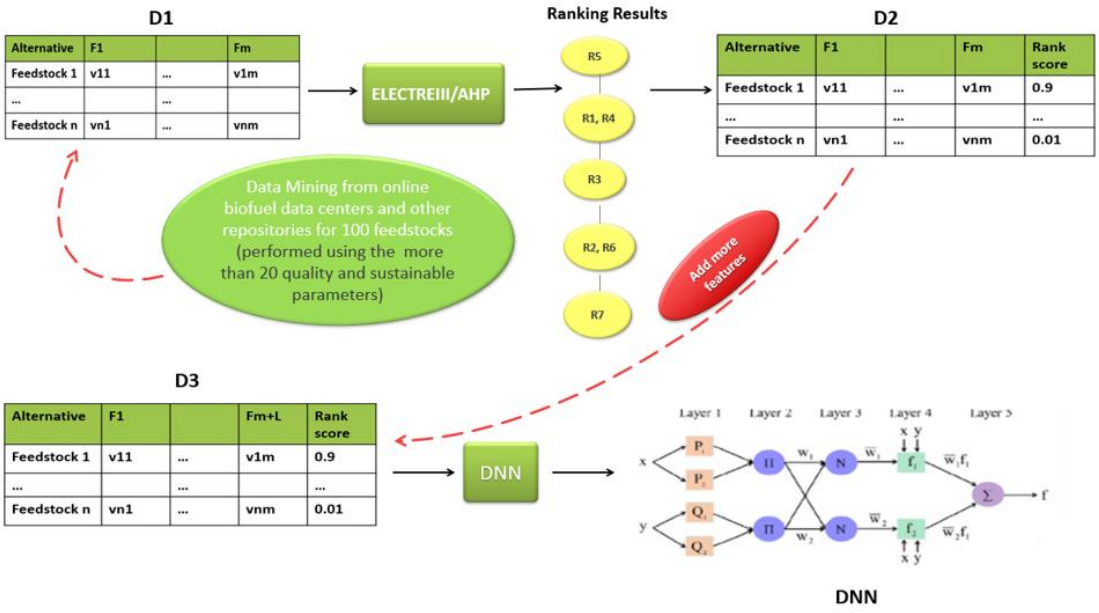



**Figure 1: Screenshot of the Provided Data**

## 2. Challenges

- It was expected that the data was going to be in a numeric format and no pre-processing or cleaning was required. Instead, the delivered data required pre-processing. Pre-processing and cleansing are done at no additional charge. A csv file (named DATA_MCDA_DNN.xlsx) with the cleansed data, in numeric format, is provided along with this report.
- The provided data was incomplete. The amount of NaN readings was counted, and it is reported in Table 3. Prior to starting the project, both parties were aware that the quality of the data may not be sufficient to obtain satisfactory results. Both parties have agreed that this project has only an exploratory purpose. The columns with 100% NaN readings are discarded, while remaining NaN readings are filled with mean values. The final outcome of the project is expected to be affected by the data quality.
- WSM, WSP and TOPSIS are chosen as MCDA methods. In WSM, WSP and TOPSIS, weight computation plays a key role. Each criterion is given a weight according to the importance the user wants to give to one category rather than another. The final results are very sensitive to weights assignment, hence the choice of such parameters is advised by experts in the field. For this project, three classes of weights have been provided and the three different ranks have been created, one per each class of weights, where each rank contains results of rankings performed through WSM, WSP, TOPSIS and an average of the three method.
- The output of the ranking system is then fed into ML and DL algorithms. Three approaches are used in an increasingly complex models.

| Column | NaN Counts | Percentage NaN values |
|---|---|---|
| Country | 0 | 0 |
| State | 0 | 0 |
| Species Type | 0 | 0 |
| Species Name (feedstocks/plants) | 1 | 0.006452 |
| Scientific Name | 18 | 0.116129 |
| fruit Yield (kg / hectare / year) | 129 | 0.832258 |
| Palmitic acid (C16:0) | 131 | 0.845161 |
| Stearic acid (C18:0) | 127 | 0.819355 |
| Oleic acid (C18:1cis) % | 124 | 0.8 |
| Lenoleic acid (C18:2) | 126 | 0.812903 |
| Lenolenic acid (C18:3) | 128 | 0.825806 |
| Eicosonoic acid | 155 | 1 |
| Behinic acid (C22:0) | 150 | 0.967742 |

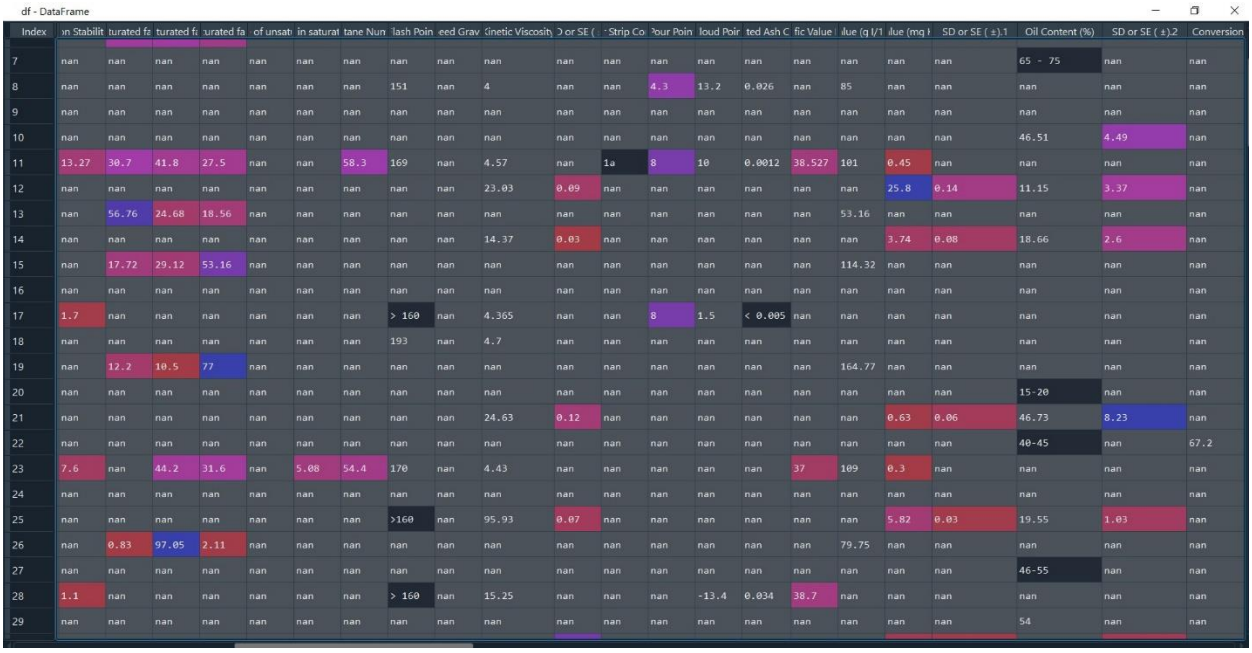
| | | |
|---|---|---|
| Lauric Acid (C12:0) | 130 | 0.83871 |
| Oxidation Stability (hour) | 125 | 0.806452 |
| Total saturated fatty acid | 139 | 0.896774 |
| Total monounsaturated fatty acid (MUFA) | 133 | 0.858065 |
| Total polyunsaturated fatty acid (PUFA) | 133 | 0.858065 |
| Degree of unsaturation | 155 | 1 |
| Long Chain saturated Factor | 149 | 0.96129 |
| Centane Number | 142 | 0.916129 |
| Flash Point | 109 | 0.703226 |
| Speed Gravity | 155 | 1 |
| Kinetic Viscosity | 89 | 0.574194 |
| SD or SE ( Â±) | 134 | 0.864516 |
| Copper Strip Corrosion | 151 | 0.974194 |
| Pour Point | 135 | 0.870968 |
| Cloud Point | 135 | 0.870968 |
| Sulphated Ash Content | 139 | 0.896774 |
| Calorific Value MJ/Kg | 129 | 0.832258 |
| Iodine value (g I/100g) (IV) | 117 | 0.754839 |
| Acid value (mg KOH/g) | 118 | 0.76129 |
| SD or SE ( Â±).1 | 133 | 0.858065 |
| Oil Content (%) | 93 | 0.6 |
| SD or SE ( Â±).2 | 130 | 0.83871 |
| Conversion Rate | 144 | 0.929032 |
| Power | 155 | 1 |
| Torque | 155 | 1 |
| Specific Fuel Consumption | 155 | 1 |
| Compression ration | 154 | 0.993548 |
| Swept Volume | 155 | 1 |
| Clearance volume | 155 | 1 |
| Power output | 155 | 1 |
| Fuel Density | 94 | 0.606452 |
| Lower heating value | 155 | 1 |
| Higher heating value (HHV) | 138 | 0.890323 |
| Mechanical Efficiency | 155 | 1 |
| Mean effective pressure | 155 | 1 |
| Road Load power | 155 | 1 |
| Indicated power | 155 | 1 |
| Effective power | 155 | 1 |
| Stroke Volume | 155 | 1 |
| Compression ratio | 155 | 1 |
| Injection pressure | 155 | 1 |
| Fuel type | 155 | 1 |

| | | |
|---|---|---|
| Power and Mechanical Efficiency | 155 | 1 |
| Fuel Air-Ratio (AFR) | 155 | 1 |
| Volumetric Efficiency | 155 | 1 |
| Specific Output | 155 | 1 |
| Specific Fuel Consumption.1 | 155 | 1 |
| Thermal Efficiency and Heat Balance | 155 | 1 |
| Exhaust Smoke and Emissions | 155 | 1 |
| Effective Pressure and Torque | 155 | 1 |
| CO2 | 155 | 1 |
| CO | 155 | 1 |
| O2 | 144 | 0.929032 |
| NOx | 155 | 1 |
| SO2 | 155 | 1 |
| Unburned hydrocarbon (THC) | 155 | 1 |
| Particulate matters (PM/soot) | 155 | 1 |
| Non-methane hydrocarbon (NMHC) | 155 | 1 |
| Seed Separation | 155 | 1 |
| Oil Refining | 155 | 1 |
| Oil Conversion | 155 | 1 |
| Biodiesel Storage | 155 | 1 |
| Biodiesel Discount | 155 | 1 |
| Biodiesel Tax | 155 | 1 |
| Climatic Conditions | 155 | 1 |
| Soil Conditions | 144 | 0.929032 |
| Stress Factors | 155 | 1 |
| Processing | 155 | 1 |
| Agronomy | 155 | 1 |

**Table 3: NaN Readings in the provided data**

## 3. Data Cleansing & Exploration

A screenshot of the provided data is given as follows. The data size is 155x82.




| Index | on Stabilit | turated fa | turated fa | urated fa | of unsati | in saturat | tane Num | lash Poin | eed Grav | Kinetic Viscosity | D or SE ( | Strip Co | Pour Poin | loud Poir | ted Ash C | fic Value | ilue (g/l | ilue (mg | SD or SE ( ±).1 | Oil Content (%) | SD or SE ( ±).2 | Conversion |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 65 - 75 | nan | nan |
| 8 | nan | nan | nan | nan | nan | nan | nan | 151 | nan | 4 | nan | nan | 4.3 | 13.2 | 0.026 | nan | 85 | nan | nan | nan | nan | nan |
| 9 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan |
| 10 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 46.51 | 4.49 | nan |
| 11 | 13.27 | 30.7 | 41.8 | 27.5 | nan | nan | 58.3 | 169 | nan | 4.57 | nan | 1a | 8 | 10 | 0.0012 | 38.527 | 101 | 0.45 | nan | nan | nan | nan |
| 12 | nan | nan | nan | nan | nan | nan | nan | nan | nan | 23.03 | 0.09 | nan | nan | nan | nan | nan | nan | 25.8 | 0.14 | 11.15 | 3.37 | nan |
| 13 | nan | 56.76 | 24.68 | 18.56 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 53.16 | nan | nan | nan | nan | nan |
| 14 | nan | nan | nan | nan | nan | nan | nan | nan | nan | 14.37 | 0.03 | nan | nan | nan | nan | nan | nan | 3.74 | 0.08 | 18.66 | 2.6 | nan |
| 15 | nan | 17.72 | 29.12 | 53.16 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 114.32 | nan | nan | nan | nan | nan |
| 16 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan |
| 17 | 1.7 | nan | nan | nan | nan | nan | nan | > 160 | nan | 4.365 | nan | nan | 8 | 1.5 | < 0.005 | nan | nan | nan | nan | nan | nan | nan |
| 18 | nan | nan | nan | nan | nan | nan | nan | 193 | nan | 4.7 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan |
| 19 | nan | 12.2 | 10.5 | 77 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 164.77 | nan | nan | nan | nan | nan |
| 20 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 15-20 | nan | nan |
| 21 | nan | nan | nan | nan | nan | nan | nan | nan | nan | 24.63 | 0.12 | nan | nan | nan | nan | nan | nan | 0.63 | 0.06 | 46.73 | 8.23 | nan |
| 22 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 40-45 | nan | 67.2 |
| 23 | 7.6 | nan | 44.2 | 31.6 | nan | 5.08 | 54.4 | 170 | nan | 4.43 | nan | nan | nan | nan | nan | 37 | 109 | 0.3 | nan | nan | nan | nan |
| 24 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan |
| 25 | nan | nan | nan | nan | nan | nan | nan | >160 | nan | 95.93 | 0.07 | nan | nan | nan | nan | nan | nan | 5.82 | 0.03 | 19.55 | 1.03 | nan |
| 26 | nan | 0.83 | 97.05 | 2.11 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 79.75 | nan | nan | nan | nan | nan |
| 27 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 46-55 | nan | nan |
| 28 | 1.1 | nan | nan | nan | nan | nan | nan | > 160 | nan | 15.25 | nan | nan | nan | -13.4 | 0.034 | 38.7 | nan | nan | nan | nan | nan | nan |
| 29 | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | nan | 54 | nan | nan |

**Figure 2: Screenshot of the Provided Data**

Data cleaning techniques are applied in order to make the data ready for usage.
Data was preprocessed taking the following actions:

- Drop columns with all NaN entries
- NaN filling
- String to float
- Data interpolation
- Data formatting
- Merge identical feedstock with values equal to their mean
- Shorten Species Names to their acronyms

The outcome is given as follows. After preprocessing, the data size is reduced to 106x32.

| Bio-diesel | (kg / hect | tic acid (C | ic acid (C | cid (C18: | eic acid (C | nic acid ( | ic acid (C | c Acid (C | n Stabilit | turated fa | turated fa | turated fa | in saturat | tane Num | lash Poin | Kinetic Viscosity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CN | 16800 | 15.1338 | 8.14286 | 38.7823 | 48.5 | 28.5 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 77 | 5.4 | 56.1977 | 142.085 | 24.63 |
| Tung | 375 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| AS | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| Peanuts | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| milkweed | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 4.9 |
| Whitewood | 5452.98 | 15.1338 | 8.14286 | 65.73 | 33.7172 | 3.9 | 0.2 | 5.2636 | 8.301 | 23.2394 | 86.05 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| Neem | 2721.5 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| DD | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 120 | 4 |
| BP | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 50.1 | 2.1 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| BS | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 30.8 | 3.76481 | 0.2 | 0.25 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 13.4305 |
| FT | 5452.98 | 21 | 16 | 52.4 | 26 | 2.68 | 0.2 | 5.2636 | 8.301 | 23.2394 | 52.4 | 31.2445 | 5.4 | 56.1977 | 142.085 | 22.33 |
| Bidwilli | 5452.98 | 42.39 | 14.37 | 40.5 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 56.76 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 142.085 | 23.03 |
| EM | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 120 | 4.5 |
| RO | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 8.301 | 23.2394 | 46.4182 | 31.6 | 5.4 | 56.1977 | 170 | 13.4305 |
| BLTO | 16000 | 15.1338 | 16.55 | 42.48 | 35.1 | 3.76481 | 0.2 | 5.2636 | 8.301 | 27.83 | 47.49 | 31.2445 | 11.64 | 56.1977 | 151 | 40.05 |
| CETBLME | 5452.98 | 13.4 | 16.5 | 40.4 | 26.2 | 0.4 | 0.2 | 0 | 13.27 | 30.7 | 41.8 | 27.5 | 5.4 | 58.3 | 169 | 4.57 |
| CS | 5452.98 | 15.1338 | 8.14286 | 38.7823 | 33.7172 | 3.76481 | 0.2 | 5.2636 | 1.7 | 23.2394 | 46.4182 | 31.2445 | 5.4 | 56.1977 | 160 | 4.365 |

**Figure 3: Screenshot of the Preprocessed Data**

Once the data has been cleansed and preprocessed, it is needed to make a choice on which criteria to include in the MCDA for ranking.

Eight criteria have been selected for the ranking, namely : 'Fuel Density', 'Higher heating value (HHV)', 'O2',' 'Oxidation Stability (hour)', 'Acid value (mg KOH/g)', 'Flash Point', 'Iodine value (g I/100g) (IV)', 'Kinetic Viscosity' with ideal values MIN, MAX, MIN, MAX, MIN, MIN, MIN, MIN, respectively.

Hence the data has been further narrowed down to a set of 106 alternatives by 8 criteria.

Data exploration was performed to get a better understanding of the data representativeness. Figure 3 illustrates a violin plot displaying the probability density of the data at different values smoothed by a kernel density estimator.

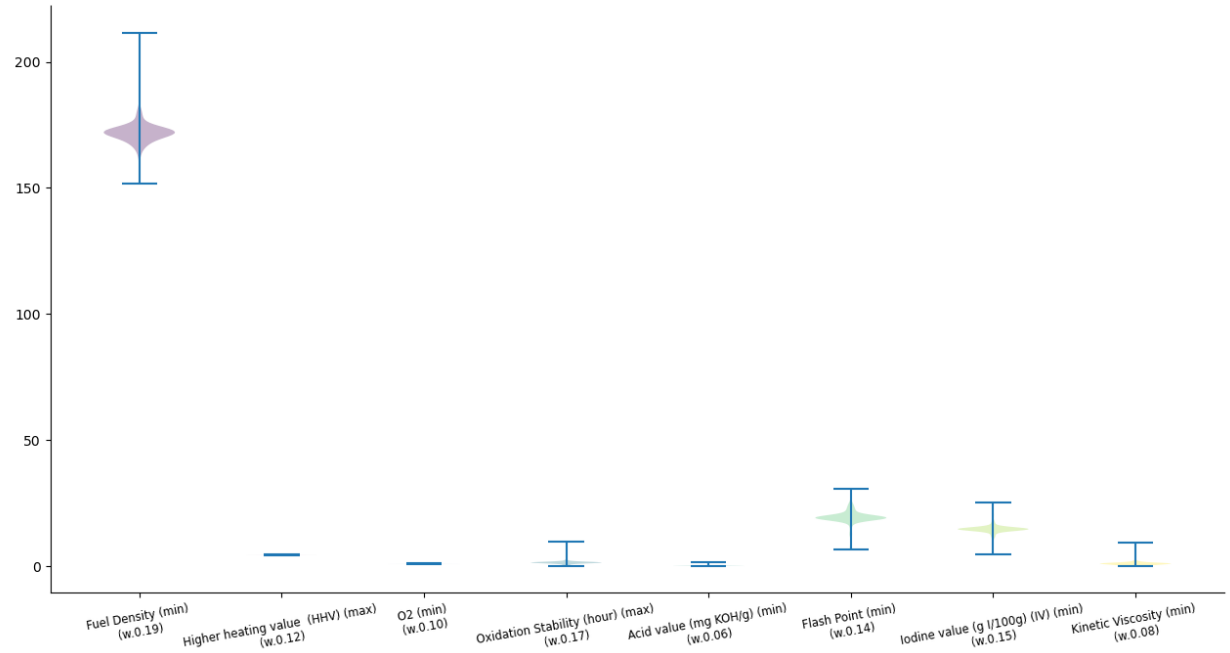



**Figure 4: Violin Plot of 106 alternatives by 8 criteria**

The chosen criteria belong to the following three categories: biodiesel properties, engine performance and engine emission. Weights have been assigned to the above criteria.

Three classes of weights are chosen, namely ‘engine’, ‘environment’ or ‘economy’. Each class gives priority to one specific aspect, in such a way the ranking systems are specified according to whether priority is given to engine, environment or economy.

The weight values are shown as follows.

```
weights_eng = [0.205, 0.045, 0.068, 0.159, 0.091, 0.182, 0.136, 0.114]
weights_env = [0.154, 0.173, 0.135, 0.115, 0.096, 0.077, 0.058, 0.192]
weights_eco = [0.192, 0.115, 0.096, 0.173, 0.058, 0.135, 0.154, 0.077]
```

**Figure 5: Screenshot showing the used classes of weights**

## 4. MCDA Deliverables

Tables 4, 5 and 6 show the heads (best 5) of the performed rankings sorted in ascending order by the average of the three ranking models. Tables 7, 8 and 9 show the tails (worst 5) of the performed rankings sorted in ascending order by the average of the three ranking models. Three csv files are generated with the script (namely decision_eng.csv, decisions_env.csv and decisions_eco.csv).

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| RETBLME | 6 | 2 | 3 | 1 |
| SFKOB | 2 | 1 | 9 | 2 |
| CETBLME | 7 | 3 | 4 | 3.5 |
| PME | 4 | 4 | 6 | 3.5 |
| RTBLME | 8 | 7 | 5 | 5 |

**Table 4: Screenshot of Rank Head with Engine Weights**

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| SFKOB | 2 | 1 | 9 | 2 |
| RETBLME | 6 | 3 | 3 | 2 |
| PME | 4 | 2 | 6 | 2 |
| CETBLME | 7 | 4 | 4 | 4 |
| RTBLME | 8 | 6 | 5 | 5 |

**Table 5: Screenshot of Rank Head with Environment Weights**

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| RETBLME | 6 | 2 | 3 | 1 |
| CETBLME | 7 | 3 | 4 | 2 |
| PME | 4 | 6 | 6 | 3 |
| SFKOB | 2 | 4 | 11 | 4 |
| KS | 8 | 5 | 8 | 5.5 |

**Table 6: Screenshot of Rank Head with Economy Weights**

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| PP | 104 | 99 | 102 | 102 |
| Bidwilli | 103 | 102 | 104 | 103.5 |
| Cordyline | 105 | 101 | 103 | 103.5 |
| SA | 101 | 104 | 106 | 105 |
| castor | 106 | 106 | 105 | 106 |

**Table 7: Screenshot of Rank Tail with Engine Weights**

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| Cordyline | 102 | 102 | 101 | 101.5 |
| PP | 104 | 103 | 102 | 103 |
| Bidwilli | 103 | 104 | 104 | 104 |
| castor | 106 | 105 | 105 | 105 |
| SA | 105 | 106 | 106 | 106 |

**Table 8: Screenshot of Rank Tail with Environment Weights**

| Bio-diesel | Rank WSM | Rank WSP | Rank TOPSIS | Avg Rank |
|---|---|---|---|---|
| PP | 103 | 97 | 102 | 102 |
| Bidwilli | 102 | 99 | 104 | 103 |
| SA | 100 | 101 | 106 | 104 |
| Cordyline | 105 | 100 | 103 | 105 |
| castor | 106 | 103 | 105 | 106 |

**Table 9: Screenshot of Rank Tail with Economy Weights**

## 5. DNN Deliverables

The outcome of the MCDA is used as an input in this phase. Figure 12, 10 and 11 illustrate an overview of MVR, DNN and MLP models used.

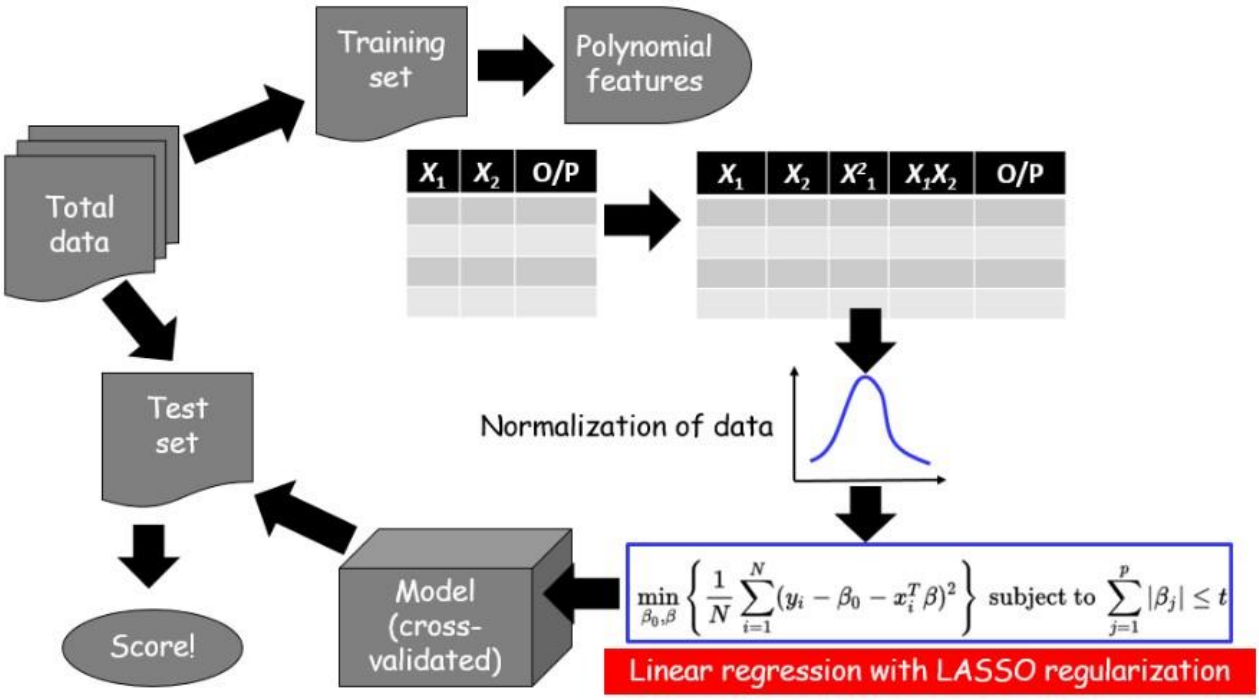



**Figure 9: MVR Methodology**

```
In [89]: model.summary()
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_32 (Dense)             (None, 50)                450
_________________________________________________________________
dropout_24 (Dropout)         (None, 50)                0
_________________________________________________________________
dense_33 (Dense)             (None, 50)                2550
_________________________________________________________________
dropout_25 (Dropout)         (None, 50)                0
_________________________________________________________________
dense_34 (Dense)             (None, 50)                2550
_________________________________________________________________
dropout_26 (Dropout)         (None, 50)                0
_________________________________________________________________
dense_35 (Dense)             (None, 1)                 51
=================================================================
Total params: 5,601
Trainable params: 5,601
Non-trainable params: 0
_________________________________________________________________
```

**Figure 10: DNN Model Summary**

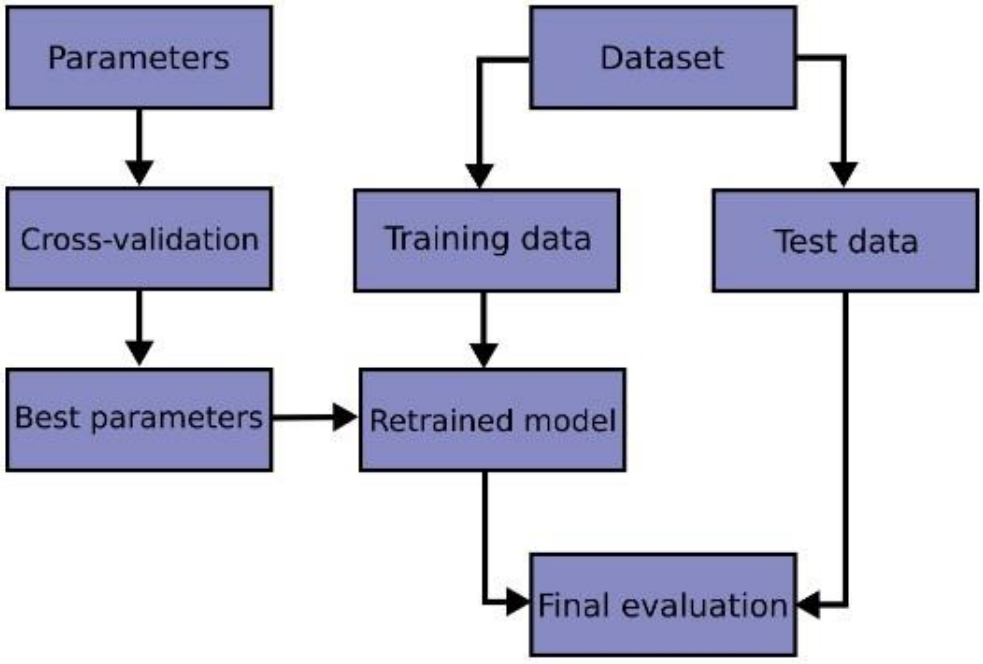



**Figure 11: MLP Methodology Overview**

A statistical overview of the data used for training is given in Table 10. Such table is also generated by the script under the filename describe_tf_train.csv.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Fuel Density** | 155 | 897.277 | 30.10763 | 790 | 890 | 897.277 | 897.277 | 1102 |
| **Higher heating value (HHV)** | 155 | 39.56118 | 0.231269 | 37.5 | 39.56118 | 39.56118 | 39.56118 | 40.12 |
| **O2** | 155 | 11.57091 | 0.242261 | 10.35 | 11.57091 | 11.57091 | 11.57091 | 14.19 |
| **Oxidation Stability (hour)** | 155 | 8.301 | 5.063658 | 0.4 | 8.301 | 8.301 | 8.301 | 56.9 |
| **Acid value (mg KOH/g)** | 155 | 3.812162 | 2.701945 | 0.16 | 3.812162 | 3.812162 | 3.812162 | 25.8 |
| **Flash Point** | 155 | 142.0848 | 21.61441 | 48 | 142.0848 | 142.0848 | 142.0848 | 228 |
| **Iodine value (g I/100g) (IV)** | 155 | 95.33895 | 13.03508 | 29.86 | 95.33895 | 95.33895 | 95.33895 | 164.77 |
| **Kinetic Viscosity** | 155 | 13.43053 | 12.29982 | 2.5 | 6.395 | 13.43053 | 13.43053 | 120.17 |

**Table 10: Statistical Overview of Training Dataset**

After implementation and training, the performance of the three above-mentioned models is assessed through prediction on both training and test set. Bar charts with predictions from MVR, DNN and MLP are reported in Figure 12, 13 and 14, respectively. The indices on the x-axes refer to the index of the biodiesel, which is unchanged from the in the initial provided data. The y axes indicate the rankings.

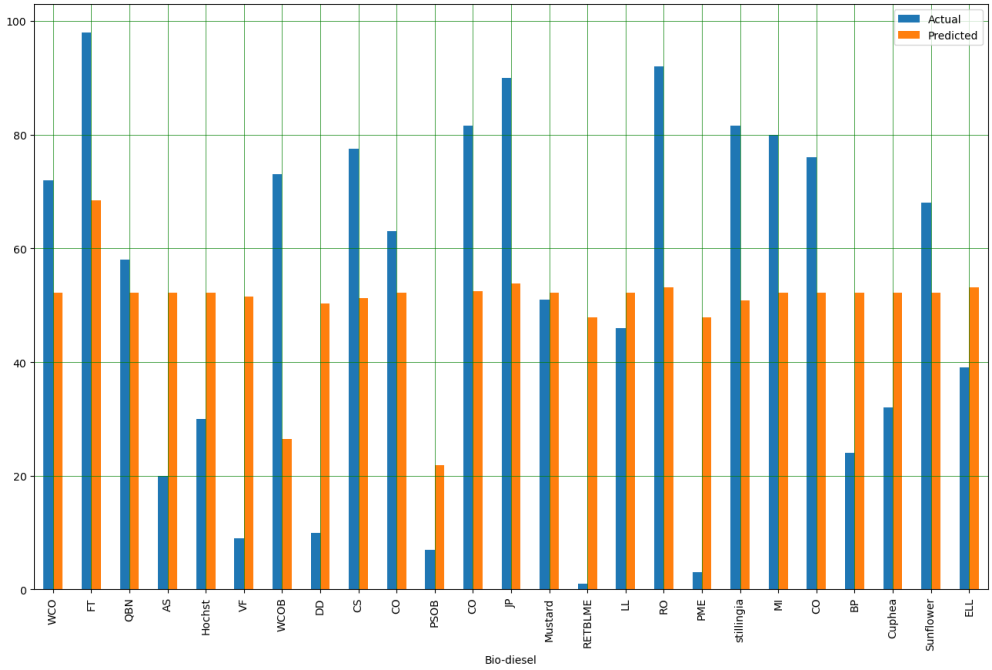



**Figure 12: MVR Predictions on Test Dataset**

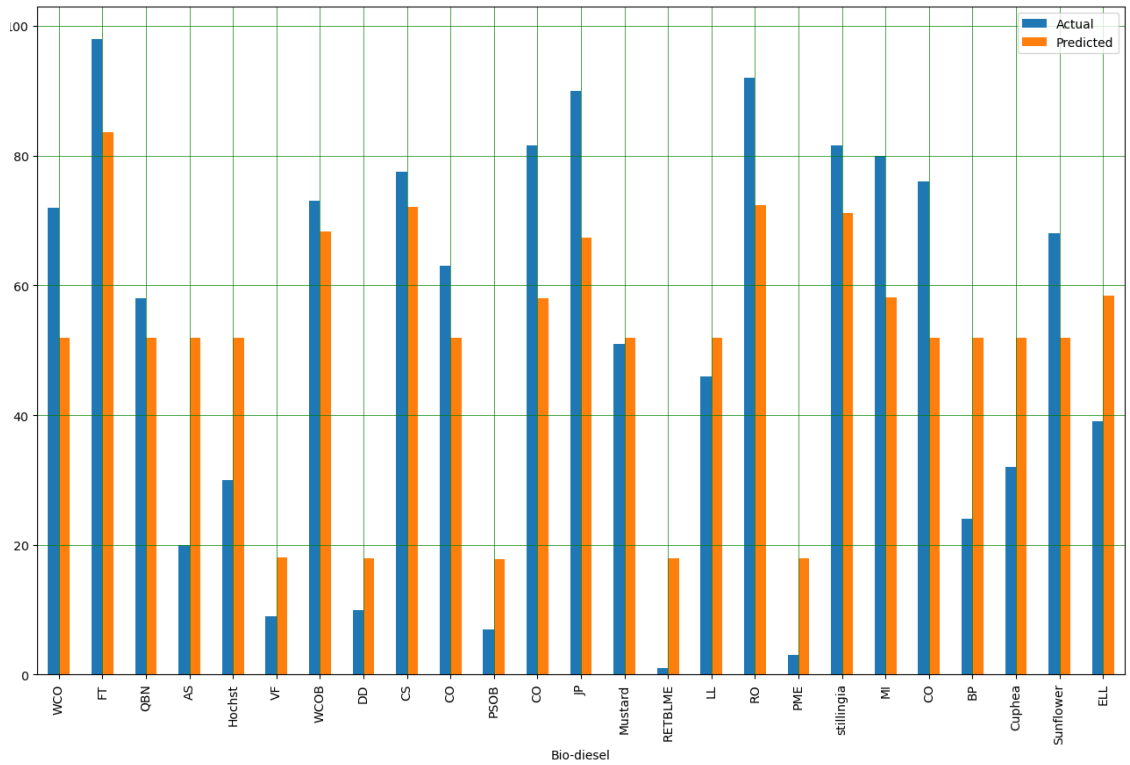



**Figure 13: DNN Predictions on Test Dataset**

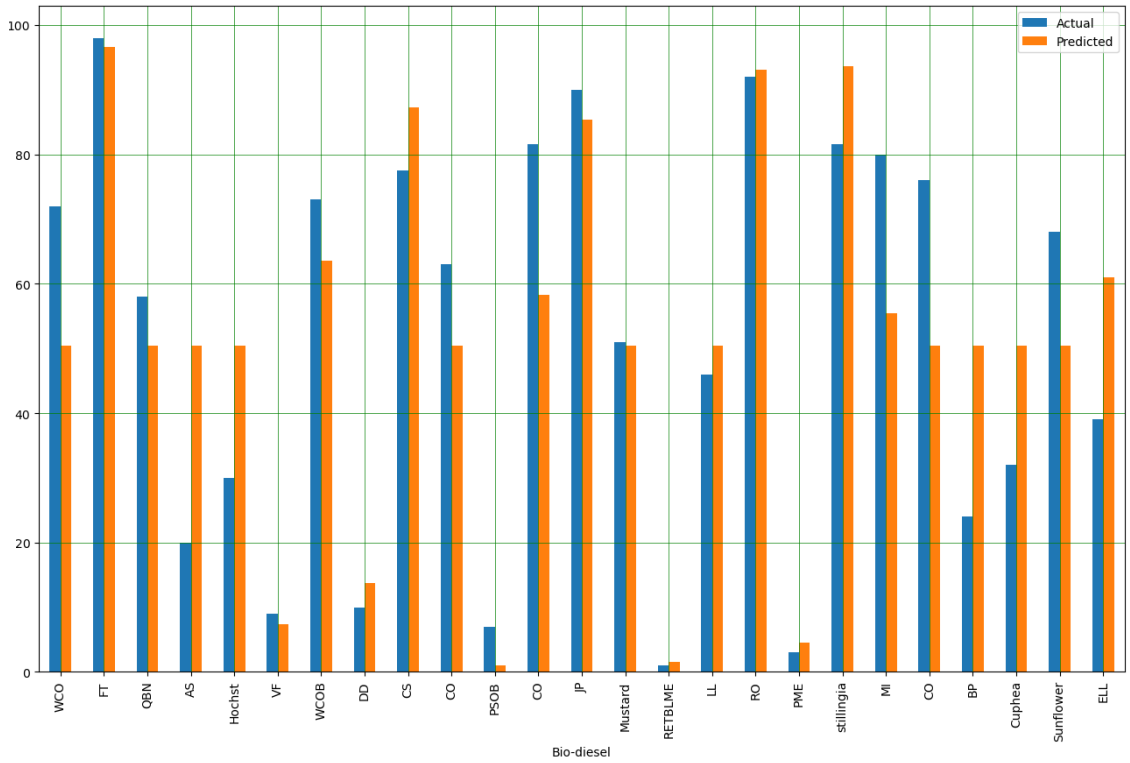



**Figure 14: MLP Predictions on Test Dataset**

The predictions displayed in Figures 12, 13 and 14 are generated with filenames predictions_mvr.csv, predicitions_dnn.csv and predictions_mlp.csv, respectively. The models for DNN and MLP are saved as dnn_model.h5 and mlp_model.pkl, respectively.

The accuracy of the used methods is evaluated through R2 scores, applied on both train and test datasets, and Mean Squared Error (MSE) and Mean Absolute Error (MAE), applied on test datasets. The obtained results are provided in Table 5 and they are also separately generated in the file accuracy.csv. From the below Table, it is clear that the best method is MLP, as it achieves the highest R2 Scores with lowest MSE and MAE.

| Method | MVR | DNN | MLP |
|---|---|---|---|
| R2 Score Train | 0.240879 | 0.696474 | 0.87189 |
| R2 Score Test | 0.152453 | 0.67604 | 0.736951 |
| MSE Test | 863.1243 | 299.2435 | 267.8833 |
| MAE Test | 26.6856 | 15.06823 | 13.15108 |

**Table 11: Accuracy overview for MVR, DNN and MLP**

## 6. Conclusions

The average amount of missing data in the entire dataset is 87.6%. After reducing the size of the dataset by filtering off the properties will all missing values, the remaining dataset, which has been used for this analysis, contained 73.3% of missing values. It is obvious that data quality had a great impact on the final outcome of the project. Before starting the project we were aware that data quality was not exhaustive enough for obtaining satisfactory models for future predictions. This project is conceived as PoC of the potential that this technology has in this industry. Through the provided results, it has been demonstrated that a Neural Network can achieve accuracy of 87.2% on training set and 73.7% on test set.
It is expected that better data quality would lead to more satisfactory results, in the sense that a generic model can be built for forecasting ranks of new biodiesels based on their characteristics.