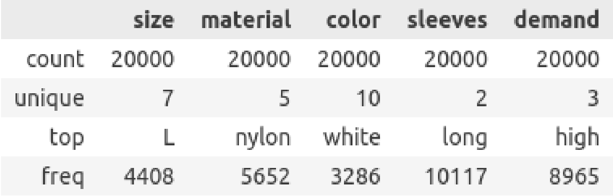
# Raport 4 - Jakub Ner

## 1. Data Exploration

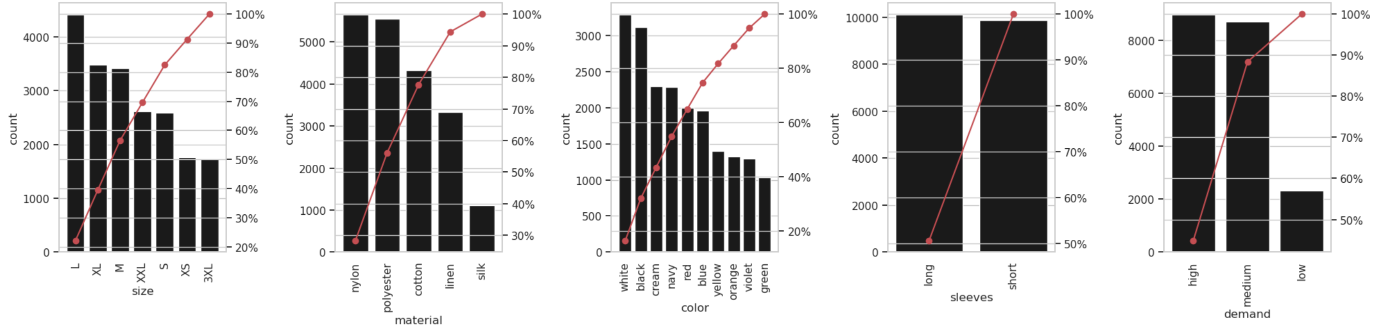
I started for inspecting the data:

- There are 5 columns and 20 000 rows.
- There are no NaN values in the dataset.
- All values are categorical



| | size | material | color | sleeves | demand |
|---|---|---|---|---|---|
| count | 20000 | 20000 | 20000 | 20000 | 20000 |
| unique | 7 | 5 | 10 | 2 | 3 |
| top | L | nylon | white | long | high |
| freq | 4408 | 5652 | 3286 | 10117 | 8965 |

In order to examine data characteristics I ploted pareto charts for each feature



The key observation is that target values lack balance, with "low" values representing only 10% of the dataset.

## 2. Data Preprocessing

### 2.1. One-hot encoding

I used one-hot encoding to convert categorical features to numerical. I dropped 1 level of each feature to avoid redundancy and multicollinearity.

> One-hot (resp. one-cold) encoding creates co-linearity if all the features are used. Simply because the following relation always holds: ∑ifi=1, So dropping one feature destroys the colinearity and it can have better results, since many models (esp. linear models) get confused with colinearities in the features. source: https://datascience.stackexchange.com/questions/96526/two-questions-about-one-hot-encoding-drop-first-and-features-with-thousands-of

### 2.2. Feature standardization

I used StandardScaler to standardize the columns according to the forumla `z = (x - u) / s` where `u` is the mean of the training samples and `s` is the standard deviation of the training samples. This operation assumes that the data is normally distributed within each feature and scales them such that the distribution is now centred around 0, with a standard deviation of 1.

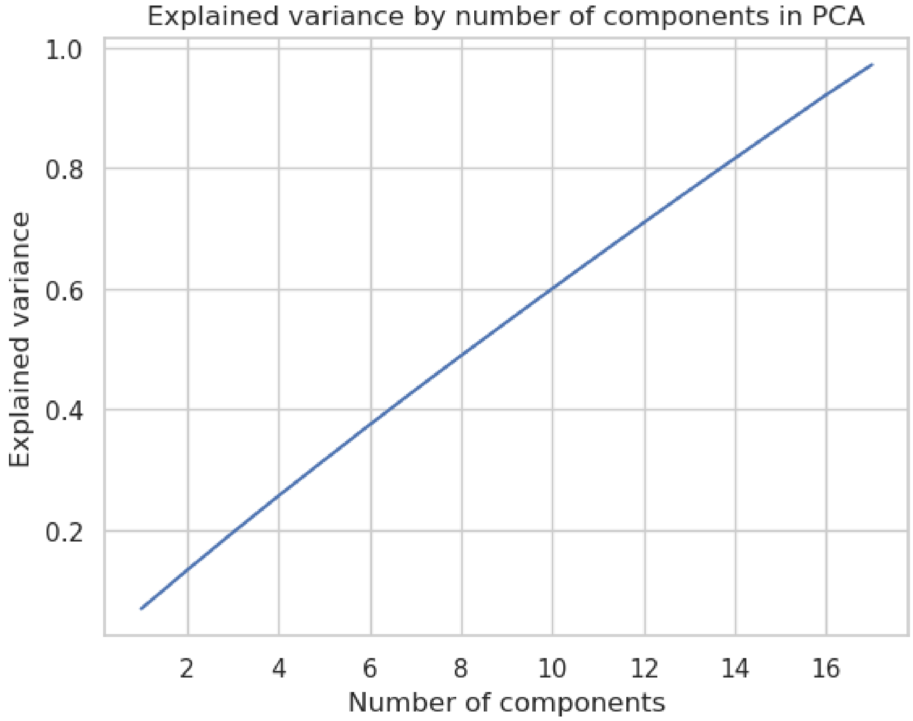
### 2.3. Samples normalization

I scaled each sample by dividing its values by euclidean norm (l2) of the sample. This operation ensures that the samples are on the same scale.

*Note: One hot encoded categorical values does not contains diversed weights (the values are 0 or 1). Moreover, it refers to all features in the dataset. Therefore those are not important for the model itself, but the PCA requires all predictors to be on the same scale.*

### 2.4. Dimensionality reduction - PCA

To reduce number of features I used Principle Component Analysis to explain 95% of variance. This lead to dimensionality reduction from 20 to 17 features.



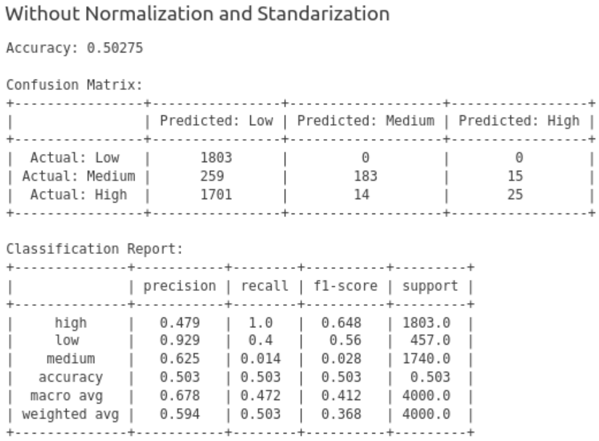
## 3. Classification

**3.1. Gaussian Naive Bayes**

Gaussian Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with naive independence assumptions between the features.

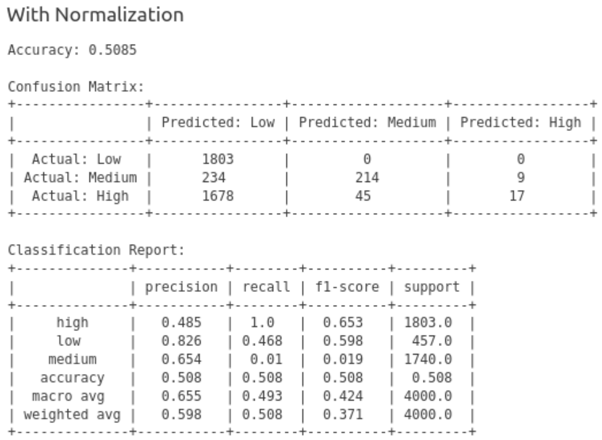
#### 3.1.1. Impact of normalization and standarization

I compared the results of the model with and without normalization and standarization. For this benchmark I didn not utilize PCA. Standarization and normalization, when applied separetely, did not have a significant impact on the model performance. However, when applied together and in the correct order - first standarization, then normalization, the model accuracy improved by 11 percentage points. After standardization, all features have the same variance, and normalization further adjusts the scale without altering the standardized distribution significantly.



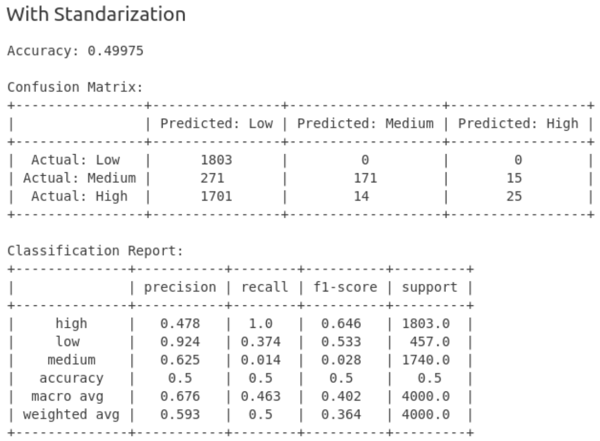
Without Normalization and Standarization

```
Accuracy: 0.50275

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1803      |         0         |        0        |
| Actual: Medium |      259       |        183        |        15       |
|  Actual: High  |      1701      |        14         |        25       |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.479   |  1.0   |  0.648   | 1803.0  |
|     low      |   0.929   |  0.4   |   0.56   |  457.0  |
|    medium    |   0.625   | 0.014  |  0.028   | 1740.0  |
|   accuracy   |   0.503   | 0.503  |  0.503   |  0.503  |
|  macro avg   |   0.678   | 0.472  |  0.412   | 4000.0  |
| weighted avg |   0.594   | 0.503  |  0.368   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



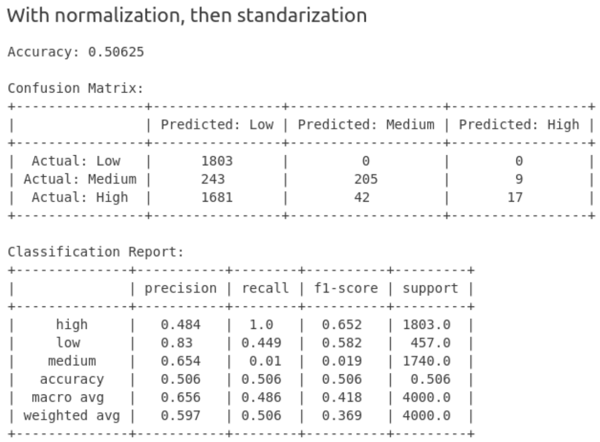
With Normalization

```
Accuracy: 0.5085

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1803      |         0         |        0        |
| Actual: Medium|      234       |        214        |        9        |
|  Actual: High |      1678      |        45         |        17       |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.485   |  1.0   |  0.653   | 1803.0  |
|     low     |   0.826   | 0.468  |  0.598   |  457.0  |
|    medium   |   0.654   |  0.01  |  0.019   | 1740.0  |
|   accuracy  |   0.508   | 0.508  |  0.508   |  0.508  |
|  macro avg  |   0.655   | 0.493  |  0.424   | 4000.0  |
| weighted avg|   0.598   | 0.508  |  0.371   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```



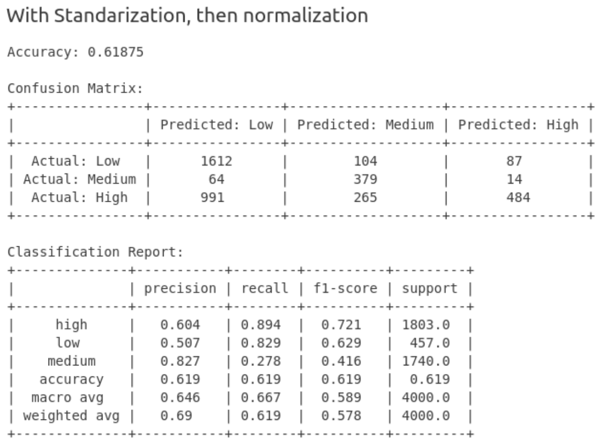
With Standarization

```
Accuracy: 0.49975

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1803      |         0         |        0        |
| Actual: Medium|      271       |        171        |        15       |
|  Actual: High |      1701      |        14         |        25       |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.478   |  1.0   |  0.646   | 1803.0  |
|     low     |   0.924   | 0.374  |  0.533   |  457.0  |
|    medium   |   0.625   | 0.014  |  0.028   | 1740.0  |
|   accuracy  |    0.5    |  0.5   |   0.5    |   0.5   |
|  macro avg  |   0.676   | 0.463  |  0.402   | 4000.0  |
| weighted avg|   0.593   |  0.5   |  0.364   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```



With normalization, then standarization

```
Accuracy: 0.50625

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1803      |         0         |        0        |
| Actual: Medium|      243       |        205        |        9        |
|  Actual: High |      1681      |        42         |        17       |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.484   |  1.0   |  0.652   | 1803.0  |
|     low     |   0.83    | 0.449  |  0.582   |  457.0  |
|    medium   |   0.654   |  0.01  |  0.019   | 1740.0  |
|   accuracy  |   0.506   | 0.506  |  0.506   |  0.506  |
|  macro avg  |   0.656   | 0.486  |  0.418   | 4000.0  |
| weighted avg|   0.597   | 0.506  |  0.369   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```

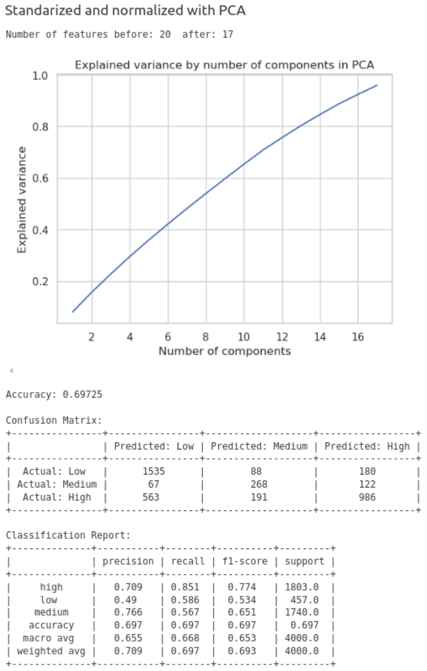


With Standarization, then normalization

```
Accuracy: 0.61875

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1612      |        104        |        87       |
| Actual: Medium|       64       |        379        |        14       |
|  Actual: High |      991       |        265        |       484       |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.604   | 0.894  |  0.721   | 1803.0  |
|     low     |   0.507   | 0.829  |  0.629   |  457.0  |
|    medium   |   0.827   | 0.278  |  0.416   | 1740.0  |
|   accuracy  |   0.619   | 0.619  |  0.619   |  0.619  |
|  macro avg  |   0.646   | 0.667  |  0.589   | 4000.0  |
| weighted avg|    0.69   | 0.619  |  0.578   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```

### 3.1.2 Impact of PCA

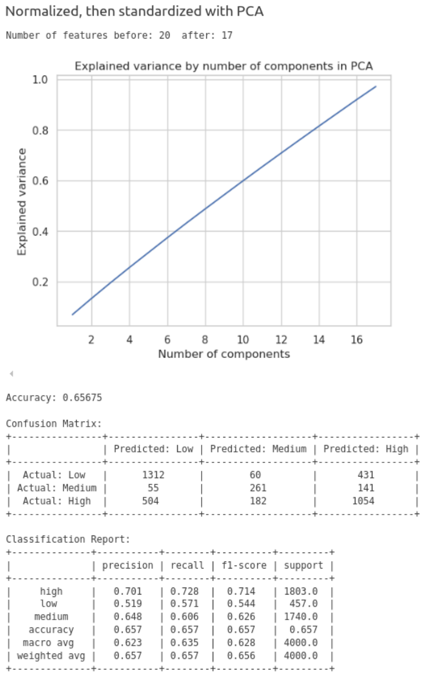
Normalization and standarization make the result worse, because those cause that less features are dropped during the PCA. It impacts the model performance.



Standarized and normalized with PCA

```
Number of features before: 20  after: 17
```

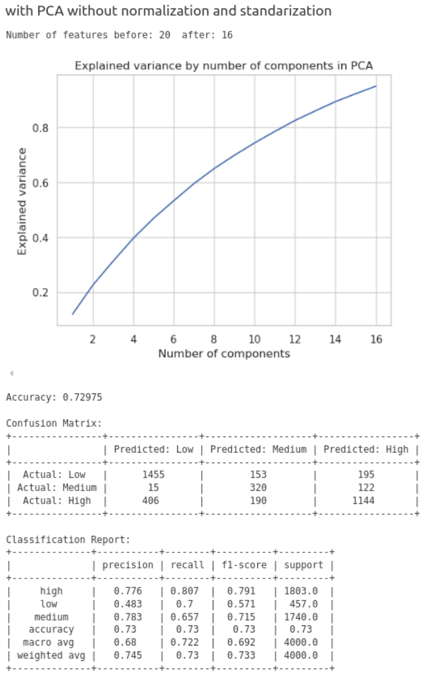
```
Accuracy: 0.69725

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1535      |        88         |       180       |
| Actual: Medium|       67       |        268        |       122       |
|  Actual: High |      563       |        191        |       986       |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.709   | 0.851  |  0.774   | 1803.0  |
|     low     |   0.49    | 0.586  |  0.534   |  457.0  |
|    medium   |   0.766   | 0.567  |  0.651   | 1740.0  |
|   accuracy  |   0.697   | 0.697  |  0.697   |  0.697  |
|  macro avg  |   0.655   | 0.668  |  0.653   | 4000.0  |
| weighted avg|   0.709   | 0.697  |  0.693   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```



Normalized, then standardized with PCA

```
Number of features before: 20  after: 17
```

```
Accuracy: 0.65675

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1312      |        60         |       431       |
| Actual: Medium|       55       |        261        |       141       |
|  Actual: High |      504       |        182        |       1054      |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.701   | 0.728  |  0.714   | 1803.0  |
|     low     |   0.519   | 0.571  |  0.544   |  457.0  |
|    medium   |   0.648   | 0.606  |  0.626   | 1740.0  |
|   accuracy  |   0.657   | 0.657  |  0.657   |  0.657  |
|  macro avg  |   0.623   | 0.635  |  0.628   | 4000.0  |
| weighted avg|   0.657   | 0.657  |  0.656   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```



with PCA without normalization and standarization

```
Number of features before: 20  after: 16
```

```
Accuracy: 0.72975

Confusion Matrix:
+---------------+----------------+-------------------+-----------------+
|               | Predicted: Low | Predicted: Medium | Predicted: High |
+---------------+----------------+-------------------+-----------------+
|  Actual: Low  |      1455      |        153        |       195       |
| Actual: Medium|       15       |        320        |       122       |
|  Actual: High |      406       |        190        |       1144      |
+---------------+----------------+-------------------+-----------------+

Classification Report:
+-------------+-----------+--------+----------+---------+
|             | precision | recall | f1-score | support |
+-------------+-----------+--------+----------+---------+
|     high    |   0.776   | 0.807  |  0.791   | 1803.0  |
|     low     |   0.483   |  0.7   |  0.571   |  457.0  |
|    medium   |   0.783   | 0.657  |  0.715   | 1740.0  |
|   accuracy  |   0.73    |  0.73  |   0.73   |  0.73   |
|  macro avg  |   0.68    | 0.722  |  0.692   | 4000.0  |
| weighted avg|   0.745   |  0.73  |  0.733   | 4000.0  |
+-------------+-----------+--------+----------+---------+
```

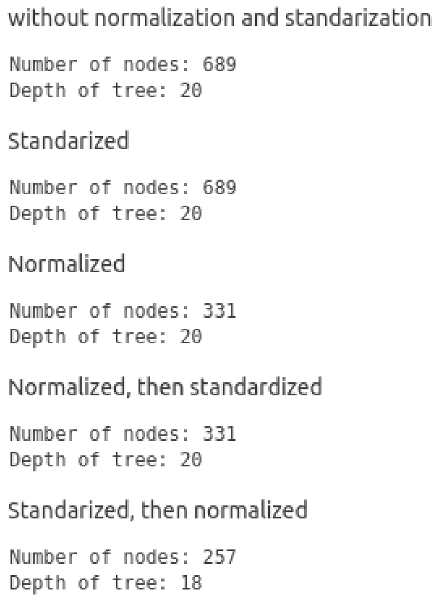
### 3.2. Decision Tree

#### 3.2.1. Impact of normalization and standarization

Normalization and standarization does not impact the model performance in terms of accuracy .

> Decision trees do not require feature scaling or normalization, as they are invariant to monotonic transformations. They can also easily handle missing values and outliers, making them suitable for raw and noisy data.

Nevertheless, it can decrease number of nodes in the tree by over a half and decrease 2 levels of depth.



without normalization and standarization

```
Number of nodes: 689
Depth of tree: 20
```

Standarized

```
Number of nodes: 689
Depth of tree: 20
```

Normalized

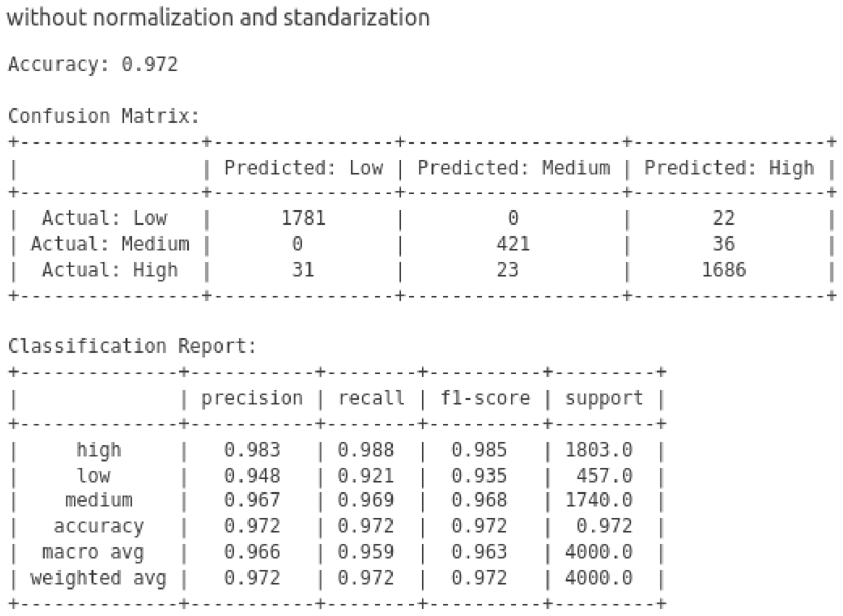
```
Number of nodes: 331
Depth of tree: 20
```

Normalized, then standardized

```
Number of nodes: 331
Depth of tree: 20
```

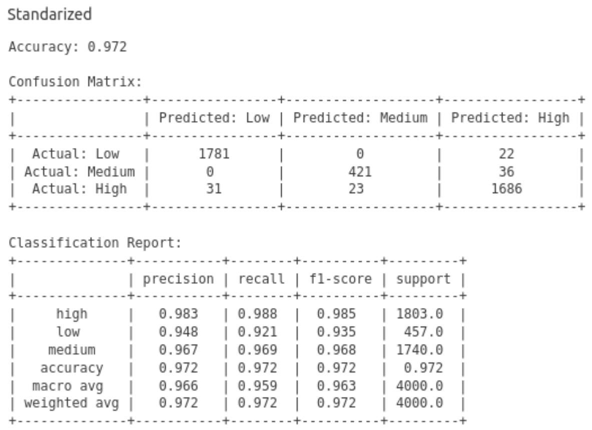
Standarized, then normalized

```
Number of nodes: 257
Depth of tree: 18
```



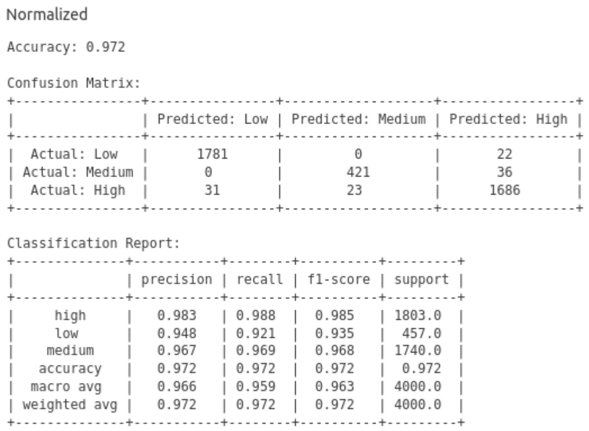
without normalization and standarization

```
Accuracy: 0.972

Confusion Matrix:
+----------------+---------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+---------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+---------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



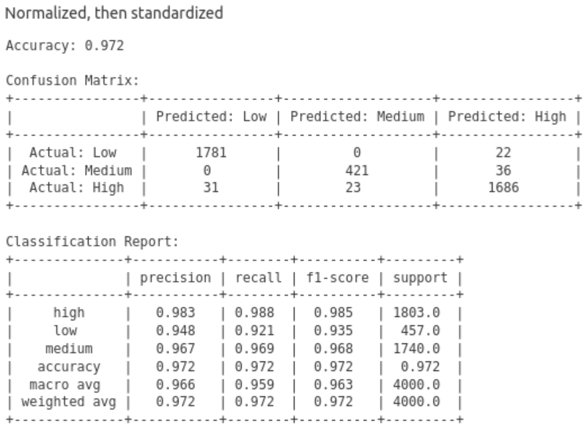
Standarized

```
Accuracy: 0.972

Confusion Matrix:
+----------------+---------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+---------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+---------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



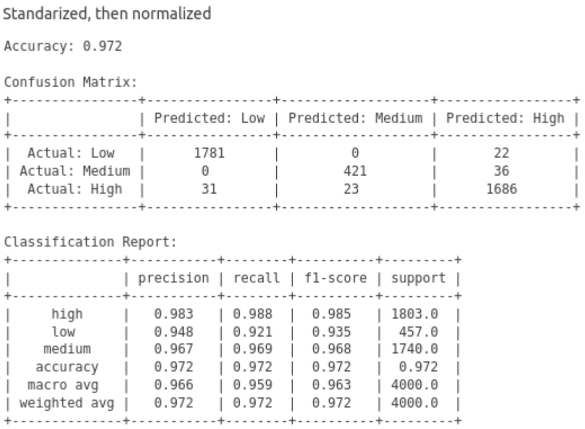
Normalized

```
Accuracy: 0.972

Confusion Matrix:
+----------------+---------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+---------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+---------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Normalized, then standardized

```
Accuracy: 0.972

Confusion Matrix:
+----------------+-----------------+-------------------+-----------------+
|                | Predicted: Low  | Predicted: Medium | Predicted: High |
+----------------+-----------------+-------------------+-----------------+
|  Actual: Low   |      1781       |         0         |       22        |
| Actual: Medium |        0        |        421        |       36        |
|  Actual: High  |       31        |        23         |      1686       |
+----------------+-----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```

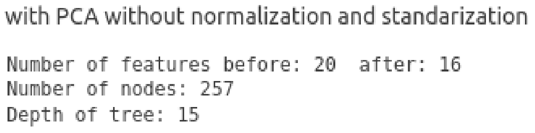


Standarized, then normalized

```
Accuracy: 0.972

Confusion Matrix:
+----------------+-----------------+-------------------+-----------------+
|                | Predicted: Low  | Predicted: Medium | Predicted: High |
+----------------+-----------------+-------------------+-----------------+
|  Actual: Low   |      1781       |         0         |       22        |
| Actual: Medium |        0        |        421        |       36        |
|  Actual: High  |       31        |        23         |      1686       |
+----------------+-----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```

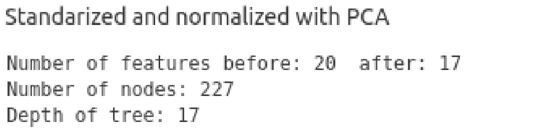
#### 3.2.2 Impact of PCA

Similarly, PCA modifies only the structure of the model, what hypotetically for a bigger model may decrease chance of overfitting and improve inference time.



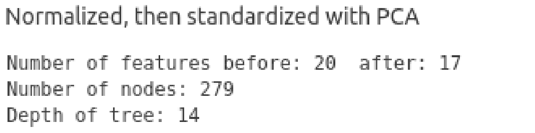
with PCA without normalization and standarization

```
Number of features before: 20  after: 16
Number of nodes: 257
Depth of tree: 15
```



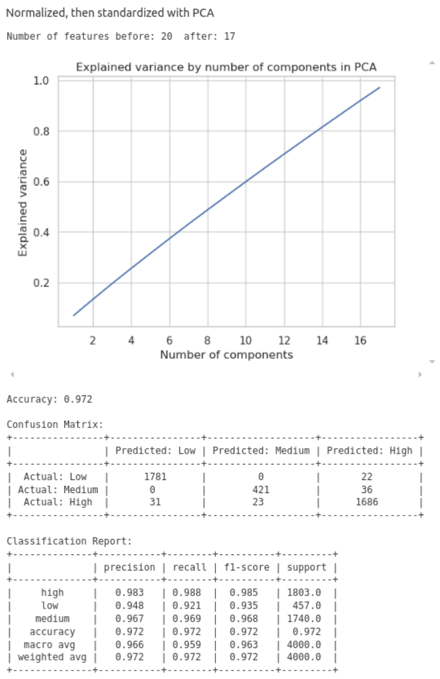
Standarized and normalized with PCA

```
Number of features before: 20  after: 17
Number of nodes: 227
Depth of tree: 17
```



Normalized, then standardized with PCA

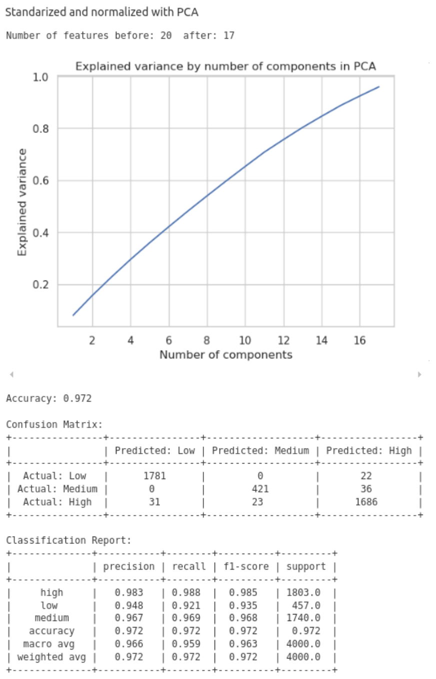
```
Number of features before: 20  after: 17
Number of nodes: 279
Depth of tree: 14
```



Normalized, then standardized with PCA

```
Number of features before: 20  after: 17
```

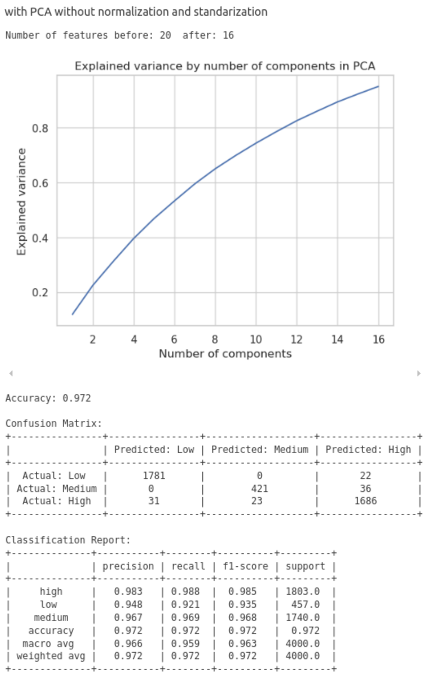
```
Accuracy: 0.972

Confusion Matrix:
+----------------+-----------------+-------------------+-----------------+
|                | Predicted: Low  | Predicted: Medium | Predicted: High |
+----------------+-----------------+-------------------+-----------------+
|  Actual: Low   |      1781       |         0         |       22        |
| Actual: Medium |        0        |        421        |       36        |
|  Actual: High  |       31        |        23         |      1686       |
+----------------+-----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Standarized and normalized with PCA

```
Number of features before: 20  after: 17
```

```
Accuracy: 0.972

Confusion Matrix:
+----------------+-----------------+-------------------+-----------------+
|                | Predicted: Low  | Predicted: Medium | Predicted: High |
+----------------+-----------------+-------------------+-----------------+
|  Actual: Low   |      1781       |         0         |       22        |
| Actual: Medium |        0        |        421        |       36        |
|  Actual: High  |       31        |        23         |      1686       |
+----------------+-----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



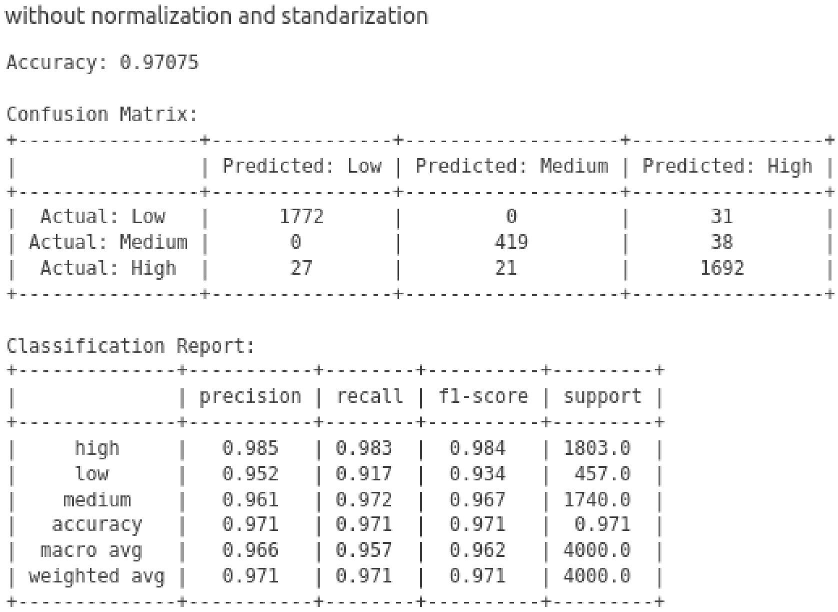
with PCA without normalization and standarization

```
Number of features before: 20  after: 16
```

```
Accuracy: 0.972

Confusion Matrix:
+----------------+-----------------+-------------------+-----------------+
|                | Predicted: Low  | Predicted: Medium | Predicted: High |
+----------------+-----------------+-------------------+-----------------+
|  Actual: Low   |      1781       |         0         |       22        |
| Actual: Medium |        0        |        421        |       36        |
|  Actual: High  |       31        |        23         |      1686       |
+----------------+-----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```

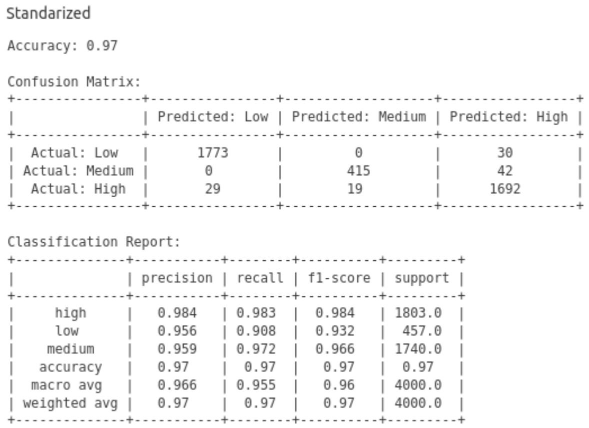
### 3.3. SVM

#### 3.3.1. Impact of normalization and standarization



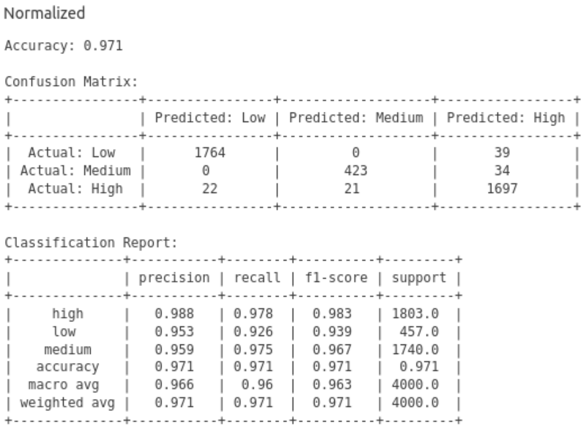
without normalization and standarization

```
Accuracy: 0.97075

Confusion Matrix:
+----------------+----------------+-------------------+------------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+------------------+
|  Actual: Low   |      1772      |         0         |        31        |
| Actual: Medium |       0        |        419        |        38        |
|  Actual: High  |       27       |         21        |       1692       |
+----------------+----------------+-------------------+------------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.985   | 0.983  |  0.984   | 1803.0  |
|     low      |   0.952   | 0.917  |  0.934   |  457.0  |
|    medium    |   0.961   | 0.972  |  0.967   | 1740.0  |
|   accuracy   |   0.971   | 0.971  |  0.971   |  0.971  |
|  macro avg   |   0.966   | 0.957  |  0.962   | 4000.0  |
| weighted avg |   0.971   | 0.971  |  0.971   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



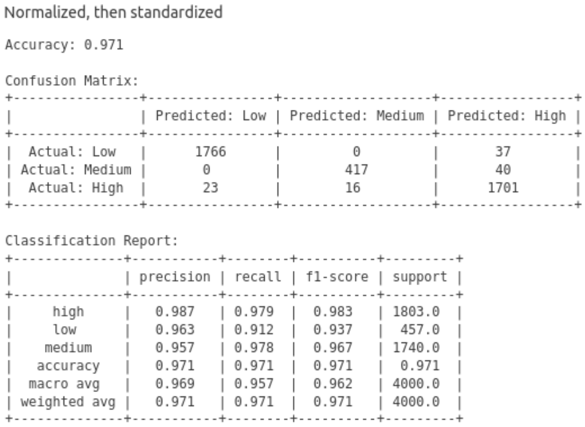
Standarized

```
Accuracy: 0.97

Confusion Matrix:
+----------------+----------------+-------------------+------------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+------------------+
|  Actual: Low   |      1773      |         0         |        30        |
| Actual: Medium |       0        |        415        |        42        |
|  Actual: High  |       29       |         19        |       1692       |
+----------------+----------------+-------------------+------------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.984   | 0.983  |  0.984   | 1803.0  |
|     low      |   0.956   | 0.908  |  0.932   |  457.0  |
|    medium    |   0.959   | 0.972  |  0.966   | 1740.0  |
|   accuracy   |   0.97    |  0.97  |   0.97   |  0.97   |
|  macro avg   |   0.966   | 0.955  |   0.96   | 4000.0  |
| weighted avg |   0.97    |  0.97  |   0.97   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



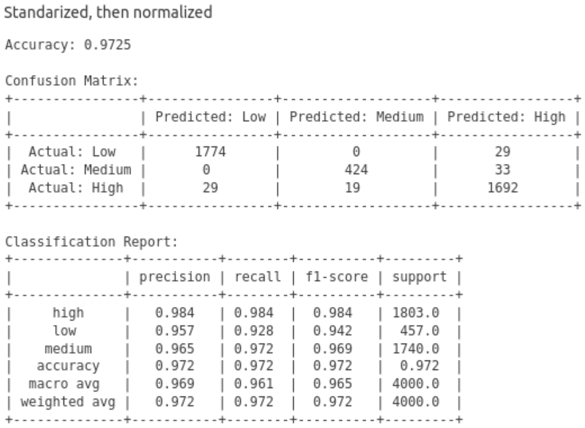
Normalized

```
Accuracy: 0.971

Confusion Matrix:
+----------------+----------------+-------------------+------------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+------------------+
|  Actual: Low   |      1764      |         0         |        39        |
| Actual: Medium |       0        |        423        |        34        |
|  Actual: High  |       22       |         21        |       1697       |
+----------------+----------------+-------------------+------------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.988   | 0.978  |  0.983   | 1803.0  |
|     low      |   0.953   | 0.926  |  0.939   |  457.0  |
|    medium    |   0.959   | 0.975  |  0.967   | 1740.0  |
|   accuracy   |   0.971   | 0.971  |  0.971   |  0.971  |
|  macro avg   |   0.966   |  0.96  |  0.963   | 4000.0  |
| weighted avg |   0.971   | 0.971  |  0.971   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Normalized, then standardized

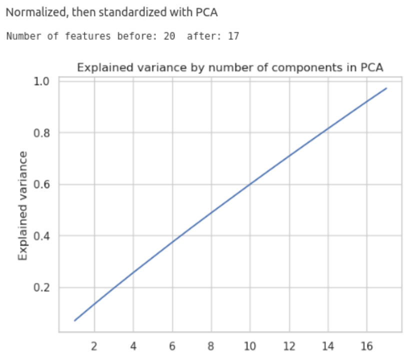
```
Accuracy: 0.971

Confusion Matrix:
+----------------+----------------+-------------------+------------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+------------------+
|  Actual: Low   |      1766      |         0         |        37        |
| Actual: Medium |       0        |        417        |        40        |
|  Actual: High  |       23       |         16        |       1701       |
+----------------+----------------+-------------------+------------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.987   | 0.979  |  0.983   | 1803.0  |
|     low      |   0.963   | 0.912  |  0.937   |  457.0  |
|    medium    |   0.957   | 0.978  |  0.967   | 1740.0  |
|   accuracy   |   0.971   | 0.971  |  0.971   |  0.971  |
|  macro avg   |   0.969   | 0.957  |  0.962   | 4000.0  |
| weighted avg |   0.971   | 0.971  |  0.971   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Standarized, then normalized

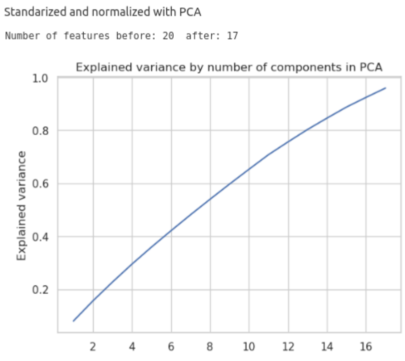
```
Accuracy: 0.9725

Confusion Matrix:
+----------------+----------------+-------------------+------------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+------------------+
|  Actual: Low   |      1774      |         0         |        29        |
| Actual: Medium |       0        |        424        |        33        |
|  Actual: High  |       29       |         19        |       1692       |
+----------------+----------------+-------------------+------------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.984   | 0.984  |  0.984   | 1803.0  |
|     low      |   0.957   | 0.928  |  0.942   |  457.0  |
|    medium    |   0.965   | 0.972  |  0.969   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.969   | 0.961  |  0.965   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```

### 3.3.2. Impact of PCA



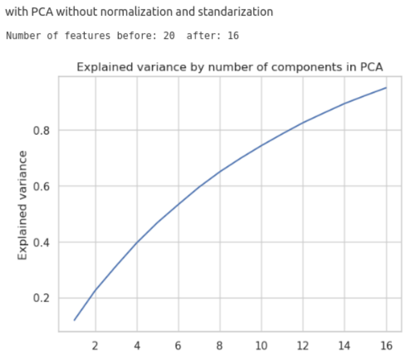
Normalized, then standardized with PCA

```
Number of features before: 20  after: 17
```



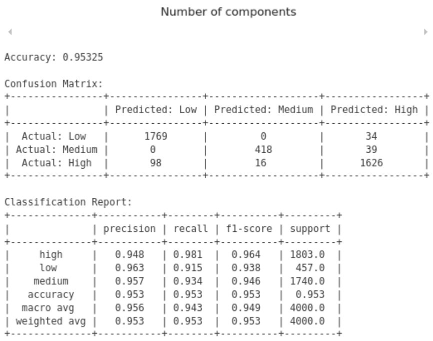
Standarized and normalized with PCA

```
Number of features before: 20  after: 17
```



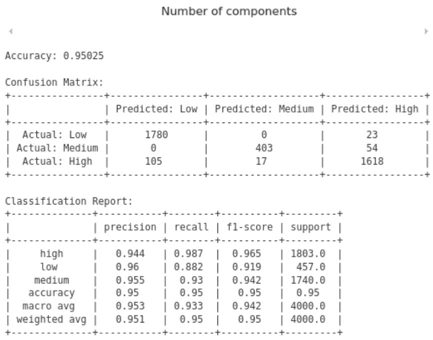
with PCA without normalization and standarization

```
Number of features before: 20  after: 16
```



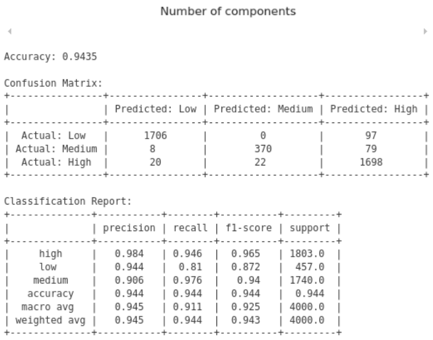
Number of components

```
Accuracy: 0.95325

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1769      |         0         |        34       |
| Actual: Medium |       0        |        418        |        39       |
|  Actual: High  |       98       |         16        |       1626      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.948   | 0.981  |  0.964   | 1803.0  |
|     low      |   0.963   | 0.915  |  0.938   |  457.0  |
|    medium    |   0.957   | 0.934  |  0.946   | 1740.0  |
|   accuracy   |   0.953   | 0.953  |  0.953   |  0.953  |
|  macro avg   |   0.956   | 0.943  |  0.949   | 4000.0  |
| weighted avg |   0.953   | 0.953  |  0.953   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Number of components

```
Accuracy: 0.95025

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1780      |         0         |        23       |
| Actual: Medium |       0        |        403        |        54       |
|  Actual: High  |      105       |         17        |       1618      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.944   | 0.987  |  0.965   | 1803.0  |
|     low      |   0.96    | 0.882  |  0.919   |  457.0  |
|    medium    |   0.955   |  0.93  |  0.942   | 1740.0  |
|   accuracy   |   0.95    |  0.95  |   0.95   |  0.95   |
|  macro avg   |   0.953   | 0.933  |  0.942   | 4000.0  |
| weighted avg |   0.951   |  0.95  |   0.95   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```



Number of components

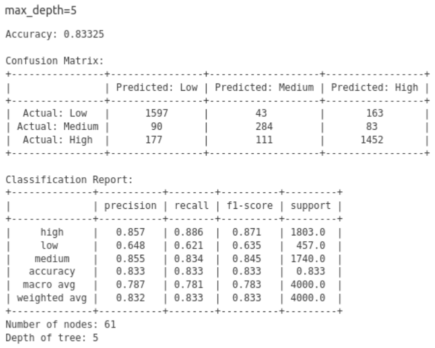
```
Accuracy: 0.9435

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1706      |         0         |        97       |
| Actual: Medium |       8        |        370        |        79       |
|  Actual: High  |       20       |         22        |       1698      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.984   | 0.946  |  0.965   | 1803.0  |
|     low      |   0.944   |  0.81  |  0.872   |  457.0  |
|    medium    |   0.906   | 0.976  |   0.94   | 1740.0  |
|   accuracy   |   0.944   | 0.944  |  0.944   |  0.944  |
|  macro avg   |   0.945   | 0.911  |  0.925   | 4000.0  |
| weighted avg |   0.945   | 0.944  |  0.943   | 4000.0  |
+--------------+-----------+--------+----------+---------+
```

#### Hyperparameters benchmarks

#### Decision Tree

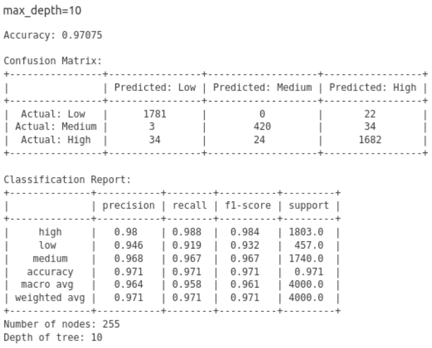
I tested the impact of the following hyperparameters on the normalized, then stadardized with PCA model:

1. max_depth: 5 vs 10 vs 15 The best results are for max_depth=15, but there is a change of overfitting.



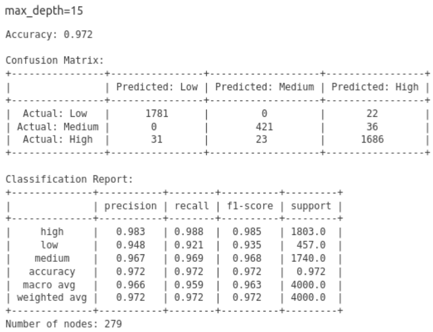
max_depth=5

```
Accuracy: 0.83325

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1597      |         43        |       163       |
| Actual: Medium |       90       |        284        |        83       |
|  Actual: High  |      177       |        111        |       1452      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.857   | 0.886  |  0.871   | 1803.0  |
|     low      |   0.648   | 0.621  |  0.635   |  457.0  |
|    medium    |   0.855   | 0.834  |  0.845   | 1740.0  |
|   accuracy   |   0.833   | 0.833  |  0.833   |  0.833  |
|  macro avg   |   0.787   | 0.781  |  0.783   | 4000.0  |
| weighted avg |   0.832   | 0.833  |  0.833   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 61
Depth of tree: 5
```



max_depth=10

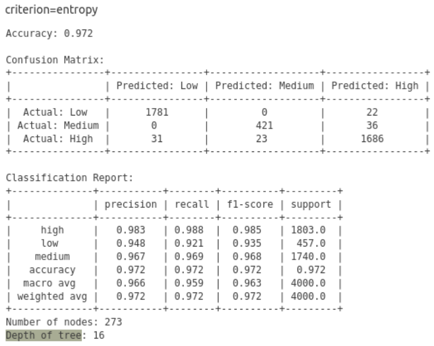
```
Accuracy: 0.97075

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       3        |        420        |        34       |
|  Actual: High  |       34       |         24        |       1682      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.98    | 0.988  |  0.984   | 1803.0  |
|     low      |   0.946   | 0.919  |  0.932   |  457.0  |
|    medium    |   0.968   | 0.967  |  0.967   | 1740.0  |
|   accuracy   |   0.971   | 0.971  |  0.971   |  0.971  |
|  macro avg   |   0.964   | 0.958  |  0.961   | 4000.0  |
| weighted avg |   0.971   | 0.971  |  0.971   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 255
Depth of tree: 10
```



max_depth=15

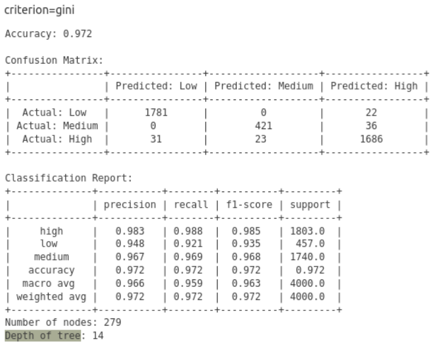
```
Accuracy: 0.972

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 279
Depth of tree: 14
```

2. criterion: entropy vs gini vs log_loss Those hyperparameters affects number of nodes and depth of the tree. The best results are for gini criterion, because



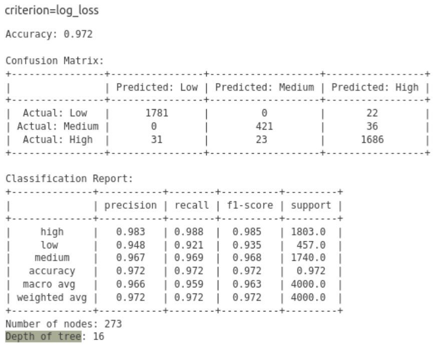
criterion=entropy

```
Accuracy: 0.972

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 273
Depth of tree: 16
```



criterion=gini

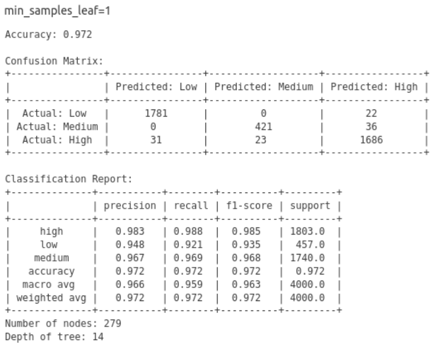
```
Accuracy: 0.972

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 279
Depth of tree: 14
```



criterion=log_loss

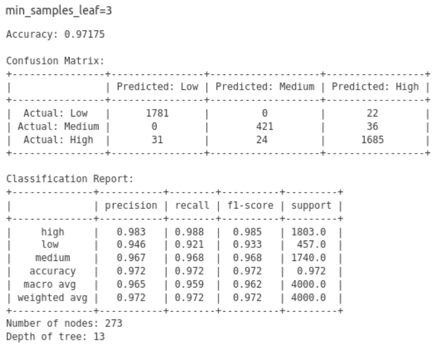
```
Accuracy: 0.972

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 273
Depth of tree: 16
```

3. min_samples_split: 1 vs 3 vs 5 There is no significant difference between the results. It is worth noting that depth decreases with the increase of min_samples_split, because requiring more samples to split a node results in fewer splits overall, leading to a simpler and shallower tree structure



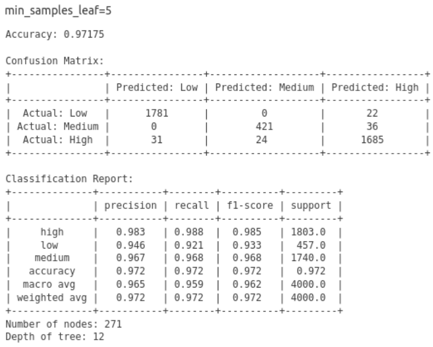
min_samples_leaf=1

```
Accuracy: 0.972

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         23        |       1686      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.948   | 0.921  |  0.935   |  457.0  |
|    medium    |   0.967   | 0.969  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.966   | 0.959  |  0.963   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 279
Depth of tree: 14
```



min_samples_leaf=3

```
Accuracy: 0.97175

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         24        |       1685      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.946   | 0.921  |  0.933   |  457.0  |
|    medium    |   0.967   | 0.968  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.965   | 0.959  |  0.962   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 273
Depth of tree: 13
```



min_samples_leaf=5

```
Accuracy: 0.97175

Confusion Matrix:
+----------------+----------------+-------------------+-----------------+
|                | Predicted: Low | Predicted: Medium | Predicted: High |
+----------------+----------------+-------------------+-----------------+
|  Actual: Low   |      1781      |         0         |        22       |
| Actual: Medium |       0        |        421        |        36       |
|  Actual: High  |       31       |         24        |       1685      |
+----------------+----------------+-------------------+-----------------+

Classification Report:
+--------------+-----------+--------+----------+---------+
|              | precision | recall | f1-score | support |
+--------------+-----------+--------+----------+---------+
|     high     |   0.983   | 0.988  |  0.985   | 1803.0  |
|     low      |   0.946   | 0.921  |  0.933   |  457.0  |
|    medium    |   0.967   | 0.968  |  0.968   | 1740.0  |
|   accuracy   |   0.972   | 0.972  |  0.972   |  0.972  |
|  macro avg   |   0.965   | 0.959  |  0.962   | 4000.0  |
| weighted avg |   0.972   | 0.972  |  0.972   | 4000.0  |
+--------------+-----------+--------+----------+---------+
Number of nodes: 271
Depth of tree: 12
```

#### References

https://towardsdatascience.com/whats-the-best-way-to-handle-nan-values-62d50f738fc
https://www.datacamp.com/tutorial/decision-trees-R https://stats.stackexchange.com/questions/399430/does-categorical-variable-need-normalization-standardization https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing