|  |  |  |
| --- | --- | --- |
| **Selected Features** | **Actual Description** | **Valuable (1, 0)** |
| Avg\_environmental\_importance |  | 1 |
| Avg\_government\_satisfaction |  | 1 |
| Avg\_education\_importance |  | 1 |
| Most\_Important\_Issue |  | 1 |
| Avg\_monthly\_expense\_on\_pets\_or\_plants |  | 1 |
| Will\_vote\_only\_large\_party | Uniform distribution across 3 states | 0 |
| Financial\_agenda\_matters | Uniform distribution | 0 |
| Looking\_at\_poles\_results | Uniform distribution | 0 |
| Married | Binarization of “Avg\_environmental\_importance” | 0 |
| Gender | Random binary | 0 |
| Voting\_Time | Random binary | 0 |
| Financial\_balance\_score\_(0-1) | Random Numerical 0-1 | 0 |
| %Of\_Household\_Income | Random Numerical 0-1 | 0 |
| Yearly\_IncomeK | Random Numerical 0-1000 | 0 |
| Overall\_happiness\_score | Random Numerical 0-10 | 0 |
| Garden\_sqr\_meter\_per\_person\_in\_residancy\_area | Random Numerical 0-100 | 0 |
| Avg residancy altitude |  | 1 |
| Yearly\_ExpensesK |  | 1 |
| %Time\_invested\_in\_work | Random Numerical 0-100 | 0 |
| Avg\_Satisfaction\_with\_previous\_vote | Linear combination of “Avg\_Satisfaction\_with\_previous\_vote” & “Avg\_monthly\_household\_cost” & “Political\_interest\_Total\_Score” | 0 |
| Avg\_monthly\_household\_cost | Linear combination of “Avg\_Satisfaction\_with\_previous\_vote” & “Avg\_monthly\_household\_cost” & “Political\_interest\_Total\_Score” | 0 |
| Political\_interest\_Total\_Score | Linear combination of “Avg\_Satisfaction\_with\_previous\_vote” & “Avg\_monthly\_household\_cost” & “Political\_interest\_Total\_Score” | 0 |
| Phone\_minutes\_10\_years | Polynomial function of “Avg\_environmental\_importance” | 0 |
| Avg\_size\_per\_room | Polynomial function of “Avg\_government\_satisfaction” | 0 |
| Weighted\_education\_rank |  | 1 |
| %\_satisfaction\_financial\_policy | Random Numerical 0-100 | 0 |
| Avg\_monthly\_income\_all\_years | Function of “Avg\_monthly\_expense\_on\_pets\_or\_plants” | 0 |
| Avg\_monthly\_expense\_when\_under\_age\_21 | Function of “Avg\_monthly\_expense\_on\_pets\_or\_plants” | 0 |
| AVG\_lottary\_expanses | Function of “Avg\_monthly\_expense\_on\_pets\_or\_plants” and “Avg\_environmental\_importance” | 0 |
| Last\_school\_grades | function of “Most\_Important\_Issue” | 0 |
| Age\_group | Uniform discrete distribution 1-3 | 0 |
| Number\_of\_differnt\_parties\_voted\_for | Binomial distribution (P=0.5, n=10) | 0 |
| Number\_of\_valued\_Kneset\_members |  | 1 |
| Main\_transportation | Uniform discrete distribution 1-4 | 0 |
| Occupation | Uniform discrete distribution 1-5 | 0 |
| Num\_of\_kids\_born\_last\_10\_years | Geometric distribution P=0.8 | 0 |
| Occupation\_Satisfaction | Uniform discrete distribution 1-10 | 0 |

**Data Separation**

* How did you split the data? Randomly? Have you stratified it? In what ratios?

**Visualization and identification of features**

* What is the type of each feature?
* Great use of visualizations throughout the work, in order to gain insights into the data.

**Categorical to Numerical**

* Good work on detecting which features are binary, which are categorical with order, and which don't have any particular order and require transforming using a one-hot-encoding scheme.

**Noise, Outliers and Imputation**

* Good choice looking at the distributions of each feature prior and after imputing by the mean value, in order to see that it's not optimal.
* It was a good choice looking which features have strong linear correlation, with other features, so that you could impute the values using this correlation, however a threshold value of 0.8, while reasonable, is probably too low, where your original value of 0.93 would have been an excellent choice.
* For the imputation of categorical features – You did ok, but it would have been preferable to find a set of k-nearest neighbors, by defining some distance metric between rows, and then impute according to the median value of the k closest rows. Since by imputing just based on the distribution of a single feature, you are completely neglecting all of the existing relationships between different features.
* It was an excellent decision to look for non-logical values, and replace them with NaN, to be later imputed.
* Why do you need to assume that most features come from normal distribution, where you can estimate the real distribution yourself, by plotting them? It would have been better to use the z-score test for outliers only on features which exhibit normal distribution.
* Although it is reasonable to assign a sample as an outlier based on only one feature, it is less than optimal approach. It would have been preferable to at least test whether a sample holds values which are considered outliers in more than one feature, since what might look like an outlier in 1D (the feature space), might be perfectly reasonable in the d-spaces of the sample.
* It was very good of you to make sure that by removing outliers you don't remove too many samples.

**Normalization and Scaling**

* Overall you did great job with the normalization, properly identifying which features should be normalized using the min-max method, and which one needs to be normalized using standard scalar method.
* In general your choices of which feature to normalize how, were good, there were a few features which you normalized not in accordance to their distribution, but those were also feature whose distribution doesn't have an immediate clear scaling, like 'Number\_of\_differnt\_parties\_voted\_for', which you've scaled with the standard scalar, although it has a binomial distribution. It is possible that it slightly hurt your results.

**Feature selection, Filter and Wrapper Methods**

* Good of you to use more than one filter method, and it is especially good that you have experimented with different threshold values in order to find the optimal one.
* it would have been better to also try at least one more different classifier with the wrapper method, as each different classifier has its own quirks, assumptions and biases underneath.
* Overall you seem to have set the hyper-parameters of the wrapper method well enough, it was especially good to see that you, again, experimented with different k values in order to get a robust assessment of the optimal number of features.
* Great choice to use the stratified k-fold.
* Great work also using MI as a filter method and adding an embedded method, which, with all methods put together, allows you to get a strong estimation, at least for the most important features.
* Great insight regarding the 3 features which you've deleted at first, it was very good to see that you recognized their importance even though they were removed by the variance test.

**Bonus**

* Excellent analysis and visualization of the feature's roles with respect to the labels, you have managed to notice almost all of the important relations!
* Good implementations of the Relief and SFS algorithms. It was especially good that you've performed ablation studies in order to test the importance of different numbers of iterations and, classifiers used with SFS, and k values – Overall it was a very thorough and complete work!

**Summary**

Kudos on an excellent work.

It is very well documented, and all of your decision and actions are explained in detail and are logical and sensible. It is clear that you have gained good understanding of the data and features' roles.

You have performed the proper pre-processing steps, and your early studies into the data has allowed you to keep vital features, even when some of you're filter methods have suggested to remove them.

It would have helped you to be just slightly more conservative, in order to get more points on the feature selection task, since you did miss on important feature.

Also, you should have included details on each feature's type and on how you have split the data, and sampled each set.

Other than that, great work.

**Final Grade:** 107