Capstone Project Report – 2023

By: Theekshana Wijesinghe

Introduction (Problem Definition)

In this project, I'm going to rank nursery school applications by applying Machine Learning methodologies and this problem can be found at the popular datasets in the UCI achieves.

The rank of the nursery school application is based off a hierarchical decision model with multiple values, thus this fall under a **multiclass classification** problem.

The goal in this task is to rank the applications depending on the features provided. The interesting aspect of this project is that all the feature parameters are **categorical**.

Data

In this problem, there are 8 categorical features and one target with multiple classes. As stated above all the feature are categorical.

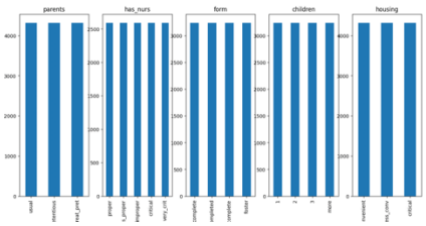
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
parents	Feature	Categorical				no
has_nurs	Feature	Categorical				no
form	Feature	Categorical				no
children	Feature	Categorical				no
housing	Feature	Categorical				no
finance	Feature	Categorical				no
social	Feature	Categorical				no
health	Feature	Categorical				no
class	Target	Categorical				no

The dataset contains 12960 records that is acceptable for model training.

Methodology

Exploratory Data Analysis

- Checking for missing data values
- Checking for duplicates
- Checking for data frequencies



Feature Engineering

- Identifying the features that has the most correlation with the target variable by applying Chi-squared test
- Identifying target classes with relatively low data that would impact classifier performance
- Using One Hot Encoding for categorical data encoding
- Assuming the rank is ordinal, ranking the target accordingly

Model building

- Start with the simplest model, Logistic Regression and using least number of features that gave p = 0 to Chi-squares test (has_nurs, health)

Evaluate model

- Using accuracy, precision, recall and f1 score to evaluate the model and using confusion matrix
- Drawing Multiclass ROC for the results

Expand the features

- By using the Chi-squared results, expand the features and evaluate the model

Feature engineering revisited

- Remove the classes that has low frequency and re-evaluate with the best scores

Model comparison

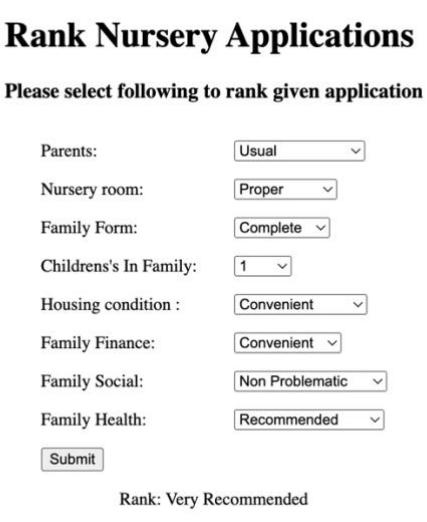
- Evaluate the Logistic regression, Random Forest Classifier, KNeighbour Classifier and Support Vector Classifier with default params with the best features selected from above steps

Hyperparameter Tuning

- Using Grid Search try to improve the best model selected from above step

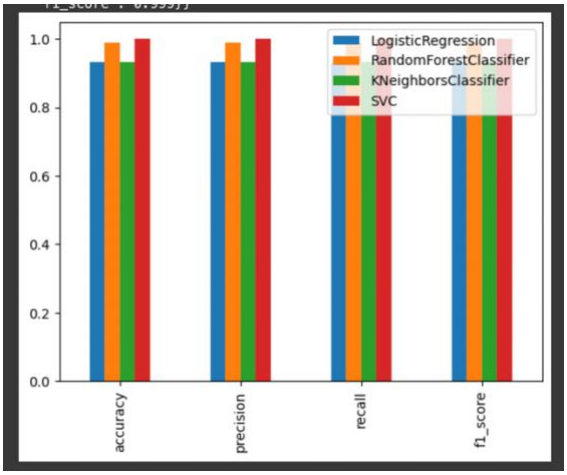
Web application

- A web application with form to get the ranking.



Results

- Chi-squared test gave p <= 0.05 for all the feature variables against target variable therefore best results were given when all the features are selected for modeling
- Out of the selected models, Support Vector Classifier is the most performing classifier for this problem
- Unoptimized SVC showed 0.999 for accuracy, precision, recall and f1 score.
- SVC with C=10 showed 1 for accuracy, precision, recall and f1 score



Conclusion

- Chi-Squared test is an accurate metric to identify correlation between categorical variables.
- All the features are required to generate best solution.
- Support Vector Classifier can be considered as the best model for this problem.

Discussion

It is interesting to note that when C is set to 10, the Support Vector Classifier (SVC) achieved a perfect score of 1 for accuracy, precision, recall, and F1 score. This outstanding performance could be labeled as an ideal model for this particular problem. Assuming potential overfitting, I ran setup by adjusting the test_size parameter, ranging from 0.2 to 0.4. However, the results remained consistent.

For precision, recall and f1 score calculation, the average=**weighted**, was used as this is a multiclass problem. This decision was based on the fact that classes are not evenly distributed even with removing recommend class in the target variable.