UFUG1601 L07 AI project Report

Zihan Lin[*1] and Jiongxu Chen[*2]

[*]HKUST(GZ)
[1]zlin368@connect.hkust-gz.edu.cn
[2]jchen074@connect.hkust-gz.edu.cn

Abstract

This report presents the findings of a comprehensive analysis conducted to predict average lifespan in different countries using data from the World Bank's World Development Indicators and the World Health Organization. The project aimed to implement and evaluate a data linkage system and a classification system by training machine learning models with World Development Indicators data.

In this project, we compare various classification algorithms, including k-NN, Decision Trees, Random Forest, XGBoost and Voting Classifier and explore the impact of feature engineering and selection on enhancing model predictive performance. What's more, we discuss not only the reliability of the models, but also what we can do to further improve the model performance, and propose additional techniques that can potentially enhance classification accuracy. We utilize the Python language and libraries such as Pandas, NumPy, Matplotlib, Seaborn, and sk-learn, to process data, implement machine learning models and visualize model performance in the project.

Overall, this project provide valuable insights into the application of machine learning models in predicting average lifespan using World Development Indicators data.

1 Introduction

Artificial intelligence and data science have been repeatedly applied in practical problems, ranging from predictive analyses to decision-making systems. And in this project, we use AI and DS knowledge to tackle a real-world problem: predicting average life expectancy across different countries using the World Development Indicators published annually by the World Bank. The project requires the implementation and evaluation of a data linkage system and a classification system, and the main task is to predict the class feature of Life Expectancy (Low, Medium, and High) by using other features and utilizing machine learning models trained on sound data science principles.

The significance of this problem lies in its demand for our capacity to exercise independent judgment in designing, training, and fine-tuning models, rather than merely implementing a predefined algorithm. This skill is essential for any successful AI practitioner, as real-world tasks often involve multiple approaches and irregularities in the data that cannot be exhaustively predefined.

2 Background and motivation

The World Bank plays a vital role in global development by publishing the World Development Indicators annually. These indicators provide a comprehensive set of high-quality, internationally comparable statistics about the fight against poverty. We want to use this wealth of information to

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

predict average life expectancy across various countries and understand how the information can be used to predict average life expectancy.

Implementing this project is of great importance in the real-world. First, analyzing the data from the World Health Organization's statistics effectively can explore the relationship between development indicators and life expectancy, which is a rich ground for exploring how economic, social, and health indicators correlate with life expectancy. Besides, accurate predictions can have profound impact for global development. By implementing machine learning algorithms, we can provide insights that could guide policy making and resource allocation, in the pursuit of improved living standards all over the world.

3 Methods

In order to implement this project smoothly and achieve prediction and evaluation, we use several methodologies based on the project requirement and our own design.

We have collected data on some features from various countries, such as "Access to electricity (% of population)", "Fertility rate, total (births per woman)" and "Individuals using the Internet (% of population)" , from the World Bank's World Development Indicators and the World Health Organization. And we obtain data on the feature "Life expectancy at birth (low, medium and high)", as shown in Figure1. Therefore, the main problem in this project is a multi-classification problem which requires us to perform classifiers to predict the class feature "Life Expectancy at Birth" using other features we have collected.

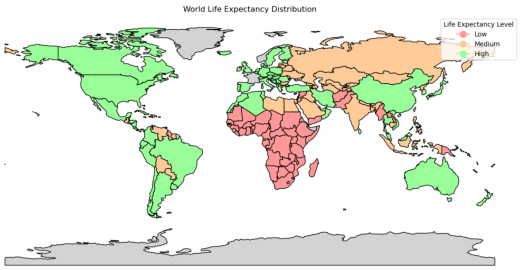


Figure 1: World life expectancy distribution

3.1 Task A: Classification

Training and testing of single models. In order to make comprehensive predictions and gain a broad understanding of the performance of various classifiers, we use different kinds of classifiers in this project, like k-NN(k=3 and k=7), Decision Tree, Random Forest and Gradient Boosting (XGBoost). To visualize and evaluate the performance of different kinds of classifiers, we take Accuracy and Confusion Matrix as metrics .

Specific investigation on k-NN model. In the k-NN classifier, in addition to setting k to 3 and 7, we also vary the value of k within the range of 1 to 11 and evaluate the overall performance of k-NN model. From this we obtain some discoveries regarding the performance of k-NN model.

Combination of multiple classifiers. Given that using a single classifier has its limitations and may lead to one-sided predictions , we attempt to combine multiple classifiers to predict the feature. Therefore, we use Voting Classifier as a new classification algorithm, which combines Random Forest, Gradient Boosting and SVC classifiers. And we also take the method of cross-validation to compare and evaluate the performance of XGBoost, Random Forest and Voting Classifier, as cross-validation can divide the dataset into multiple subsets, then use some for training and the remainder for testing. This method can reduce the variance in model assessment results and enhance the stability and reliability of the evaluation.

Further methodology to evaluate models. We combine cross-validation and learning curves to visualize the performance of Voting Classifier, because learning curves can demonstrate how a

model's performance metrics change with varying numbers of training samples,with an effect that this combination can help for a comprehensive evaluation of the model.

3.2 Task B:Feature engineering and selection

Feature-importance. Decision Tree, Random Forest and XGBoost are tree-based machine learning models, so feature-importance is a common attribute in them, from which we can further explore the field of feature selection in our project and investigate the performance of models deeply. Therefore, we take the top 6 most important feature in each model and calculate the weights of feature importance.

Feature engineering and selection. The main operation in this task is implementing feature engineering and selection. So we produce a new dataset with 211 features (20 original features, 190 features generated by interaction term pairs and 1 feature generated by clustering). Then we implement the feature selection using three different methods. The first method is selecting 4 features by SelectKbest from the new dataset. The second method is selecting thought PCA by taking the first four principal components from the new dataset. And the last method is taking the first four features from the original dataset.

Use of classifiers. We choose k-NN(k=3) and Decision Tree classifiers to predict the feature "Life Expectancy at Birth", using three methods above separately, and then analyze the performance of models and further the impact of feature selection methods.

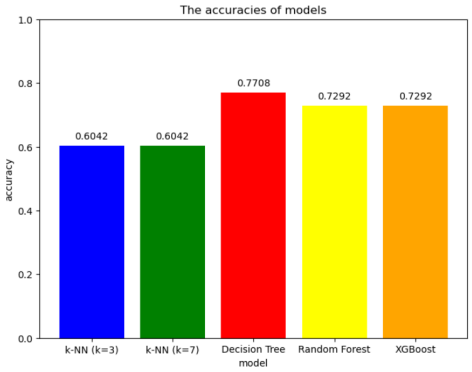
Feature selection methodologies. Then we discuss the methods we used to select features, especially SelectKBest and PCA. To investigate SelectKBest, we obtain 4 best features selected by it, and compare these features with important features in tree-based machine learning models. As for PCA method, first, we use PCA dimensionality reduction to Visualize the contrast between clustering results and actual labels, with purposes to intuitively display clustering results and evaluate the performance of clustering algorithms, thereby further guiding the selection and optimization of clustering algorithms. Next, we check the correlation between the features after PCA transformation and the original labels, to understand which principal components might be more useful for predicting the target variable.

4 Result and analysis

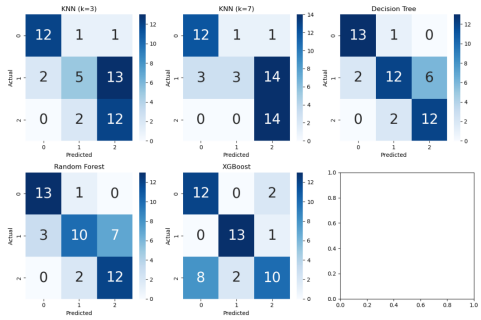
4.1 Classification result

Training and testing of single models. First, we analyze the performance of each single classifier though metrics like Accuracy and Confusion Matrix. As shown in Figure2a, the accuracies of different classifiers fall within the range of 0.6 to 0.8, among which Decision Tree has the highest accuracy of 0.77. Hence, based on this metric, Decision Tree demonstrates the best performance among these classifiers. And from the Confusion Matrix images in Figure2b, we can see that Decision Tree, Random Forest and XGBoost classifiers have higher values on the diagonal, which means they have correctly classified most of the samples. Decision Tree and Random Forest performed similarly in this test, but Random Forest typically improves performance and stability by integrating multiple decision trees. XGBoost enhances performance by using gradient boosting and regularization, which is also reflected in the image.

Specific investigation on k-NN model. Due to the poor performance of k-NN(k=3 and k=7) in the above tests, we analyze the impact of the value of k on the performance of the k-NN model. As shown in Figure3a, regardless of the value of k taken from 1 to 11, the accuracy of the k-NN model remains low, with the highest accuracy reaching only about 0.65 (when k=4). Hence, k-NN model is not suitable to implement the prediction in this project. What's more, from the Confusion Matrix image in Figure3b, we can found that the classification accuracy for the Medium category in k-NN(k=4) model is generally low, which is likely the main factor contributing to the poor performance of k-NN model. The reason for the low classification accuracy of the medium category will be mentioned and explained in 4.2(PCA evaluation).

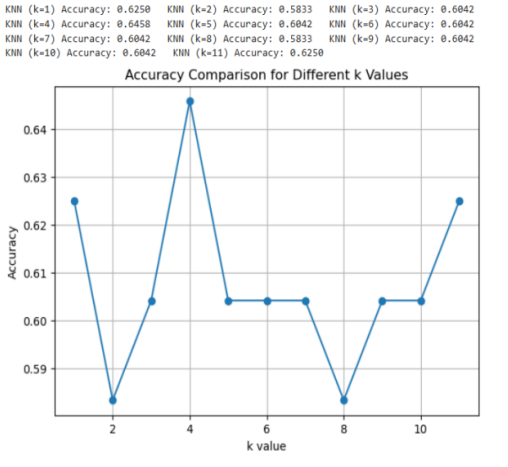


(a) Accuracies of different models

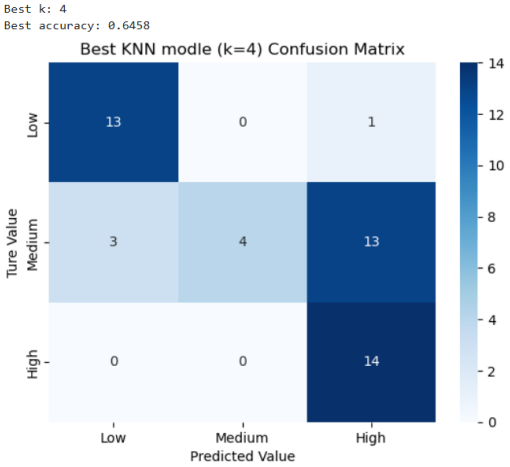


(b) Confusion Matrix images of different models

Figure 2: Performance of different models



(a) Accuracy comparison for different k values



(b) Performance of Best k-NN

Figure 3: k-NN performance

Combination of multiple classifiers and crossing-validation. After generating Voting Classifier as a combination of other classifiers, we analyze the effect of cross-validation on evaluating the performance of XGBoost, Random Forest and Voting Classifier. After employing this method, as shown in Figure4, we obtain the accuracy of Random forest and XGBoost, both of which are about 0.75, slightly higher than the accuracy mentioned above without using this method. What's more, the accuracy of Voting Classifier reach 0.78, close to 0.8, which is higher than the accuracy of using other single classifiers. Therefore, the Voting Classifier not only combines other single classifiers to make predictions more comprehensive, but also has better predictive performance than other single classifiers.

```
accuracy of XGBoost:0.7466666666666667
accuracy of Random Forest:0.7466666666666667
accuracy of Voting Classifier:0.7888888888888889
```

Figure 4: Accuracies of using crossing-validation

Learning curve. The next step is using the learning curve to study the performance of Voting Classifier in detail. As shown in Figure5, the training score remains almost at 1.0, indicating that the model performs very well on the training set and can almost perfectly fit the training data , and the cross-validation score generally increases with the number of training examples,which means that the performance of Voting Classifier slightly improves with the increase of training examples. Clearly, the performance of this learning curve also has some shortcomings, such as the significant

gap between the training scores and cross-validation scores, which is usually a sign of over-fitting. This is exactly where we need to further improve in the future, by employing additional techniques, such as increasing the amount of data, adjusting model complexity, and applying regularization.

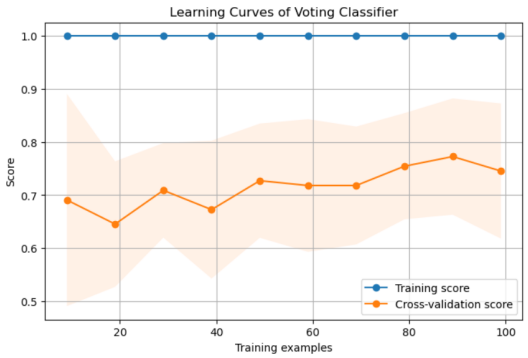


Figure 5: Learning curve of Voting Classifier

4.2 Feature engineering and selection result

Feature-importance. From three figures (Figure6 about top 6 most important features of Decision Tree, Random Forest and XGBoost, we can find that the most important features of the three models share many similarities, and two of these features are common to all three models, which are "Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions" and "People using at least basic drinking water services". So there is a strong correlation between there important features and the target variable "Life Expectancy at Birth". Besides, in the feature importance of Decision Tree, only 5 features have weights greater than 0, indicating that the distribution of feature importance weights is concentrated, while the weight distributions of feature importance are relatively scattered and average in Random Forest and XGBoost model. So this is one reason why performance of Decision Tree is better than that of Random Forest and XGBoost, as mentioned in task A.

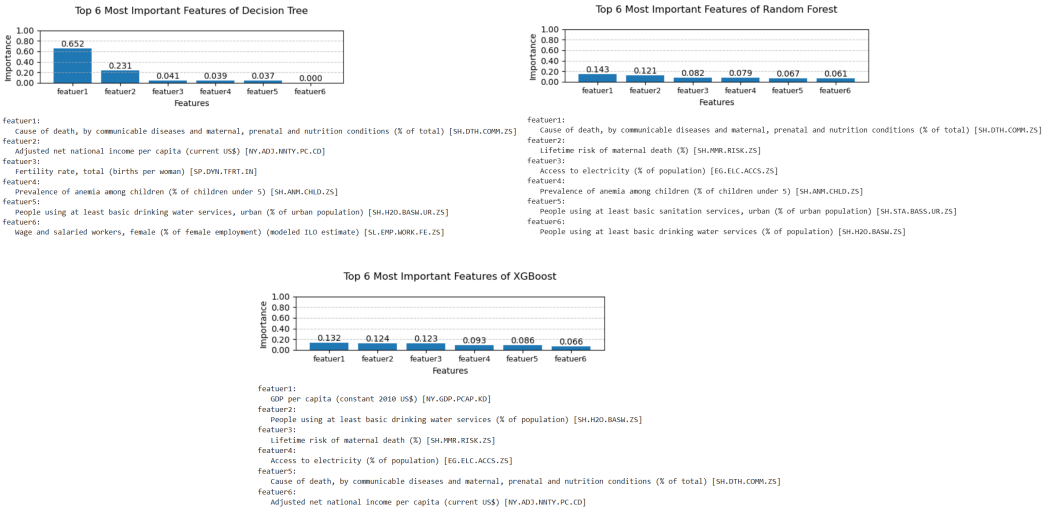


Figure 6: Top 6 most important features

Feature selection effect and classifiers performance. Next, after using 3 methods to select features and implementing k-NN(k=3) and Decision Tree model, we gain the result of Figure7. From the figure, we can learn that the model performance using SelectKBest to select features is the best, while the performance using PCA is poor. And the specific reasons are detailedly explained next.

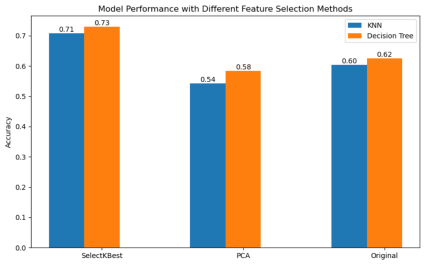


Figure 7: Feather selection effect

SelectKBest evaluation. As shown in Figure8, the four features selected by SelectKBest are "Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions", "People using at least basic drinking water services", "People using at least basic sanitation services,urban" and "Prevalence of anemia among children", which all are the important features of tree-based models mentioned above and have a strong correlation with the target variable. That is exactly the reason why the model performance using SelectKBest to select features is the best.

That is to say, in order to achieve better model performance, it is indeed necessary to select features that have strong correlations with the target variable. However, the new features we obtain through interaction term pairs contain numerous extraneous components. Therefore, we can conclude that this method of generating new features without selectivity is not suitable for the problem in this project.

```
Features selected by SelectKBest:
Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions
People using at least basic drinking water services
People using at least basic sanitation services, urban
Prevalence of anemia among children
```

Figure 8: Features selected by SelectKBest

PCA evaluation. As shown in Figure9 of PCA dimensionality reduction, the region of the Medium part basically coincides with the region of the other two parts, which actually explains the reason why the performance of k-NN model in predicting the medium part is generally poor, with a result that the overall accuracy of k-NN model is not ideal.

As shown in Figure10, the correlation between each principal component and the target variable after dimensionality reduction by PCA is very low (one is between 0.4 and 0.5, and the rest are around 0.1, which is much less than 1), so the performance of model trained by using these principal components is also not ideal. Then we can conclude that the model performance using PCA is poor.

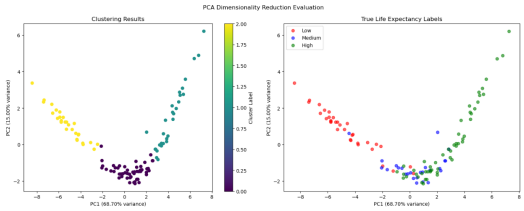


Figure 9: PCA dimensionality reduction

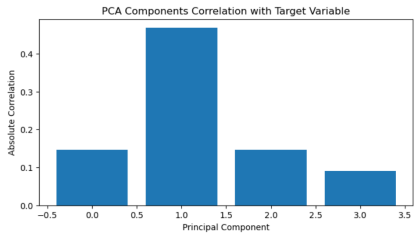


Figure 10: PCA components correlation with target variable

5 Conclusion

In this project, we use different kinds of classifiers and various methods of feature selection to predict the target variable "Life Expectancy at Birth" and evaluate the performance of models.

For the feature selection, it is best to use SelectKbest for feature selection in this project, which can improve the performance of the model, instead of using PCA dimensionality reduction or direct selection.

For the classifier selection, we try to combine multiple models, so we use Voting classifier that combined Random Forest, XGBoost and k-NN. The performance of Voting Classifier is better than that of Random Forest, XGBoost and Decision Tree when performing k-fold verification, indicating that combining multiple different classifiers can play a greater advantage. Besides, the performance of single Decision Tree model is also great, and it is not cumbersome to build with its simple structure.

This project is of great significance in the real-world, especially the prediction of life expectancy. And the design and methodologies used in the project can help for the prediction in the real-world and further contribute to the improvement of living standards all over the world.

6 Future work

Data preprocessing. In this project, we use the median to fill missing values before implement the main tasks. In the future, we will consider more ways to preprocess the data, such as filling missing values with the mean, directly deleting countries with missing values, etc.

Feature settings. From the Figure1, we can observe that the distribution of the feature "Life Expectancy" on the map exhibits a very regular pattern, and there is an obvious trend in the life expectancy of each continent. Therefore, we believe the continent to which a country belongs can be used as a new feature in future research.

Model optimization. For the model used in the project, the accuracy of Voting Classifier is very high (close to 0.8), so in the future we will continue to improve Voting Classifier model and try to effectively combine it with SelectKbest to further improve the performance and achieve better prediction results. At the same time, we noticed that some features have great impact on the results, so we will try to combine these important features in a certain way to form new features for training and discard the features that have little impact on the results with a pursuit of improving model performance and prediction accuracy.