CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*)

The Cross-Industry-Standard-Process for Data Mining is very important for data science projects. This is a widely used technique in the industry for data mining. It has a well-structured methodology and systematic approach to tackling a business problem. It is a flexible and iterative process that allows adjustments and changes as new data is available. It provides a valuable framework for managing and executing data-driven projects.
For the same reason, I have used this methodology for my project that predicts death rates using different features. The dataset was already provided to me by the professor, and it was already split into training and test data.

Business Understanding:

1. **Business-objective:**
   Cancer is one of the most deadliest disease that has a very high death rate, when diagnosed with it many people think about their chances of survival. Many health insurance companies are using different machine learning approaches to figure out, if they can predict the death rate accurately, so that they can provide the health cover accordingly. Most people nowadays are aware of health insurances and the government has done a tremendous job of making people aware of the importance of health insurance. But as the demand for health insurance increases, more companies will set their business. This makes the market more competitive. Most people buy those insurances that are affordable and at the same time provides the most coverage. But these pricing depends on various factors like the persons health history, age, employment status, etc. It is important for them to accurately predict death rate, as if they predict it wrong, it can potentially lead to losing customers.

2. **Data Understanding:**
   Using the pandas data frame, the CSV file was converted into a data frame. This helps to work on columns easily. The data contained 2438 rows and 35 columns. Most of the columns were of numeric type and contained information like median age, employment status, unique id and many more. Columns that were not identified as numeric were dropped in the data pre-processing stage. The data covered range of features from age, gender, employment status, marital status and ethnicity.

3. **Data Preparation:**
   For the initial task, the feature that I selected was ‘MedianAgeMale’ and ‘MedianAgeFemale’, the purpose of selecting these was to see how I can predict death rate from age and gender because this is a kind of data that every person can provide. The error score from the features was high, but the difference between the train and test set was marginal, stating that the model had fitted well in spite of not cleaning the data. For the second task and the third task same number of were selected, but in the third task, the outliers were removed using IQR (Inter Quartile Range).

4. **Modelling:**
   In the modelling stage, many key insights were observed, where in the first task, the MSE scores were very high at 800, this implies that a single feature cannot provide high accuracy data prediction, but the small difference between the trainset score and testset score suggests that the model was balanced. For the second task, there was a distinct improvement in the model's performance, reducing the MSE score significantly. The difference between the trainset score and test set score suggests that the model was slightly overfitting. For the third task, the model was run two times, one with removing outliers and one-without outliers. The models ran in this test were Linear Regression, Elasticnet, Decision Tree and Random Forest Regressor. A significant over-fitting was observed when the outliers were removed.

5. **Evaluation:**
   Task 1: The selected features had a very low correlation with the target variable, thereby decreasing the model accuracy. It was not possible to accurately predict the target, hence more features were selected in the further tasks.
   Task 2: Adding features helped in improving the model prediction and reducing the errors, stating that multiple features are required to get higher model accuracy and performance. It also stated that the model fitted well with the training set, indicating the features selected are reliable and clean.
   Task 3: When comparing the model performance with outliers and without outliers, a key insight was derived that some values in outliers are important, and a different regularisation technique must be used to improve the performance. The least MSE score was for Linear Regression, stating that the given dataset fits well with the algorithm. I was expecting other algorithms to perform better but that did not happen. Possibly because some important features might be not cleaned well or the hyperparameters were not tuned to fit the model well. Due to Elasticnet regularisation, a very high MSE values was observed, indicating that elasticnet is very sensitive to outliers and possibly use a less sensitive regression model. Most of the models in task 3 showed overfitting suggesting that feature engineering is important to increase performance.

6. **Deployment:**
   The provided data can help in predicting the death rate using linear regression model, but the performance scores suggest that further improvements are needed in the data processing and feature selection. It is not necessary that the more the number of features, the better the prediction, it is equally important to select which data is important for the target variable. In the insurance industry, all the business growth depends on how accurate the predictions are and hence deployment can only be done when the model performance and accuracy are very high. Through this analysis, I believe my model is not yet ready for deployment and needs improvisation in the above-mentioned areas. Once the necessary improvements are performed, the model is ready for deployment.

References:

This Dataset consolidated from census data in the USA. The information is based on information related to US counties.The dataset contains 33 different features (demography, medical information). All the credits goes to the sole Authors.