Results

Effect of dietary patterns on the cause (ACR) as well as the mortality of Chronic Kidney Disease (CKD) patients

Data Analytics: Major Research Project

Sayed Ahmed

MSc in Data Science and Analytics

Ryerson University

500869723

Supervisor

Dr. Youcef Derbal

Information Technology Department

Ted Rogers School of ITM

Ryerson University

# Results

## Food Groups and Mortality

Experiments to find association between food groups and CKD mortality using PCA and Regression show that Grains (-0.84) and Fruits (-0.43) have strong negative correlations with CKD mortality. Data exploration also reflects that i.e. data shows that when Grains and Fruits are taken less, mortality is high. As correlation for fruits is -0.43 i.e. not very high, hence, Fruits can be thought of somewhat associated.

|  |  |
| --- | --- |
|  |  |
|  |  |

Vegetables show positive (0.58) correlation i.e. when vegetables intakes are high, mortality is also high. As the correlation is 0.58, hence, this is not a very strong conclusion. Data shows this relation in older adults. Even though ratios of food intake to high-recommended amounts (as well as actual intake amounts) were used; age might have biased the correlation. This does not show conformity with doctors’ recommendation to take more vegetables for CKD patients. If this analysis has to be right: moderate (not high) or average recommended amount of vegetables intake might be a recommendation for the older adults.

|  |  |
| --- | --- |
|  |  |
|  |  |

## Food Subgroups and Mortality

Experiments to find association between food subgroups and CKD mortality using PCA and Regression show that Other vegetables (0.68), Red and orange vegetables (0.55), and Starchy vegetables (0.44) have positive correlations with mortality. Data Exploration and the plots as given below also show that mortality is low when the intake amounts are low, and mortality is high when intake amounts are high.

|  |  |  |
| --- | --- | --- |
| positive_subgroup_line_1.png | positive_subgroup_line_2.png | positive_subgroup_line_3.png |

Food subgroups such as Alcoholic beverages (-0.79), Added Sugars/Sugars and sweets (-0.64), Whole grains (-0.61), and ‘Nuts, Seeds, and Soy Products’ (-0.55) show the most negative correlation with CKD mortality. Data exploration also shows negative correlations as shown in the charts below.

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| negatively_subgroup_line_1.png | negatively_subgroup_line_2.png | negatively_subgroup_line_3.png |

## Food Groups, Food Nutrients, and Albumin Creatinine Ratio association

The experiments using PCA and Regression showed negligible effect on ACR with food groups and subgroups. However, some food groups and/or nutrients such as Dairy, and ‘Sugars, Sweets, and Beverages’, have higher and positive though negligible (0.02) affect than the others where Fruits (-0.01) showed negative effect. For nutrients Poly unsaturated fatty acids (-0.02), iron (-0.02) have negative correlation where Choline (0.02) showed positive correlation.

As the correlations are not significant further analysis can be done on the data especially for food groups and nutrients that are found important (using PCA) in the data as provided below:

Dairy, ‘Fats, oils, and salad dressings’,  Fruits, Grains, Protein, ‘Sugars, sweets, and beverages’, Vegetables, Avg\_energy\_kcal,  avg\_protein\_gm, avg\_carbohydrate\_gm, avg\_total\_fat\_gm, avg\_total\_saturated\_fatty\_acids\_gm, avg\_total\_monounsaturated\_fatty\_acids\_gm,  avg\_total\_polyunsaturated\_fatty\_acids\_gm, avg\_lutein\_zeaxanthin\_mcg, avg\_thiamin\_vitamin\_B1\_mg, avg\_riboflavin\_Vitamin\_B2\_mg, avg\_Niacin\_mg, avg\_Calcium\_mg,  avg\_Phosphorus\_mg, avg\_Magnesium\_mg, avg\_Iron\_mg, avg\_Zinc\_mg, avg\_Copper\_mg, avg\_Sodium\_mg, avg\_Potassium\_mg, avg\_Selenium\_mcg, Hexadecenoic\_gm, Octadecenoic\_gm

## Food Subgroups and Albumin Creatinine Ratio association

The experiments show ‘Milk desserts, Sauces, Gravies’ (0.22), and Alcoholic beverages (0.087) have more positive correlation with ACR than the other food subgroups i.e. taking more of these food subgroups results higher ACR values that reflect the current medical knowledge. However, the correlation is very low. Low values might still explain a correlation where ACR values might depend on other factors in together than only these food subgroups. Fruits and juicy baby foods show negative correlation (-0.04) though not significant i.e. taking high amount shows lower ACR values.

## Predictability of ACR values based on Dietary patterns

Experiments were conducted to discover if ACR values can be predicted using the food intake patterns. Machine Learning (ML) Approaches such as Regression, Polynomial Regression, Random Forest Regression, Bayesian, with or without 10 fold cross validations were applied on Food Subgroups intake dataset. Only the food subgroups that were found to be important using PCA were used for the ML approaches.

**Target Variables:**

Absolute ACR values and ACR Category were used as the target variables. For ACR category, ACR < 30 is assigned to class 0, and ACR > 30 is assigned to class 1.

Outcome

The best test set accuracies are found using the approaches such as: 10 Fold Cross Validation Polynomial Regression, Polynomial Bayesian with Cross Validation (68%), Polynomial Regression (57%), Bayesian on Polynomial fit (41%), Cross Validation with Polynomial Random Forest Regression (21%)

A list of the best performing approaches and the outcome are provided below. Complete list can be seen on outcome\_machine\_learning\_for\_acr\_values.xlsx

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature Data** | **Target** | **Approach used** | **MSE Train** | **MSE Test** | **RMSE Train** | **RMSE Test** | **R2 Score Train** | **Accuracy: Test R2 Score if not mentioned** |
| Data Not Normalized | ACR Value | 10 Fold Cross Validation Polynomial Regression | | |  |  |  | -0.957 - cross\_val\_score |
| Data Not Normalized | ACR Value | Polynomial Bayesian with Cross Validation |  |  |  |  |  | -0.682 - cross\_val\_score |
| Data Not Normalized | ACR Value | Polynomial Regression | 90965 | 52946 | 301 | 301 | 0.359 | -0.579 – r2\_score – test data |
| Data Not Normalized | ACR Value | Bayesian on Polynomial fit | 93047 | 47431 | 305 | 305 | 0.344 | -0.414 r2\_score on test data |

### Outcome when ACR category is used

After regression, y > 0.5 is assigned to category 1. Test accuracies are 88%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Target** | **Approach** | **Train Confusion (total, correct : [ TP, FN, FP, TN])** | **Test Confusion Mat.** |
| Not Normalized | acr\_category | Linear Regression | 6927, 87 :[6032,0,895,0] | 770, 88: [692, 0, 78, 0] |
| Normalized | acr\_category | Linear Regression | 6927, 87:[6032,0,895,0] | 770, 88: [692, 0, 78, 0] |

The high prediction accuracies might relate to the fact that only 10 to 14% population has ACR > 30. However, as cross validations also show high accuracy, it can be concluded that ACR values can be well predicted using Machine Learning approaches.

@todo : Include the corresponding regression result and line chart plots. Include PCA outcome in the appendix. Also, include all ML outcomes in the appendix.

# Appendix

## More plots to show positive or negative correlations:

Vegetables and CKD Mortality

|  |  |
| --- | --- |
|  |  |
|  |  |

**Food Subgroups and CKD Mortality**

|  |  |
| --- | --- |
|  |  |
|  |  |  |

## Grain and Mortality: Negative

|  |  |
| --- | --- |
|  |  |
|  |  |

## ML Approach Results

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data** | **Target** | **Approach used** | **MSE Train** | **MSE Test** | **RMSE Train** | **RMSE Test** | **R2 Score Train** | **Accuracy**  **/performance:**  **Test R2 Score** |
| Not Normalized | ACR Value | **10 Fold Cross Validation Polynomial Regression** | | | |  |  | -0.957 |
| Not Normalized | ACR Value | **Polynomial Bayesian with Cross Validation** |  |  |  |  |  | -0.682 |
| Not Normalized | ACR Value | Polynomial Regression | **90965.000** | 52946.43 | 301.604 | 301.604 | 0.359 | -0.579 |
| Not Normalized | ACR Value | Bayesian on Polynomial fit | **93047.912** | 47431.675 | 305.038 | 305.038 | 0.344 | -0.414 |
| Not Normalized | ACR Value | **Cross Validation with Polynomial Random Forest Regressor** | | | | |  | -0.394 |
| Normalized | ACR Value | **10 Fold Cross Validation with Linear Regression** | | | |  |  | -0.134 |
| Not Normalized | ACR Value | **10 Fold Cross Validation with Linear Regression** | | | |  |  | -0.134 |
| Normalized | ACR Value | LinearRegression | **139019.96** | 36128.194 | 372.854 | 190.074 | 0.020 | -0.077 |
| Not Normalized | ACR Value | LinearRegression | **139019.96** | 36128.194 | 372.854 | 190.074 | 0.020 | -0.077 |
| Not Normalized | ACR Value | Bayesian | **139091.64** | 35430.367 | 372.950 | 372.950 | 0.019 | -0.056 |
| Data Not Normalized | acr\_category | **10 Fold Cross Validation with Linear Regression** | | | |  |  | -0.053 |
| Data Normalized | acr\_category | |  |  |  |  |  | -0.053 |
| Data Not Normalized | ACR Value | Polynomial RandomForestRegressor | **77719.250** | 34344.340 | 278.780 | 278.780 | 0.452 | -0.024 |
| Data Not Normalized | acr\_category | LinearRegression | **0.112** | 0.092 | 0.335 | 0.303 | 0.005 | -0.008 |
| Data Normalized | acr\_category | LinearRegression | **0.112** | 0.092 | 0.335 | 0.303 | 0.005 | -0.003 |

# References

As used in earlier reports such as reports for exploration phase and reports for experiment phase. As well as all data and code used and submitted before.