

**Experimentation of AITA Project**

**Experimentation of the Mini Project**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EXPERIMENT NUMBER** | **M.L MODEL** | **EXPERIMENT PERFORMED** | **RESULTS** | **CONCLUSION** |
| 1 | Linear Regression | Target – Seasonal Vaccine, Data Sampling – 70%, All Features | AUC – 0.833, C.A – 0.762, F1 – 0.761, Precision – 0.761, Recall – 0.762 | With all the features (numerical and categorical) the accuracy achieved is 83 % |
| 2 | Linear Regression | Features – employment \_occupation, behavioural large gather, behavioural mask | AUC – 0.832, C.A – 0.759, F1 – 0.759, Precision – 0.759, Recall – 0.759 | No major change in AUC hence we can say these features are not dominant in the prediction. |
| 3 | Linear Regression | Ignored Features – race, sex, income, poverty, employment status. | AUC – 0.841, C.A – 0.764, F1 – 0.764, Precision – 0.764, recall – 0.764 | No Major change in accuracy but the accuracy increased by 1% meaning one of the ignored features is affecting the accuracy of prediction by 1%. |
| 4 | Linear Regression | Kept Features – chronic med condition, marital status, hhs\_geo\_region,behavioural large gatherings, race , income property , large gatherings, income Property , employment status | AUC – 0.646, C.A – 0.608, F1 – 0.604, Precision- 0.607 | So, we can see that all the ignored features are not that dominating the prediction model. |
| 5 | Linear Regression | Only selected feature – income property | AUC – 0.529, C.A – 0.531, F1 – 0.368, Precision – 0.28, Recall – 0.531 | While I only selected one feature the accuracy dropped to 52 %, se we can say this one feature is highly dominating in the model. |
| 6 | Naïve Bayes | All features selected, Target Variable – Seasonal\_vaccine | AUC – 0.828, C.A – 0754, F1 – 0.754, Precision – 0.756, recall – 0.754 | With all the features (numerical and categorical) the accuracy achieved is 82 % |
| 7 | Naïve Bayes | Selected Features – behavioural\_avoidance,behavioural \_facc\_mask,behavioural\_wash\_hand,large gatherings,target\_seasonal\_vaccine | AUC – 0.57, C.A – 0.549, F1- 0.533, Precision – 0.545, recall – 0.549 | We can see the AUC decreases to 57% which means one of the features ignored is a dominant feature. |
| 8 | Naïve Bayes | Only numerical feature, target variable – seasonal vaccine | AUC – 0.811, C.A – 0.741, F1 – 0.741, Precision – 0.743, recall – 0.741 | N.A |
| 9 | Naïve Bayes | Only categorical feature, target variable – seasonal vaccine | AUC – 0.723, C.A – 0.667, F1 – 0.667, Precision – 0.666, recall – 0.667 | So, we can say only numerical are dominant and only categorical features are less dominant than numerical features |
| 10 | Naïve Byes | Features – doctor\_recc\_seasonal, behavioural large gatherings, opinion\_seas\_vacc , target – seasonal\_vaccine | AUC – 0.766, C.A – 0.712, F1 – 0.711, Precision – 0.719, recall – 0.712 | N.A |
| 11 | Neural Network | All features considered and target – seasonal\_vaccine | AUC – 0.777, C.A – 0.714, F1 – 0.715, Precision – 0.715, recall – 0.714 | With all the features (numerical and categorical) the accuracy achieved is 77 % |
| 12 | Neural Network | Only categorical feature, target variable – seasonal vaccine | AUC – 0.831, C.A – 0.761, F1 – 0.761, Precision – 0.761, recall – 0.761 | N.A |
| 13 | Neural Network | Only numerical feature, target variable – seasonal vaccine | AUC – 0.635, C.A – 0.601, F1 – 0.601, Precision – 0.601, recall – 0.601 | So, we can say only categorical are dominant and only numerical features are less dominant than categorical features |
| 14 | Neural Network | High\_risk\_cat,doctor\_recc\_seasonal,behavioural wash hand , Chronic\_med\_condition,household\_children | AUC – 0.758, C.A – 0698, F1 – 0.698, Precision – 0.698, recall – 0  .698 | With these features we can say that they are highly dominant and after ignoring all the features still we get around 75% of the accuracy. |
| 15 | Adaboost | Behave\_score , h1n1 concern - Highly Correlated | 0.763 |  |
| 16 | Adaboost | Behav\_to\_risk , opinion\_seas\_risk – Highly corelated | 0.0737 |  |
| 17 | Adaboost | Behave\_score,h1n1\_knowledge | 0.415 |  |
| 18 | Adaboost | Behave\_score , h1n1 concern | 0.763 | Most highly correlated , The AUC value with only most correlated features means these features are most dominant in the |
| 19 | Adaboost | Household\_adults,opinion\_seas\_rsik | 0.001 | Least corerelated |

* **Highly correlated functions are not good for prediction model best way is to get rid of one of them or use dimension reduction technique such as PCA.**