**Predicting Male Infertility Using KNN Classifier**

**Problem**

For several decades, male fertility rate is declining. As Male infertility accounts for 40% of all infertility cases, this downward trend is quite worrying. Low fertility rates are usually associated with diminished economic growth particularly due to a decline in workers per capita. This decrease in labor supply would especially be felt in countries where the elderly are already retired and receive social security from the government.

The factors that determine fertility rates are varied, ranging from [culture, cost of raising children, and urbanization. These factors all require long term solutions which will be quite costly. One factor that can be tackled straightforward though is medical infertility, as factors such as those previously mentioned affect a persons choice and not their ability to have children.

People who are infertile for medical reasons can be given treatment to help them become fertile. This task would be much easier for governments to do as compared to changing a country's culture or providing education subsidies. However, to be able to treat infertility we must be able to diagnose it.

Conducting a mass fertility test on an entire population can be quite costly, both in the collection phase as well as the testing phase. However, If we are able to gather information that has a high correlation with fertility then it would simply be a matter of databasing patient records and running algorithms to identify potential infertile subjects. The factors we look at in this notebook are environment, medical history, and lifestyle. These factors have been studied with [artificial intelligence and have shown promising results.

The problem would then be to see if there are any variables that are strongly correlated with infertility and if a Machine Learning algorithm would be able to accurately predict if a person is fertile or infertile.

**Highlights**

Three Highlights of the k-NN Classifier are:

1. Smoking has the highest sensitivity that can deteriorate fertility in men.

2. Historic medical complications affect fertility as high as smoking does.

3. Alcohol intake didn’t yield high sensitivity to influence male fertility.

Prolonged high fever, child diseases (mumps, measles, and others), surgical intervention.

**Data Acquisition and Cleaning**

The available data from <https://archive.ics.uci.edu/ml/datasets/Fertility> is cleaned already and ready to be as input in k-nearest neighbors algorithm.

**Implementation**

Custom Functions

1. knn\_dataset

2. plot\_accuracy

We can observe that at around 2-3 n\_neighbors, standard deviations of the training and test accuracy values begin to overlap signaling that the mean difference is approaching insignificance. At higher n\_values (> 10) the overlap becomes more consistent and prominent.

Looking closely, a plot for the first 10 n\_neighbers is shown below.





From these results, we chose k = 7 nearest neighbors for the KNN implementation below to show the exact test and train accuracy values for a random trial. However, it can be seen from the plots that any choice of k beyond that until k = 50 will still yield a desired accuracy above 80%.

From one random trial above, we can observe that the accuracies are already much higher than the PCC=0.7888. Replicating this result in more random trials at k = 7 will continue to yield accuracy values within the standard deviation shown in the plots above. These values are well above the target of 80% accuracy of the classifier.

Our group observed that the 1.25Pcc is unbeatable using the original dataset. Thus in the section \*\*KNN Classification with Oversampling below we employed other methods to try to beat the 1.25Pcc.

ensitivity Analysis

Approach A: Individual KNNs for each Feature

To determine the impact of each feature to our model, we conducted a sensitivity analysis by performing individual KNNs (one for each feature-target pair) and solved for the accuracy scores using the cross\_val\_score() funtion. The results are shown below.

****

Approach B: Accuracy dips

Another approach to this sensitivity analysis is by investigating on accuracy dips as we remove features on by one.

****

This sensitivity analysis can now be a basis on which features are least impactful as predictor in the classification. For example, in the case of our dataset, the feature 'season' showed a very low score on the sensitivity analysis thus functioning as a 'weaker' predictor.

Remarkable differences among accuracies of each feature can be seen in the Approach A are:

1. Sitting hours is shown to have a higher accuracy in predicting fertility in men than alcohol consumption

2. Smoking ranked as high as other medical conditions in terms of its accuracy in prediciting fertility in men.

**KNN Classification with Oversampling**

Our oversampling approach was effective in reducing the Pcc and 1.25Pcc values of the new dataset. However, we have reservations to the validity of our approach since the observed plot here no longer follows the behavior shown in the other plots above. Specifically, the cost of beating the 1.25Pcc is drastically lower accuracy (< 80%). The technique we employed thus requires further improvement and scrutiny.

****

**Discussion of Results**

From the sensitivity analysis, it can be seen that the following features are most correlated to the male infertility:

1. Smoking. Several studies have linked tobacco to be correlated with decreased sperm concentration, decreased motility, fewer normally shaped sperm, and increased sperm DNA damage. This is reflected on the model wherein the infertility is sensitive to smoking features of the dataset.

2. Childish Disease.Some childish diseases (such as mumps, chicken pox) can cause orchitis later in adulthood that are known to impair the sperm production in the testis. [healthline.com](https://www.healthline.com/health/orchitis)

3. High fevers. It is shown that the testicles is held further away from the body to regulate its temperature. However, in times when the testicles cannot maintain this ideal temperature (during high fevers), sperm production can stop or be hindered for a short time.

4. Surgical intervention. Surgical interventions on the reproductive parts have also shown to have high sensitivity on the sample dataset. However, there have been no definitive studies that might correlate surgeries with male infertility. This might necessitate to have further researches.

It was also found that the male infertility are less sensitive to the following features:

1. Age. While it is known that male can still bear children even at an older age, a greater age reduces the overall chances of pregnancy due to sperm quality decreases.

2. Number of hours sitting. Increased number of hours sitting have a higher risk of Oligospermia which is a male fertility issue characterized by a low sperm production. This is correlated to the increased temperature of the scrotum when sitting on extended period of hours.

3. Alcohol consumption. Temporary decreases in sperm counts and testosterone levels have been correlated to increased alcohol consumption. The good news is that the effects are reversible once alcohol consumption stops.

4. Accidental Trauma. Serious accidents and traumas (but non testicular trauma are used in the dataset) also have less sensitivity to male infertility. There are also no insightful papers that can confirm trauma as contributory to sperm production and should be further researched.

5. Season/ time of year. It was found that men with normal sperm production had the better sperm (in terms of swimming speed and quality) in the winter compared to other seasons of the year.

References:

1. Fertility Dataset; UCI Machine Learning Repository; https://archive.ics.uci.edu/ml/datasets/Fertility

2. "Semen Parameters can be Predicted from Environmental Factors and Lifestyle using Aritificial Intelligence Methods" by: Jose L. Girela1, David Gil2, Magnus Johnsson, María José Gomez-Torres, and Joaquín De Juan