# Programming for Data Analysis

Task 3

What physical symptoms are most primarily correlated to smokers?

By Vyshnavi Muthumula (st20219772)

## Contents

## List of figures

## Introduction

The main motive of the project is to find the effects of smoking. Recent days everyone is addicted to smoking because they think tobacco gives relaxation and reliefs from stress. Temporarily tobacco gives relaxation but later it gives so much health issues like sleep disorder, breath problems, heart related diseases, cancer etc.,. Tobacco is a mixture of nicotine, benzene, arsenic and formaldehyde because of this smoker can feel relaxation and stress free for some time. By using this study researcher wants to find out the consequences of smoking behaviour with the help of exploratory data analysis (EDA) and Machine learning models. The dataset used for this research contains 55,692 people data in this 20,455 are smokers and remaining are non-smokers. In this there are total 35,401 males and 20,291 females.

## Rationale Behind Choosing Models

The purpose of this study is used to find the effects of smoking behaviour by using EDA and machine learning models. Density plots are used to plot the categorical data in a distributive format. A box is used to indicate the upper and lower quartiles, and a line is used to represent the middle quartile, which represents the centre 50% of the distribution. By using all these 5 values the dataset description will be very easy to understand (citeseerx.ist.psu.edu, n.d.). The researcher used density, box and correlogram plots to define the effects of smoking behaviour under EDA techniques.

The question can also answer through machine learning models such as KNN, Random forest, Decision tree, Logistic regression and Support vector machine learning (SVM). Researcher performed 5 basic machine learning models and concluded the feature importance for the output with the best accuracy model i.e., SVM. Now a day's Machine learning models are used more because of the higher performance and for predicting future un-test data. One of the best supervised machine learning algorithms is support vector machine (SVM) and lazy algorithm is k-nearest-neighbour (K-NN) (Lim et al., 2020).

SVMs are effective for machine learning, neural networks, and binary classification tasks (Zeebaree, n.d.).Choosing hyper parameter values for SVM is a big challenge, for this research the hypermeters selected for better accuracy are kernel = rbf, C= 10 and gamma = 0.1. These hyper parameters are used to remove the over fitting/ under fitting from the noise and outliers. Reason for choosing kernel as Radial basis function as it is similar to K-NN algorithm and it reduces the space complexity issues. Here gamma represents the influence of single training data set and lower gamma will lead better choice for SVM. The 'C'is the regularization or optimization parameter which is used to control the trade off point (Friedrichs and Igel, 2005).

Decision trees are able to discover missing data in a data set, extract text from a data set, and in the healthcare profession to replace statistical methods. (Myles et al., 2004).

Logistic regression is used to find the probability of the output variable from the data set (Stoltzfus, 2011).

In order to obtain a high level of predicted accuracy, the Random Forest methodology is used and it is a regression tree method which aggregates bootstrap data and randomises predictors (Rigatti, 2017). Smoking increases the levels of serum creatinine, cholesterol, triglycerides, and low-density lipoprotein (LDL), while decreasing the levels of antiatherogenic high-density lipoprotein (Joshi et al., 2013). Changes in artery wall damage, coagulation status, and cholesterol and lipoprotein levels are only a few examples of how it affects physiological parameters.

# Challenges:

There are some columns which are not helpful to answer the question, so unwanted columns are dropped (Example: ID). As this dataset contains large amount of data, visualization and predicting the output is difficult but researcher answered the question perfectly with machine learning models by taking 1000 random samples. For SVM selecting kernel is difficult but as this is a non-linear problem researcher selected Radial basis function as kernel (Norkin and Keyzer, 2009). However, since Random forest and KNN have fluctuating accuracies but SVM always gives better accuracy. The performance of the SVM depends on hyper parameters. Gamma and C values play very important role in this model. The gamma value influences the performance, if this value is over estimated then the kernel performance will be exponential change to linear(Al-Mejibli, Alwan and Abd, 2020). These gamma and c are the optimising parameters and there are no such rules to choose these values. By using hard test, researcher needs to find the best suited values for better accuracy (Duan, K.B. and Keerthi, S.S., 2005).After multiple iterations it is found that the gamma value 0.1 and 'c' value 10 is providing better accuracy. By performing multiple tests on polynomial, sigmoid and rbf it is proven that rbf is best suited kernel.

## Methods Effectiveness

The methods to conclude on answering the question will derive from analysis on the EDA and then the conduction towards accuracy analysis for my 5 chosen models as to conclude the most suitable algorithm to choose for the further analysis towards proving the most problematic and correlated symptom linked to smokers. Once the most efficient model has been discovered then analysis towards concluding the most correlated feature will include the conduction of feature importance calculations and analysis the accuracy of each feature individually with the output.

**EDA results:**

Correlogram results saying that Hemoglobin, relaxation and serum creatinine levels are highly correlated to smoking as mentioned in .ipynb file.

**Hemoglobin levels:**

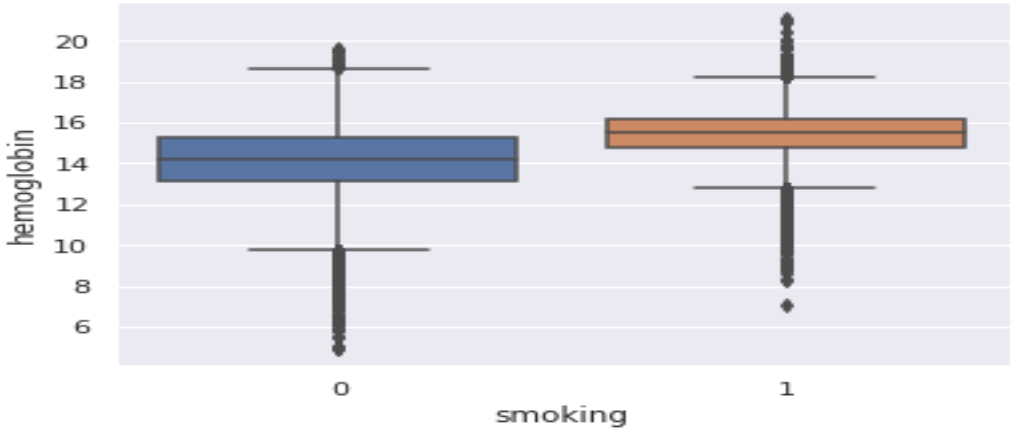

Figure 1: Hemoglobin levels

**Relaxation levels:**

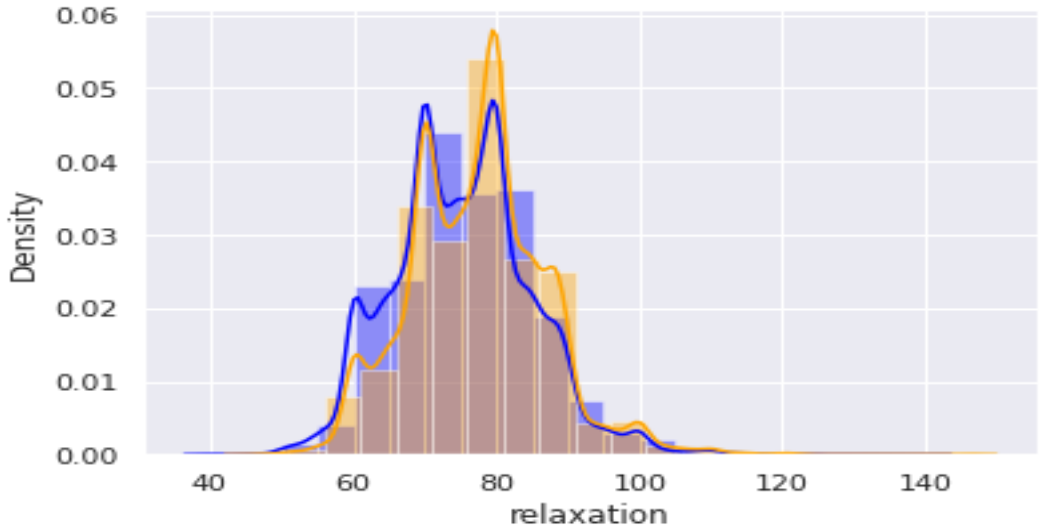


Figure 2: Relaxation levels

**Serum creatinine level:**

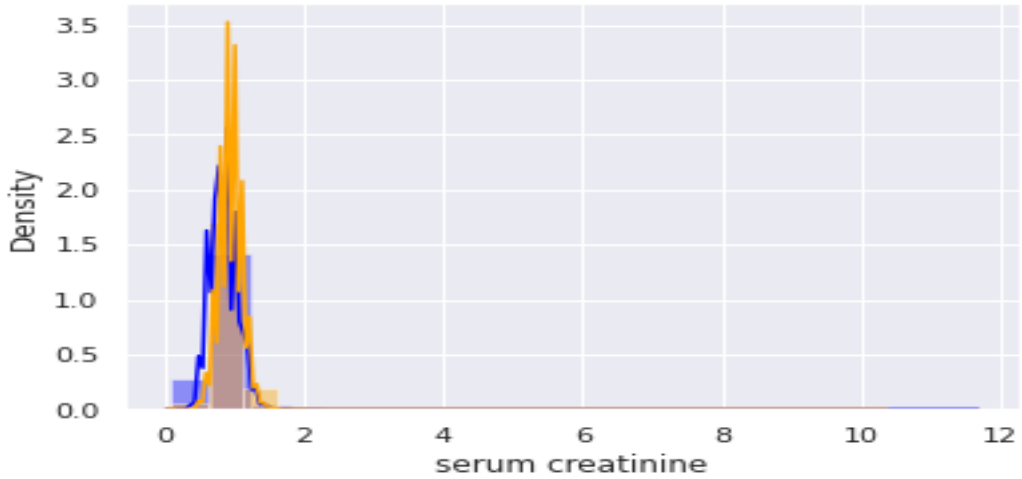


Figure 3: Serum creatinine

The main purpose of machine learning in this study is to predict the accuracy of the smokers based on few features such as systolic, serum creatinine, fasting blood sugar, triglyceride, LDL, Urine protein, serum creatinine, hemoglobin, AST, LDL, relaxation, Gtp.
However, even though SVM performed the best with 93.33% accuracy than KNN[92.15%] and random forests[91.52%] both performed surprising well without the hassle of tinkering and analysing of so many hyper parameters that is issued with SVM. Therefore, the methods effectiveness of SVM may be the best in this example but due to speed than KNN or RF can both perform almost as efficiently without the negatives/barriers of calculation speed/cost seen with SVM. However, for the purpose of the study then SVM will be utilised.

```
[237] best_model = pd.DataFrame(best_acc, columns = ["Model", "Accuracy"])
      best_model
```

| | Model | Accuracy |
|---|---|---|
| 0 | Logistic Regression | 81.212121 |
| 1 | Decision Tree | 80.303030 |
| 2 | Random Forest | 91.515152 |
| 3 | KNN | 92.424242 |
| 4 | SVM | 93.333333 |

Figure 4: All Models

**Feature importance by using SVM model**

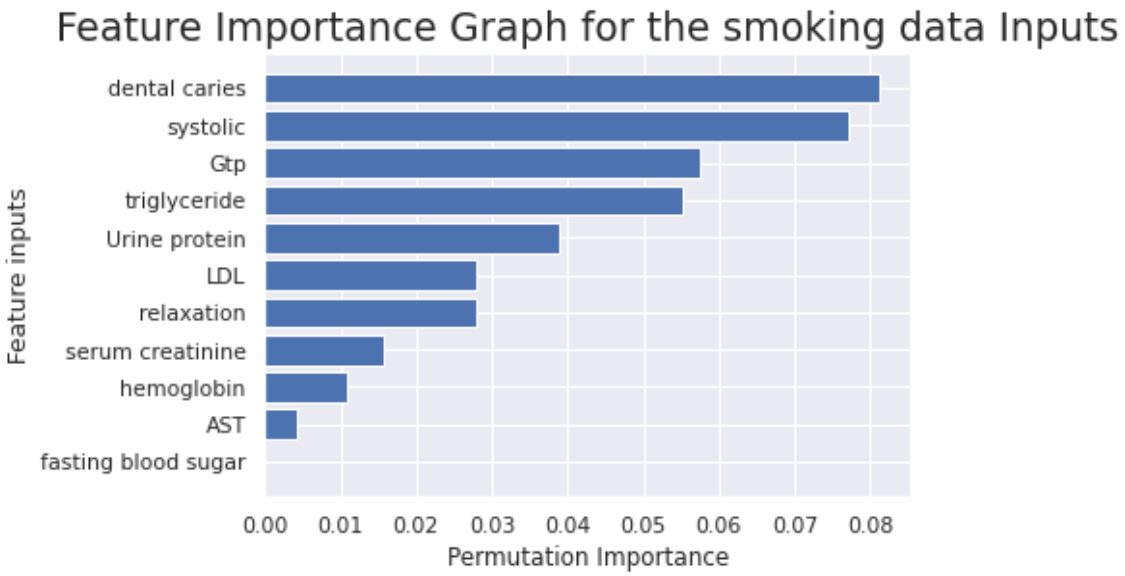


Figure 5: Feature importance

The above graph shows that dental caries, systolic, Gtp and triglyceride are more affected by smoking.

**Accuracy for the individual inputs for SVM model:**

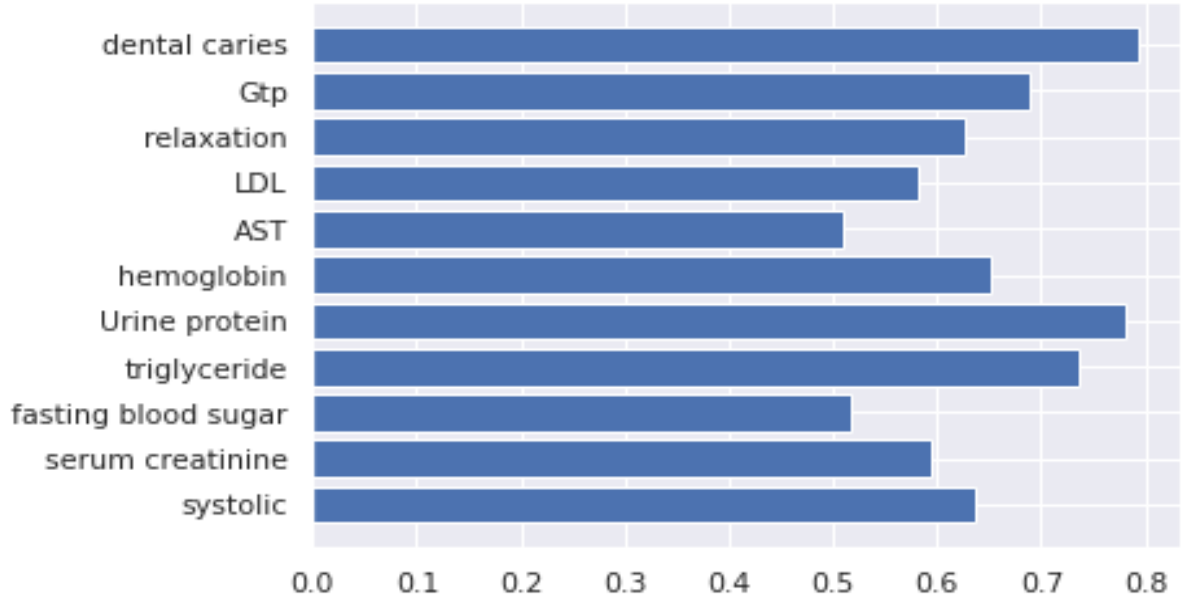


Figure 6: Individual accuracies

Dental caries is directly correlated to smoking as it has highest accuracy and low correlated value is AST.

## Methods Effectiveness Conclusion and Lessons Learnt:

From the above case study these are the physical factors Triglyceride, Relaxation, Serum creatinine, Fasting blood sugar, Urine protein, haemoglobin, LDL and Systolic that are affected by smoking. By using EDA it is proven that consumption of smoking causes several health issues. Heavy smoking causes heart stroke and cancer as the above input parameters are directly correlated to smoking. The correlogram results shown that haemoglobin levels are highly correlated to smoking. When they consume tobacco (which contains nicotine) their anxiety levels and oxygen supply increases so haemoglobin also increases (Shah et al., 2013).

The problems occur when human have high haemoglobin are Dehydration, Heart failure, Kidney cancer, Liver cancer and Emphysema (Sari, Maharani and Damayanti, 2021).
Weight is also positively correlated to smoking parameter. Smoker's weight is less than non-smokers because their appetite is decreased. But smokers will gain weight when they stop smoking it's because of nicotine present in the body which increases appetite (Aubin et al., 2012). Smokers are getting more relaxation than non-smokers after consumption of tobacco it's because of the ingredients present in tobacco i.e. nicotine which relaxes the body (Rambali, Vleeming and Opperhuizen, 2002).

Machine learning algorithms are performed to find the feature importance as EDA always not gives correct predictions in future. Out of five algorithms it is proven that the Support vector Machine model is the best algorithm by comparing accuracy with other algorithms. The accuracy for the SVM model is 93.3% which is very good accuracy with this accuracy we can say in future the predictions can be done perfectly (Gold and Sollich, 2003).

By using feature importance for Support vector machine Dental caries, systolic, Gtp and triglyceride are more affected top 4 factors for smokers. The least correlated values are AST and fasting blood sugar.
Smoking cigarette leads to increase in the concentration of serum creatine, cholesterol, triglyceride, low density lipoprotein (LDL) and fall in levels of antiatherogenic high density lipoprotein (Joshi et al., 2013). It results in altered physiological parameters such as changes in arterial wall damage, coagulation status, cholesterol and lipoprotein content.
In researcher's point of view since dental has the highest individual accuracy via the feature importance calculations then dental problems are the highest correlated factors for smokers.
After research it is found that feature importance values will be changed based on the chosen models. And the chances for re order of the values also changes sometimes because of stochastic nature.

## Recommendation:

Since Dental caries and systolic has a higher correlation to smoking as shows that smokers have a higher chance of mouth cancer and heart stroke then the next step towards maintaining individual's health is to create additional algorithms to discuss if smoking would result in a higher possibility of dental caries as (Benedetti et al., 2012)
Triglyceride also has a higher coloration to smoking as shows that smokers have a higher chance of heart attacks (Savdie, Grosslight and Adena, 1984). Higher blood pressure could result in a higher chance of strokes (Pocock et al., 2001).

# References


Emerson, J.W., Green, W.A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H. and Wickham, H. (2012). The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics*, 22(1), pp.79–91. doi:10.1080/10618600.2012.694762.

Psu.edu. (2019). *CiteSeerX — Unknown file type*. [online] Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.66&rep=rep1&type=pdf. [Accessed 16 Aug. 2022].

Abdulqader, D.M., Abdulazeez, A.M. and Zeebaree, D.Q., 2020. Machine learning supervised algorithms of gene selection: A review. Machine Learning, 62(03), pp.233-244.

T. Horvat and J. Job, "Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods," Elektroteh. Vestn., vol. 86, no. 4, pp. 197–202, 2019.

Friedrichs, F. and Igel, C., 2005.Evolutionary tuning of multiple SVM parameters.*Neurocomputing*, *64*, pp.107-117.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D., 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), pp.275-285.

Stoltzfus, J.C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, 18(10), pp.1099–1104. doi:10.1111/j.1553-2712.2011.01185.x.

Rigatti, S.J., 2017. Random forest.*Journal of Insurance Medicine*, *47*(1), pp.31-39.

Joshi, N., Shah, C., Mehta, H.B. and Gokhle, P.A., 2013. Comparative study of lipid profile on healthy smoker and non smokers.*International Journal of Medical Science and Public Health*, *2*(3), pp.622-626

Allen, A.M., Oncken, C. and Hatsukami, D. (2014). Women and Smoking: The Effect of Gender on the Epidemiology, Health Effects, and Cessation of Smoking. *Current Addiction Reports*, 1(1), pp.53–60. doi:10.1007/s40429-013-0003-6.

Chiolero, A., Faeh, D., Paccaud, F. and Cornuz, J. (2008). Consequences of smoking for body weight, body fat distribution, and insulin resistance. *The American Journal of Clinical Nutrition*, 87(4), pp.801–809. doi:10.1093/ajcn/87.4.801.

Murty, M.N. and Raghava, R., 2016. Kernel-based SVM. In *Support vector machines and perceptrons* (pp. 57-67). Springer, Cham.

Norkin, V.I. and Keyzer, M.A. (2009). Asymptotic efficiency of kernel support vector machines (SVM). *Cybernetics and Systems Analysis*, 45(4), pp.575–588. doi:10.1007/s10559-009-9125-1.

Al-Mejibli, I.S., Alwan, J.K. and AbdDhafar, H., 2020.The effect of gamma value on support vector machine performance with different kernels.*International Journal of Electrical and Computer Engineering*, *10*(5), p.5497.

Duan, K.B. and Keerthi, S.S., 2005, June. Which is the best multiclass SVM method? An empirical study.In *International workshop on multiple classifier systems* (pp. 278-285).Springer, Berlin, Heidelberg.

psycnet.apa.org. (n.d.). *APA PsycNet*. [online] Available at: https://psycnet.apa.org/record/1979-30528-001 [Accessed 16 Aug. 2022].

Shah, B., Nepal, A., Agrawal, M. and Sinha, A. (2013). The effects of cigarette smoking on hemoglobin levels compared between smokers and non-smokers. *Sunsari Technical College Journal*, 1(1), pp.42–44. doi:10.3126/stcj.v1i1.7985.

Nordenberg, D., Yip, R. and Binkin, N.J., 1990.The effect of cigarette smoking on hemoglobin levels and anemia screening.*Jama*, *264*(12), pp.1556-1559.

Aubin, H.-J. ., Farley, A., Lycett, D., Lahmek, P. and Aveyard, P. (2012). Weight gain in smokers after quitting cigarettes: meta-analysis. *BMJ*, [online] 345(jul10 2), pp.e4439–e4439. doi:10.1136/bmj.e4439.

Rambali, B., Vleeming, W. and Opperhuizen, A. (2002). The role of nitric oxide in cigarette smoking and nicotine addiction. *Nicotine & Tobacco Research*, 4(3), pp.341–348. doi:10.1080/14622200210142724.

Gold, C. and Sollich, P. (2003). Model selection for support vector machine classification. *Neurocomputing*, 55(1-2), pp.221–249. doi:10.1016/s0925-2312(03)00375-8.

Joshi, N., Shah, C., Mehta, H. and Gokhle, P. (2013). Comparative study of lipid profiel on healthy smoker and non smokers. *International Journal of Medical Science and Public Health*, 2(3), p.622. doi:10.5455/ijmsph.2013.210420131.

Benedetti, G., Campus, G., Strohmenger, L. and Lingström, P. (2012). Tobacco and dental caries: A systematic review. *Acta Odontologica Scandinavica*, 71(3-4), pp.363–371. doi:10.3109/00016357.2012.734409.

Savdie, E., Grosslight, G.M. and Adena, M.A. (1984). Relation of alcohol and cigarette consumption to blood pressure and serum creatinine levels. *Journal of Chronic Diseases*, [online] 37(8), pp.617–623. doi:10.1016/0021-9681(84)90111-5.

Pocock, S.J., McCormack, V., Gueyffier, F., Boutitie, F., Fagard, R.H. and Boissel, J.-P. . (2001). A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. *BMJ*, 323(7304), pp.75–81. doi:10.1136/bmj.323.7304.75.