Deliverables

1. Lay out an approach plan, consisting of:

a. Your understanding of data, based on a preliminary exploratory analysis

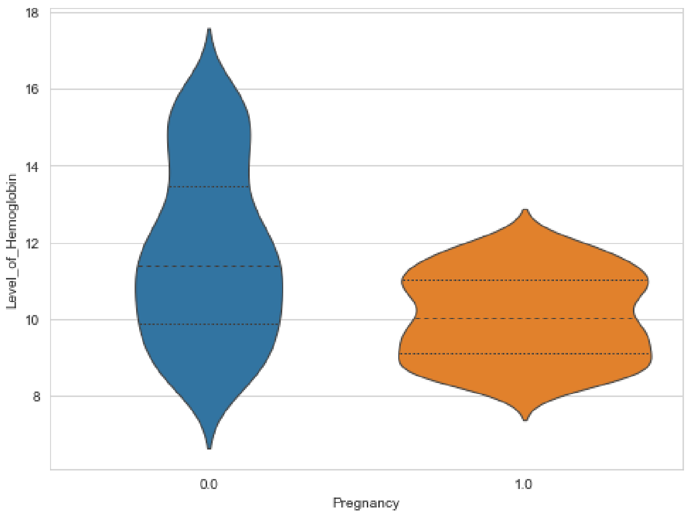
b. Different traditional as well as state-of-the-art statistical/Machine Learning techniques, which you are going to use to come up with different models to meet the objective

EDA - Exploratory data analysis:

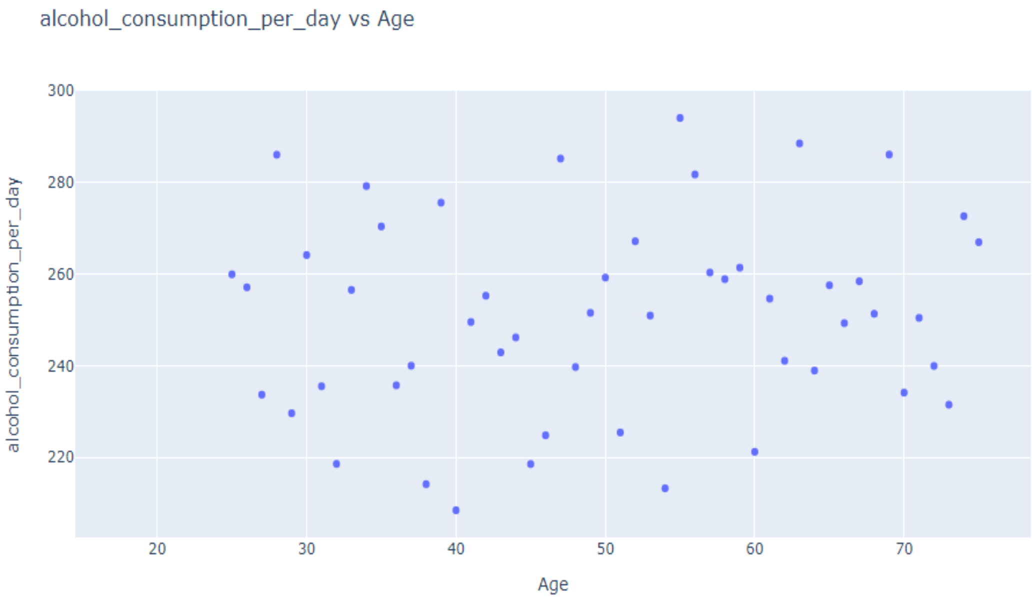
1. Understanding which variables are continuous and non-continuous.

2. Describing the data to understand various statistical measures, such as mean, standard deviation, max, etc. This concluded that various features continuous features are required to be scaled.

3. Checking for null values in that data. The following variables have null values and were imputed in the following ways:
   a. Imputing null values for Pregnancy:
      - The graph below shows the density of points at different values of Level_of_Hemoglobin for those who are pregnant and those who aren't, we call it the distribution.
      - We can see that for those who are pregnant the large majority had Level_of_Hemoglobin of around 10. But then for those who aren't pregnant can see that the distribution of Level_of_Hemoglobin is much more spread out, but the median is higher.
      - If the woman is pregnant around 60% of them don't smoke, else 44% of them don't smoke.
      - The above finding shows that if the woman is pregnant there is less chance of her smoking. Moreover, we can utilize Level_of_Hemoglobin and Smoking variables to impute for pregnancy missing values in women, and for men we can replace missing values with another class '2'.



b. Imputation of null values for alcohol_consumption_per_day



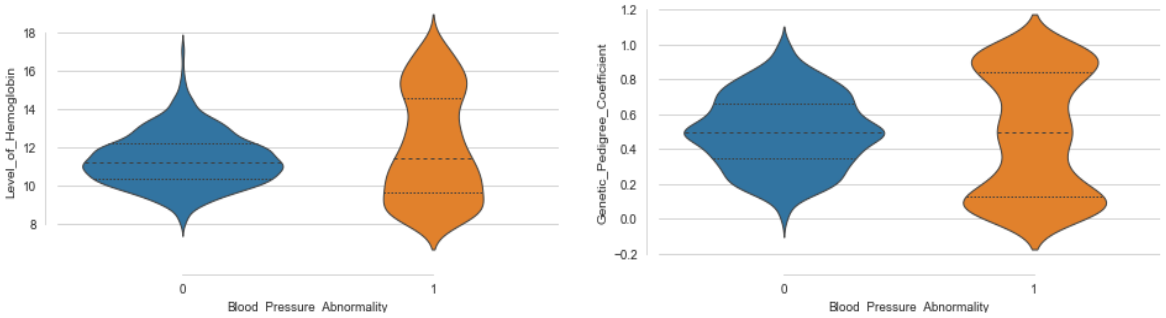
- For ages below 25 the alcohol_consumption_per_day is null, and it's safe to assume that legal drinking age for the region is 25. Hence, imputing alcohol_consumption_per_day as 0 for ages less than 25.

c. Imputation of null values for Genetic_Pedigree_Coefficient:

- As the missing values for Genetic_Pedigree_Coefficient are less than 5%, we can take mean to impute missing values.

4. Checking correlation among the variables. None of the variables are highly correlated except Chronic_kidney_disease and Adrenal_and_thyroid_disorders with target variable.

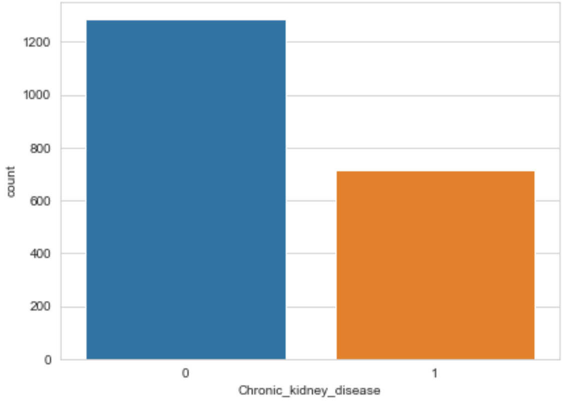
5. EDA for continuous variables:

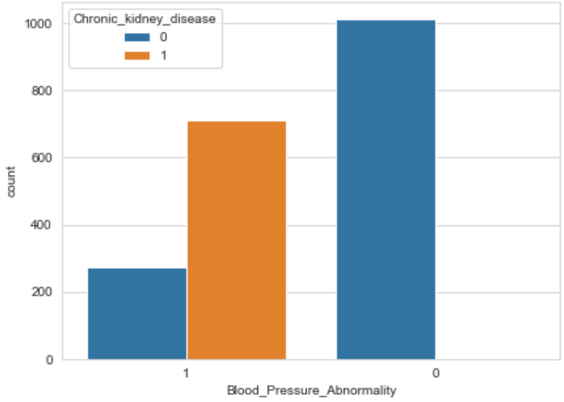


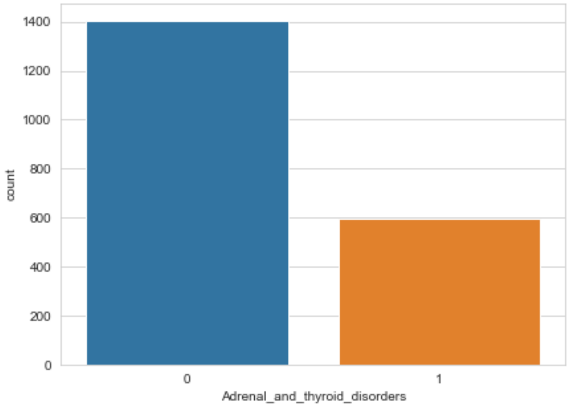
- A general conclusion to make from all of this is that it seems the people with Blood_Pressure_Abnormality had much more variation in their Level_of_hemoglobin and Genetic_Pedigree_Coefficient than for those who didn't.

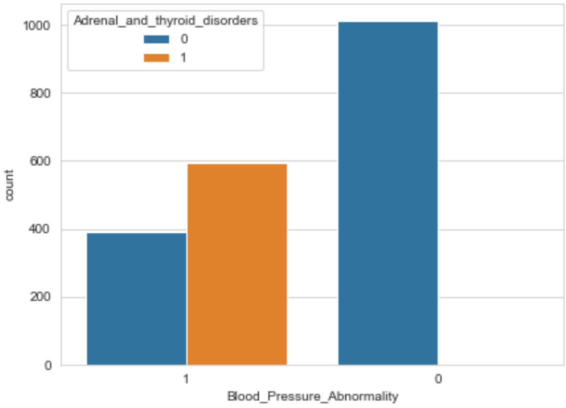
- Level_of_hemoglobin data points are skewed and much more varied when blood pressure abnormality is there.

6. EDA for non-continuous variables:









- Out of all the categorical variables only Chronic_kidney_disease and Adrenal_and_thyroid_disorders seem to have direct significance with target variable. People with Chronic_kidney_disease and Adrenal_and_thyroid_disorders have higher chances of having Blood_Pressure_Abnormality.

7. Doing one hot encoding for Pregnancy variable.

8. Splitting the data into test and train test.

9. Standardising the numerical variables.

Choosing the best model:

1. Logistic regression:
   - Benchmark model for this classification problem is taken as Logistic regression.
   - Calculation of optimal inverse of lambda to give best AUC value.

- Best Inverse lambda value calculated is 0.001(C).
- Possibility of some the variables being linearly separable.
- Only one nominal variable is there pregnancy (after transformation).

2. XGBoost model:
   - Based on EDA it suggests that features like are Level_of_hemoglobin and Genetic_Pedigree_Coefficient are skewed, not linearly separable and contains some outliers as well.
   - Good number of categorical variables are there.
   - Features having null values.
   - Best parameters selected are, number of trees as 300 and learning rate of 1.

2. Contrast the pros and cons of applying each technique on this problem

Logistic Regression:

- Works well with linearly separable variables.
- Doesn’t deal with data having null values.
- Doesn’t work well with skewed data.
- Can’t deal with nominal categorical data.
- Doesn’t deal with outliers well.
- More efficient for large datasets.

XGBoost Model:

- Doesn’t deal well with linearly separable variables.
- Deals well with data having null values.
- Deals well with skewed data.
- Deal well with nominal categorical data.
- Deals deal with outliers well.
- Less efficient for large datasets.

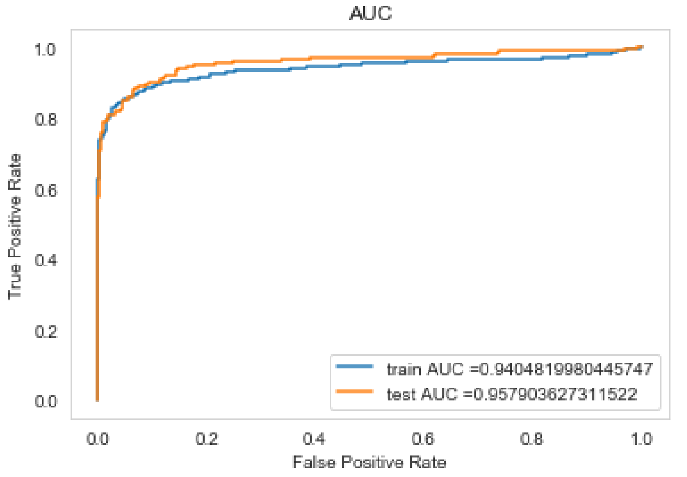
Conclusion – XGBoost is a better choice of model as its cons align with the given dataset based on our EDA results.

3. Build model(s) using the most promising technique on the dataset

- Model were built using above mentioned techniques, but as per assessment the champion model is XGBoost for the given dataset.

4. Validate the model(s) with the appropriate technique(s)

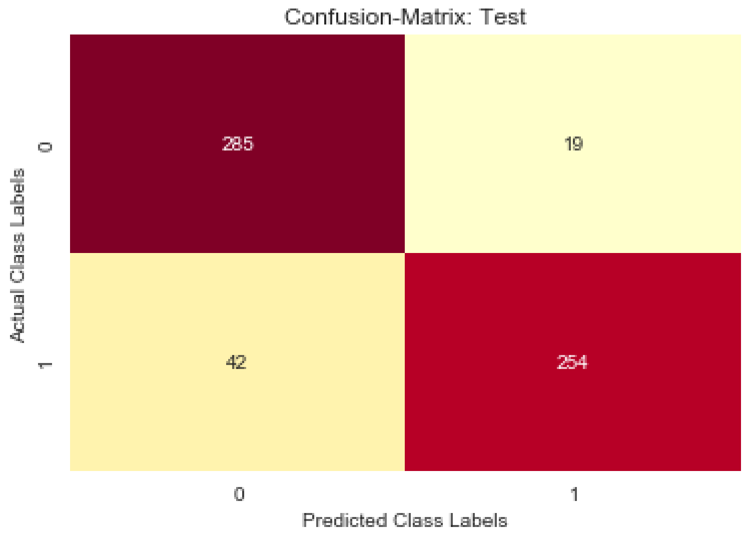
- **Logistic Regression:**
  - **AUC Curve**



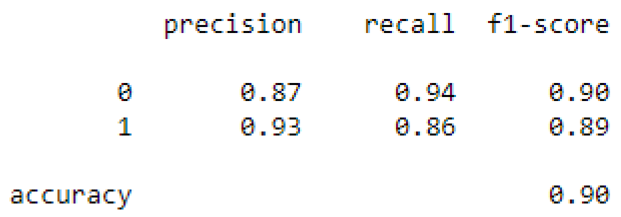
```
Logistic Regression - Test AUC:  0.957903627311522
```

- **Confusion Matrix**

```
Confusion matrix: Test data
```

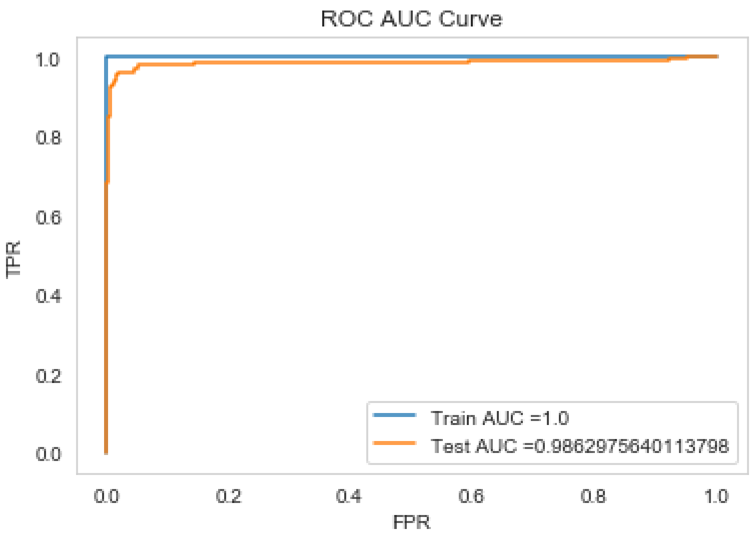


- **Other Validation Factors**



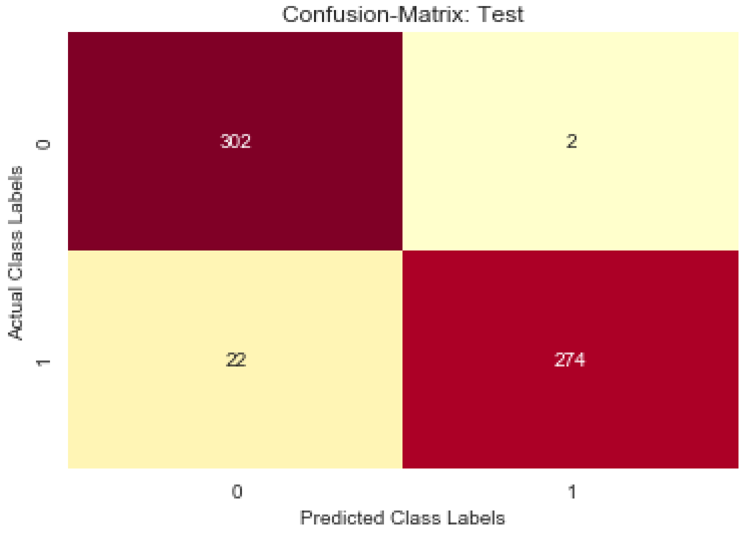
```
              precision    recall  f1-score

           0       0.87      0.94      0.90
           1       0.93      0.86      0.89

    accuracy                           0.90
```

- **XGBoost:**
  - **AUC Curve**



- **Confusion Matrix**

```
Confusion matrix: Test data
```



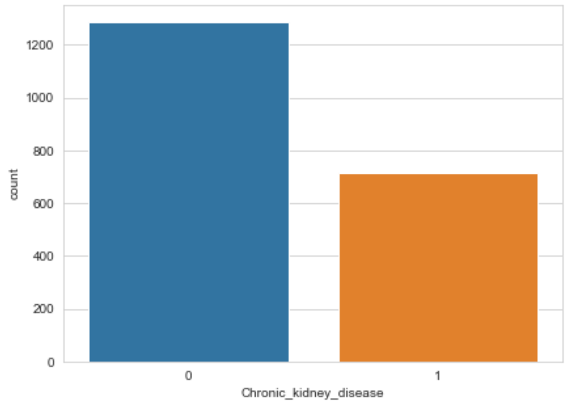
- **Other Validation Factors**

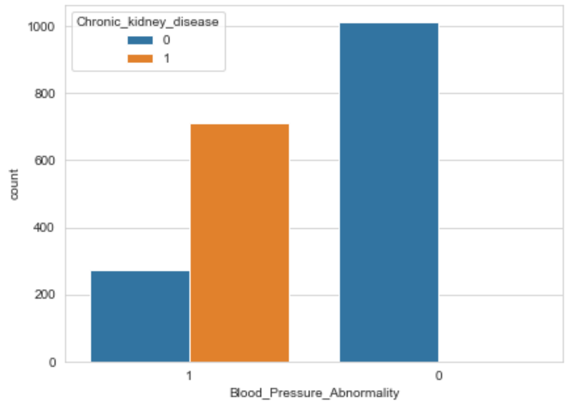
```
              precision    recall  f1-score

           0       0.93      0.99      0.96
           1       0.99      0.93      0.96

    accuracy                           0.96
   macro avg       0.96      0.96      0.96
```

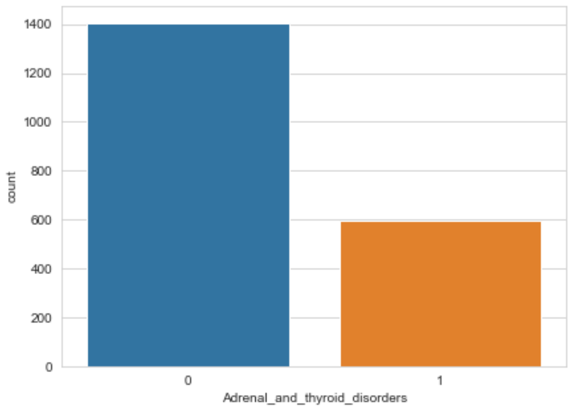
Conclusion – Based on the above validation parameters XGBoost out performs logistic regression.

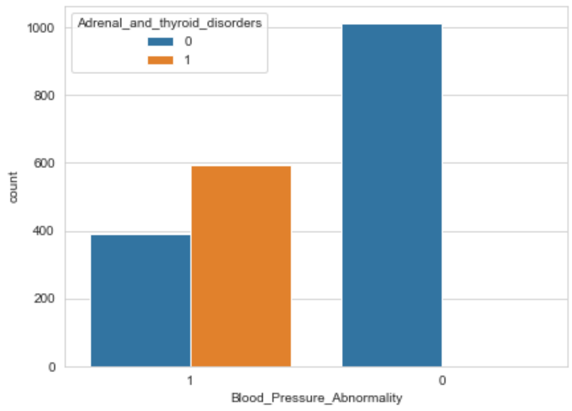
5. What would be your approach, if there were other variables also in the data:

Smoking, obesity (BMI), Lack of physical activity, salt content in the diet, alcohol consumption per day, Level of Stress, Age, Sex, Pregnancy, Chronic kidney disease and Adrenal & thyroid disorders.

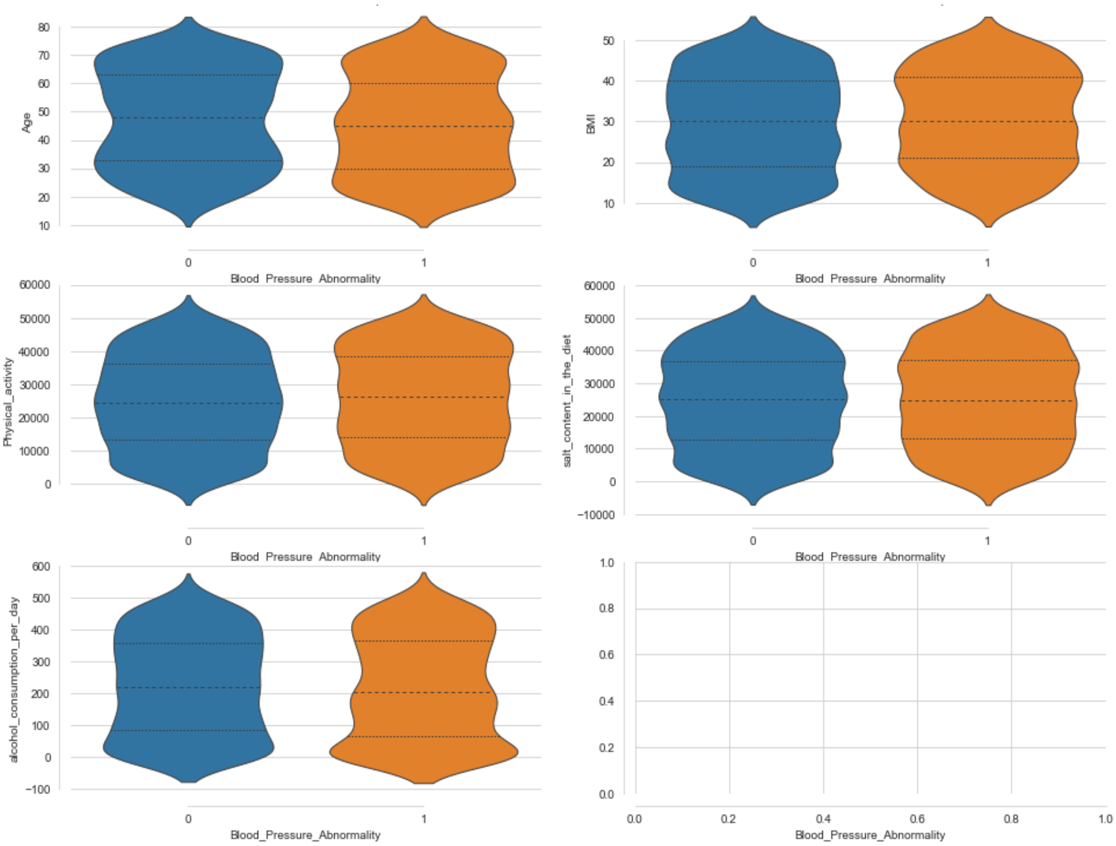








- Out of all the categorical variables only Chronic_kidney_disease and Adrenal_and_thyroid_disorders seem to have direct significance with target variable. People with Chronic_kidney_disease and Adrenal_and_thyroid_disorders have higher chances of having Blood_Pressure_Abnormality.



- For categorical features no significance conclusions can be made based on violin plots with target variables. But we can we can check for more trends and patters among the features. In EDA it was concluded that pregnancy is dependent on Smoking, Level_of_Hemoglobin and Sex, and these features can be utilised to fill null values.