Birla Institute of Technology & Science, Pilani

Work-Integrated Learning Program Division

First Semester 2020-2021

M.Tech (Data Science and Engineering)

Assignment

Course No.: DSECF ZC415

Group: G352

Contribution Table: Group: G352

Sl. No.	Name (as appears in Canvas)	ID NO	Contribution
1	YEDDI NAVEEN KUMAR	2020FC04926	100%
2	SRINIVASARAO LADI	2020FC04927	100%
3	IMMADISETTY AKHILA	2020FC04925	100%

Problem statement

- Identify the factors causing chronic kidney disease.
- Build a model that can help to determine if a patient is suffering from kidney chronic disease or not.

Data Understanding

- The data is gathered for two months from patients at a hospital. You need to make utilization of the features presented in the data set for your task. The data set and a document containing the information about the attributes are attached with the assignment problem statement.
- Make yourself familiar with these attributes as these might help you in determining the patients with kidney chronic disease

Dataset columns

There are 25 columns in the data set 24 + class(predict)

11 are Numeric variables:

- Age
- Blood Pressure(bp)
- Blood Glucose Random(bgr)
- Blood Urea(bu)
- Serum Creatinine(sc)
- Sodium(sod)
- Potassium(pot)
- Hemoglobin(hemo)
- Packed Cell Volume(pcv)
- White Blood Cell Count(wbcc)
- Red Blood Cell Count(rbcc)

5 Yes/No columns:

- Hypertension (htn)
- Diabetes Mellitus(dm)
- Coronary Artery Disease(cad)
- Pedal Edema(pe)
- Anemia(ane)

Dataset columns cont..

- 9 Categorical columns:
  - Specific Gravity(sg)
  - Albumin(al)
  - Sugar(sc)
  - Red Blood Cells(rbc)
  - Pus Cell(pc)
  - Pus Cell clumps(pcc)
  - Bacteria(ba)
  - Appetite(appet)

- Class: this variable is predict (Y) variable contains ‘ckd’, ‘notckd’
  - ckd: person contains Chronic Kidney Disease
  - notckd: person does not have Chronic Kidney Disease

Data Cleaning and Manipulation

- Identified some ‘?’ in dataset and replace ‘?’ with null values
- “Red Blood Cells(rbc)” column have more no. of null values around 38% , hence dropped the “Red Blood Cells” column
- Dropped null values rows for categorical variables.
- For numerical variables replace the null values with “mean”(bgr, bu,hemo,pcv,wbcc,rbcc) and “mode”(bp, sc, sod,pot) values based on the distribution plots.

Data Cleaning and Manipulation cont…

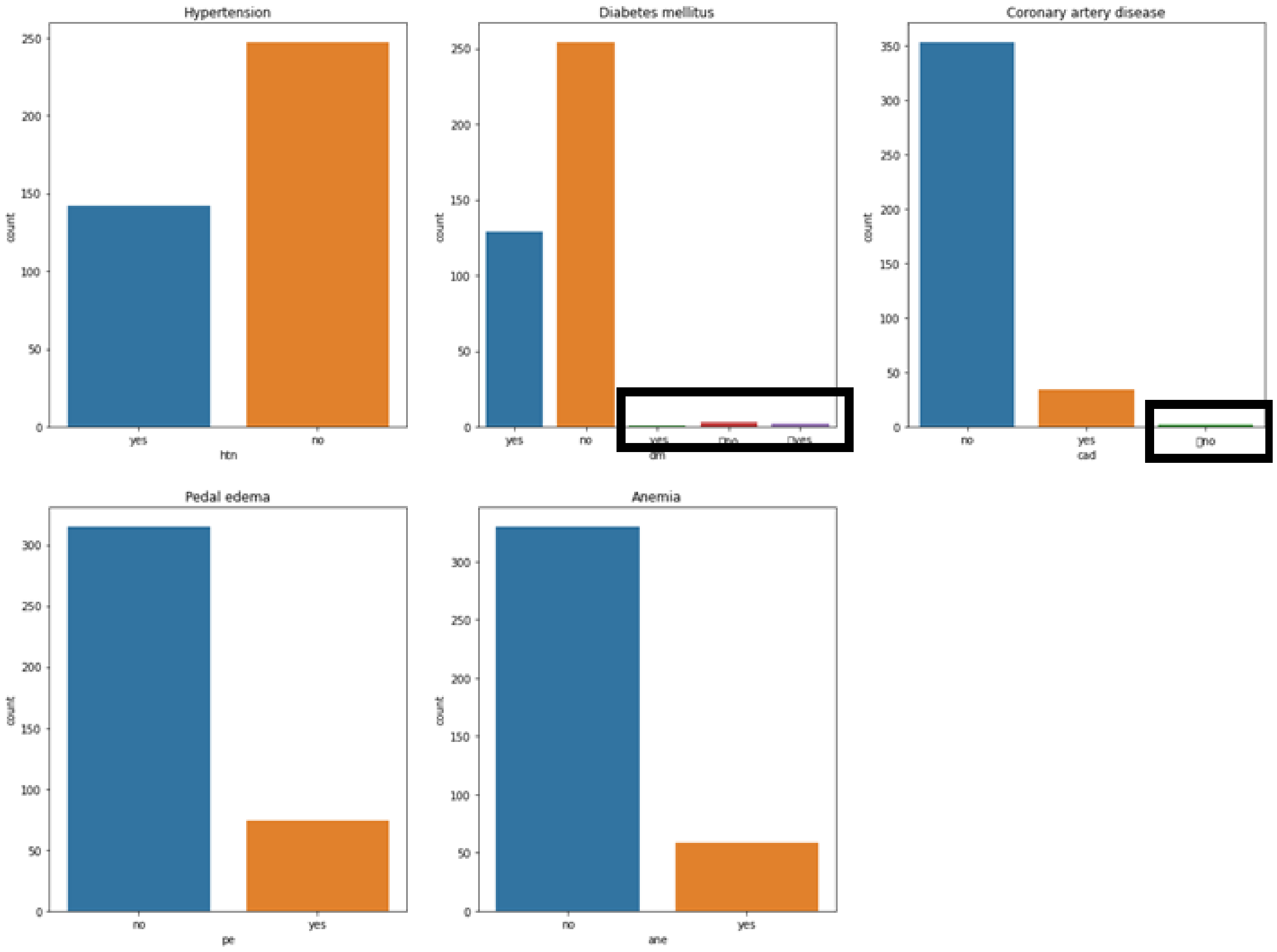
- Converted ‘yes’-> 1 and ‘No’ -> 0 for below columns
  - Hypertension (htn)
  - Diabetes Mellitus(dm)
  - Coronary Artery Disease(cad)
  - Pedal Edema(pe)
  - Anemia(ane)
- Predicted variable(class) and other categorical variables convert to numeric variables using pandas ‘**pd.get_dummies**’ method and dropping the first column

Dummy variables

class_ckd	class_notckd
1	0
1	0
1	0
1	0
1	0

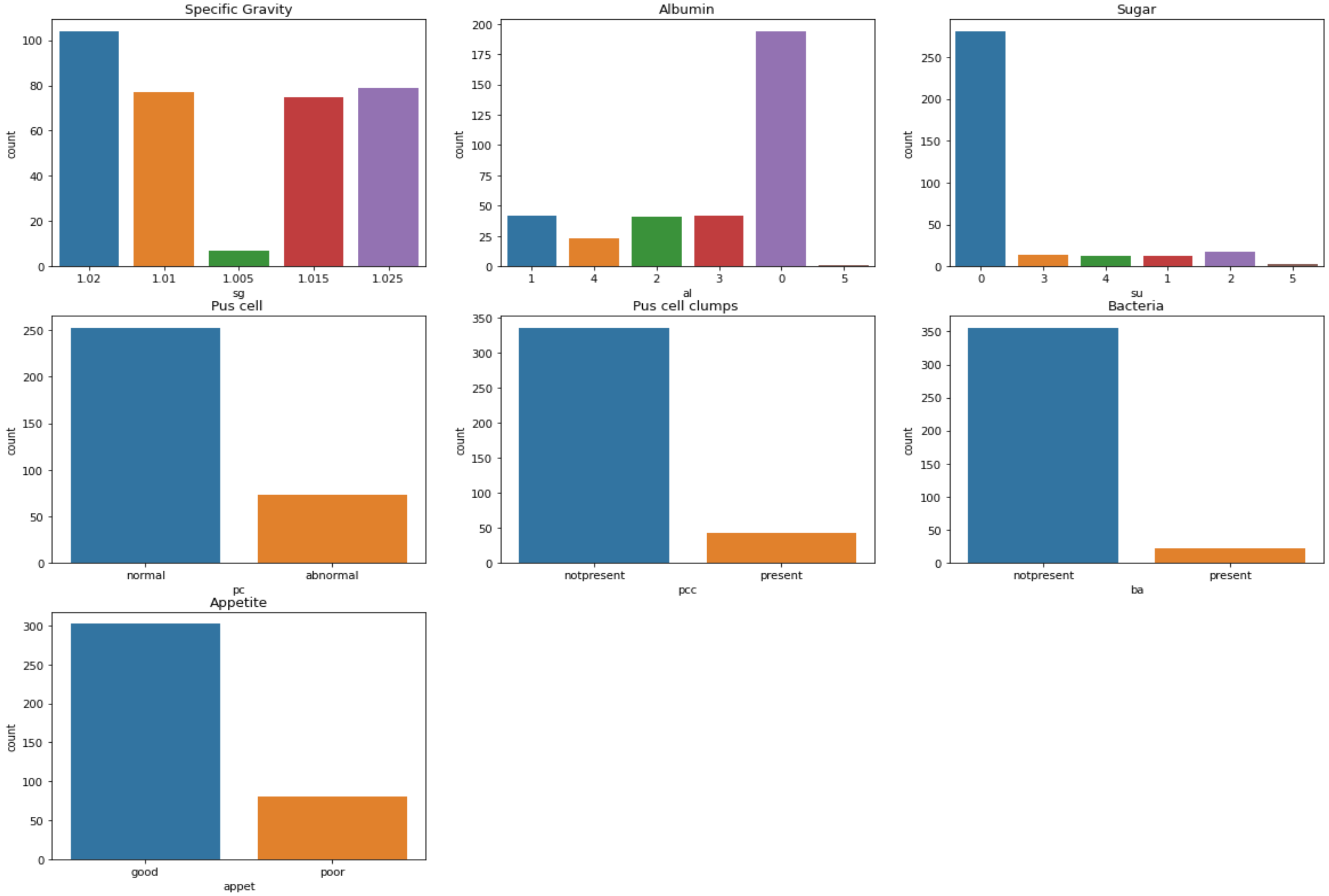
- Converted class data to dummy variables
- It has created two columns class_ckd, and class_notckd
- Instead of using both features one can be enough to use, as for the dummy variable condition n-1 features need to be considered.
- Same condition applied for all other categorical variables

Yes-No columns

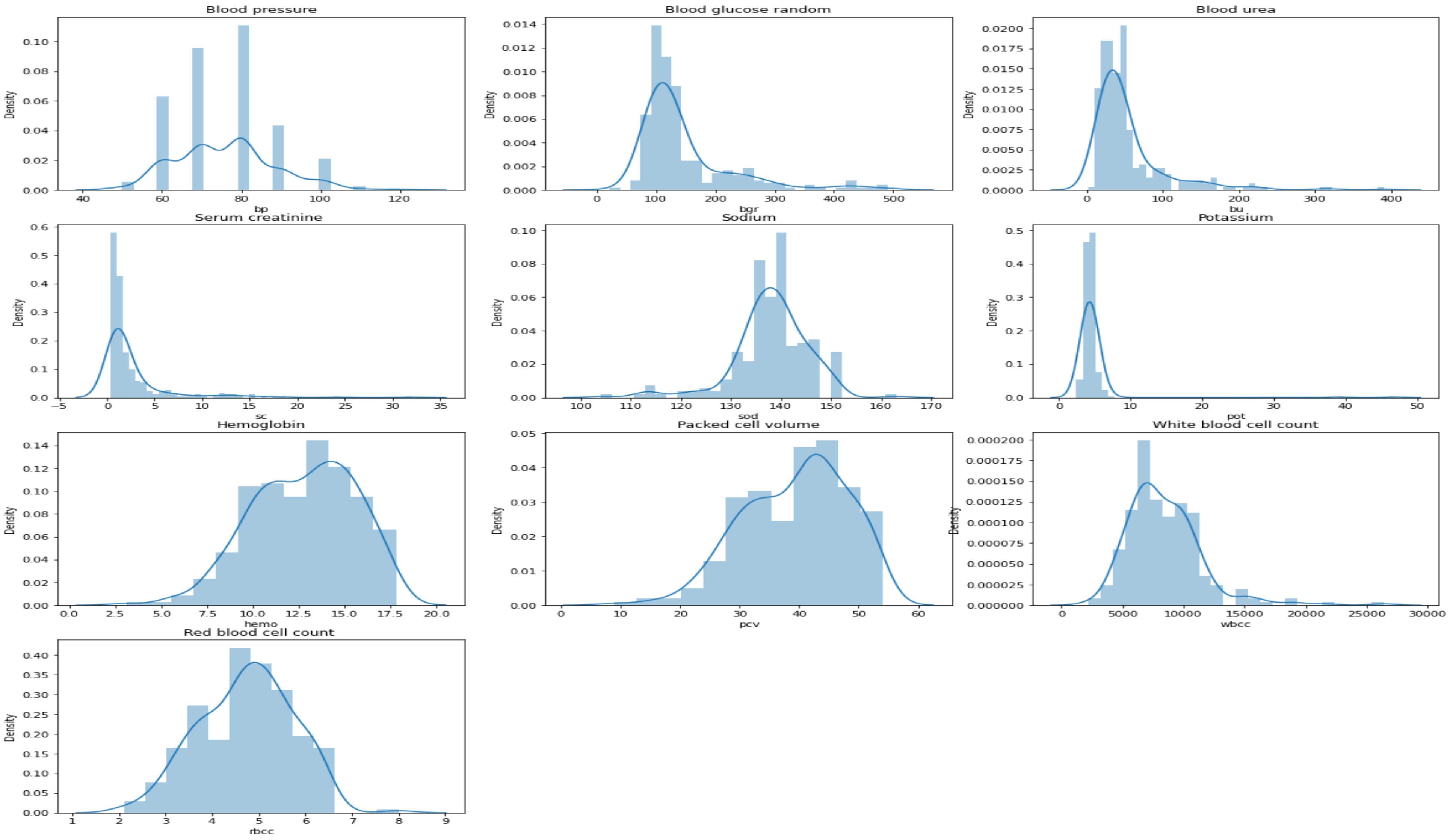


- Observed there are some data which is not matched ‘yes’ and ‘no’ text.
- Those data also removed form the data set.

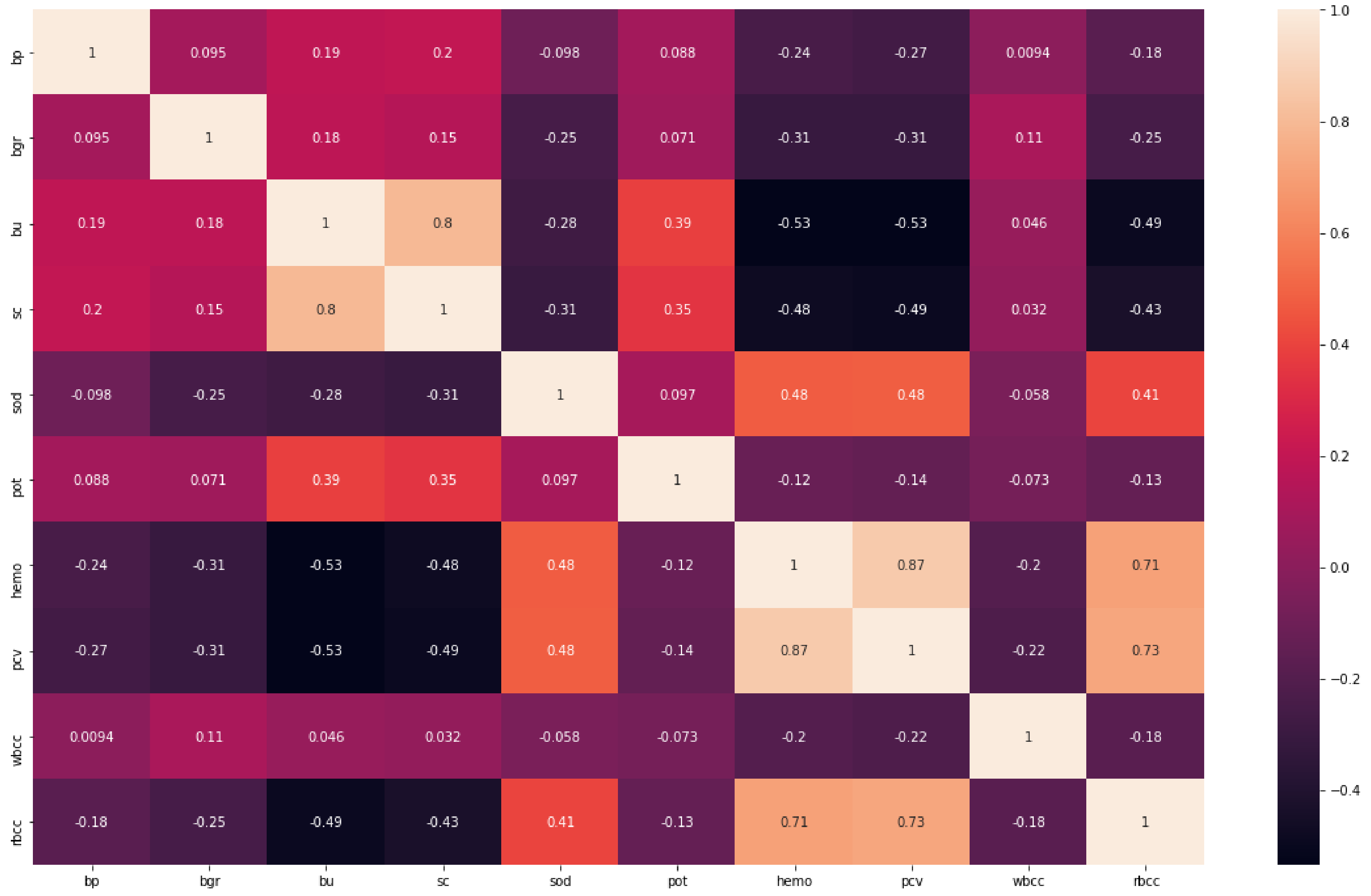
Categorical variables count plot



Numerical variables distribution plots



Numerical variables heatmap



Model Selection- Logistic regression

- Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem
- Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.
- It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

RFE

- Feature ranking with recursive feature elimination.
- Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Model building

- Considering 15 features from the given data applying recursive feature elimination(RFE)
- P-value of "appet_poor" is 0.785 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "pcv" is 0.778 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "sg_1.01" is 0.749 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "dm" is 0.481 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "pe" is 0.311 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "sc" is 0.395 more hence dropping of this feature. And p-value should <= 0.05
- P-value of "sg_1.015 " is 0.072 more hence dropping of this feature. And p-value should <= 0.05
- All the P-values are < 0.05, so that we can select the remaining features for validation

Model building

```
==============================================================================
Dep. Variable:              class_ckd   R-squared:                       0.779
Model:                            GLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     47.82
Date:                Tue, 09 Mar 2021   Prob (F-statistic):           3.52e-58
Time:                        16:43:03   Log-Likelihood:                 11.288
No. Observations:                 219   AIC:                             9.423
Df Residuals:                     203   BIC:                             63.65
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6794      0.116      5.858      0.000       0.451       0.908
bgr            0.0490      0.020      2.445      0.015       0.009       0.088
sc            -0.0190      0.019     -0.975      0.331      -0.057       0.019
sod           -0.0631      0.019     -3.249      0.001      -0.101      -0.025
hemo          -0.1192      0.039     -3.022      0.003      -0.197      -0.041
pcv            0.0113      0.039      0.291      0.771      -0.065       0.088
htn            0.0825      0.049      1.667      0.097      -0.015       0.180
dm             0.0352      0.051      0.684      0.495      -0.066       0.136
pe             0.0409      0.050      0.819      0.414      -0.058       0.139
appet_poor     0.0132      0.048      0.274      0.785      -0.082       0.109
sg_1.01        0.0370      0.117      0.317      0.752      -0.193       0.267
sg_1.015       0.1172      0.116      1.008      0.314      -0.112       0.346
sg_1.02       -0.3428      0.119     -2.879      0.004      -0.578      -0.108
sg_1.025      -0.4434      0.120     -3.684      0.000      -0.681      -0.206
al_1           0.2237      0.052      4.282      0.000       0.121       0.327
al_3           0.1015      0.056      1.813      0.071      -0.009       0.212
==============================================================================
Omnibus:                       28.710   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.909
Skew:                           0.972   Prob(JB):                     1.59e-08
Kurtosis:                       3.400   Cond. No.                         28.2
==============================================================================
```

Model building

```
                            GLS Regression Results
==============================================================================
Dep. Variable:              class_ckd   R-squared:                       0.773
Model:                            GLS   Adj. R-squared:                  0.765
Method:                 Least Squares   F-statistic:                     89.52
Date:                Tue, 09 Mar 2021   Prob (F-statistic):           2.05e-63
Time:                        16:59:14   Log-Likelihood:                 8.2626
No. Observations:                 219   AIC:                             1.475
Df Residuals:                     210   BIC:                             31.98
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7714      0.032     24.056      0.000       0.708       0.835
bgr            0.0537      0.018      3.018      0.003       0.019       0.089
sod           -0.0587      0.019     -3.112      0.002      -0.096      -0.022
hemo          -0.1058      0.024     -4.340      0.000      -0.154      -0.058
htn            0.1080      0.045      2.403      0.017       0.019       0.197
sg_1.02       -0.4247      0.045     -9.390      0.000      -0.514      -0.336
sg_1.025      -0.5250      0.050    -10.546      0.000      -0.623      -0.427
al_1           0.2148      0.052      4.167      0.000       0.113       0.316
al_3           0.1077      0.054      1.981      0.049       0.001       0.215
==============================================================================
Omnibus:                       30.414   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               38.555
Skew:                           0.998   Prob(JB):                     4.25e-09
Kurtosis:                       3.490   Cond. No.                         5.60
==============================================================================
```

Model evaluation

```
In [91]: y_train_pred_final.head()
```

Out[91]:

	class	class_pred	predicted
187	1	0.867121	1
352	0	0.195613	0
254	0	0.216092	0
119	1	0.779420	1
41	1	0.784642	1

```
In [92]: confussion = metrics.confusion_matrix(y_train_pred_final['class'], y_train_pred_final.predicted)
```

```
In [93]: confussion

Out[93]: array([[ 87,   0],
                [ 10, 122]], dtype=int64)
```

```
In [94]: print("Accuracy : ",metrics.accuracy_score(y_train_pred_final['class'], y_train_pred_final.predicted))

         Accuracy :  0.954337899543379
```

Accuracy of the model is 0.95433

Model Testing

```
In [104]: y_pred_final.head()
```

Out[104]:

	class_ckd	class_prob	final_predicted
225	1	1.334385	1
271	0	0.108690	0
108	1	0.745258	1
1	1	0.418736	0
292	0	0.130190	0

```
In [105]: print("Accuracy : ",metrics.accuracy_score(y_pred_final.class_ckd, y_pred_final.final_predicted))

          Accuracy :  0.9368421052631579

In [106]: confusion2 = metrics.confusion_matrix(y_pred_final.class_ckd, y_pred_final.final_predicted )
          confusion2

Out[106]: array([[47,  0],
                 [ 6, 42]], dtype=int64)
```

The accuracy of the model is 0.95 for training data, and for test is 0.93

Results

- The accuracy of the model is 0.95 for training data, and for test it is 0.93

Selected features for building the model are listed below.

  - Blood Glucose Random(bgr)
  - Sodium(sod)
  - Hemoglobin(hemo)
  - Hypertension(htn)
  - Specific Gravity(sg_1.02)
  - Specific Gravity(sg_1.025)
  - Albumin(al_1)
  - Albumin(al_3)

- For identifying Kidney disease, we are recommending the above features.