Chronic kidney disease with machine learning

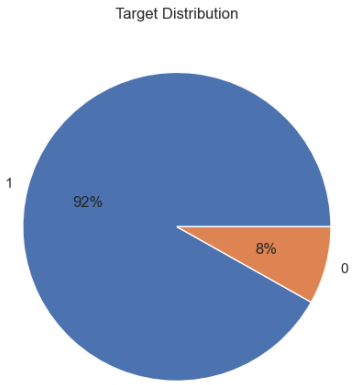
Luka Nikolić

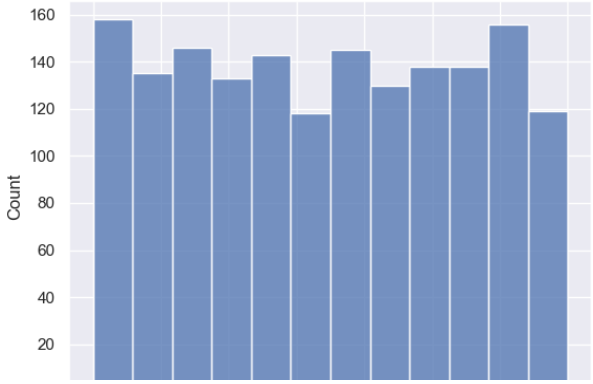
1. Motivation

Doctors of all specializations are in short supply and their time is worth quite a lot. To combat this a machine learning model can be created to help in chronic kidney disease diagnosis.

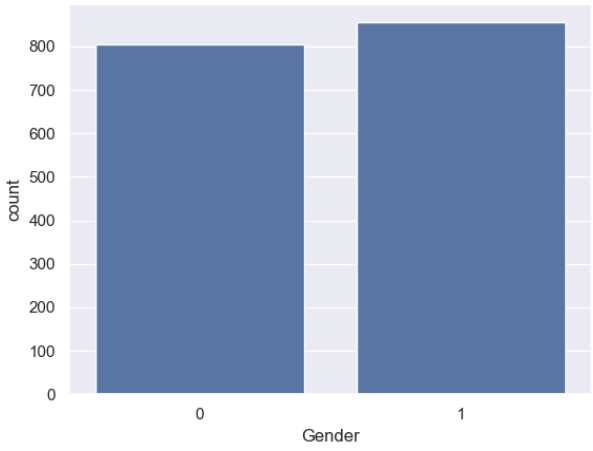
2. Research questions

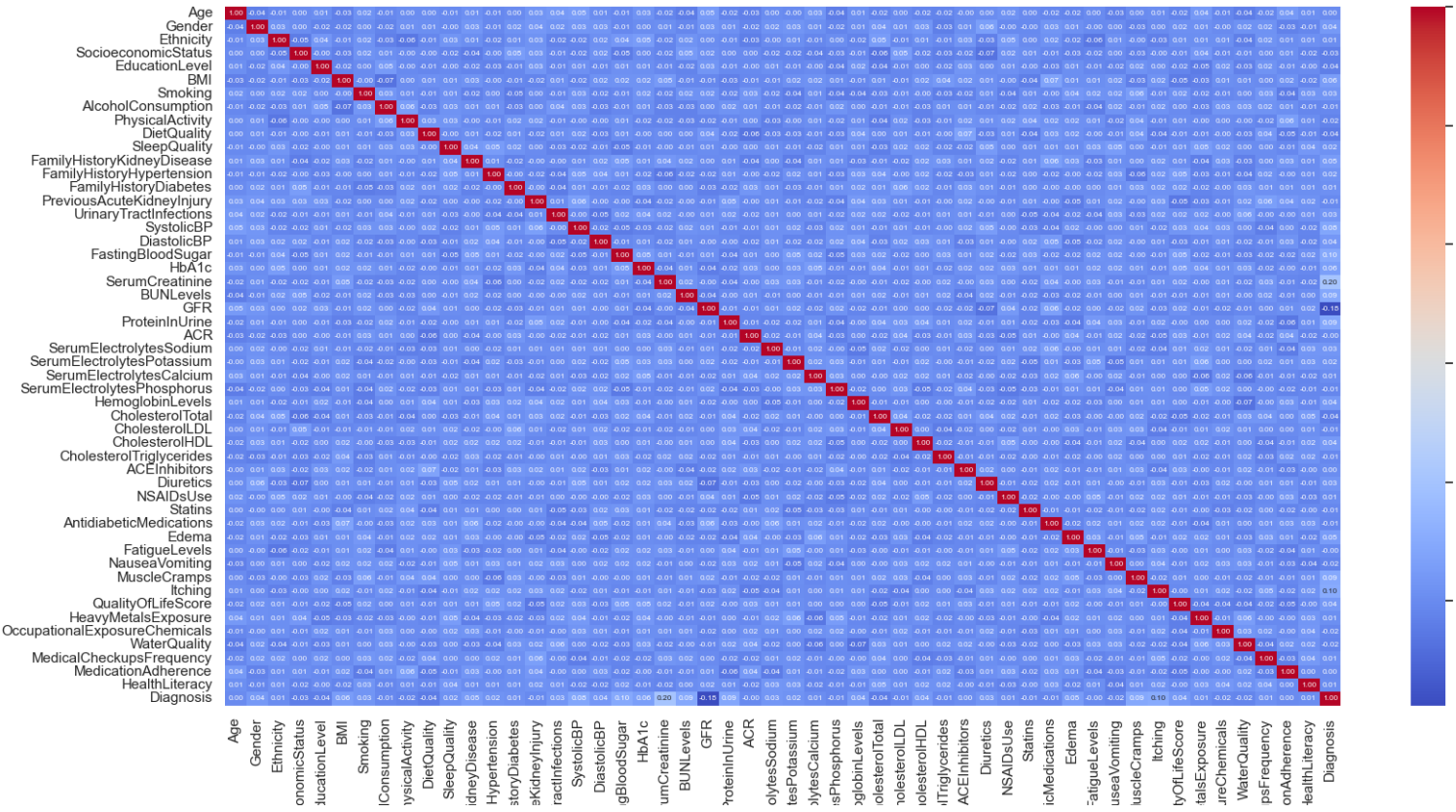
Specifically, i want to make a machine learning model for classification of chronic kidney disease based on various parameters, these include: Age, Gender, Ethnicity, SocioeconomicStatus, EducationLevel, BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality, FamilyHistoryKidneyDisease, FamilyHistoryHypertension, FamilyHistoryDiabetes, PreviousAcuteKidneyInjury, UrinaryTractInfections, SystolicBP, DiastolicBP, FastingBloodSugar, HbA1c, SerumCreatinine, BUNLevels, GFR, ProteinInUrine, ACR, SerumElectrolytesSodium, SerumElectrolytesPotassium, SerumElectrolytesCalcium, SerumElectrolytesPhosphorus, HemoglobinLevels, CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides, ACEInhibitors, Diuretics, NSAIDsUse, Statins, AntidiabeticMedications, Edema, FatigueLevels, NauseaVomiting, MuscleCramps, Itching, QualityOfLifeScore, HeavyMetalsExposure, OccupationalExposureChemicals, WaterQuality, MedicalCheckupsFrequency, MedicationAdherence, HealthLiteracy, Diagnosis. We will now look at how many patients have the diagnosis and some statistics.

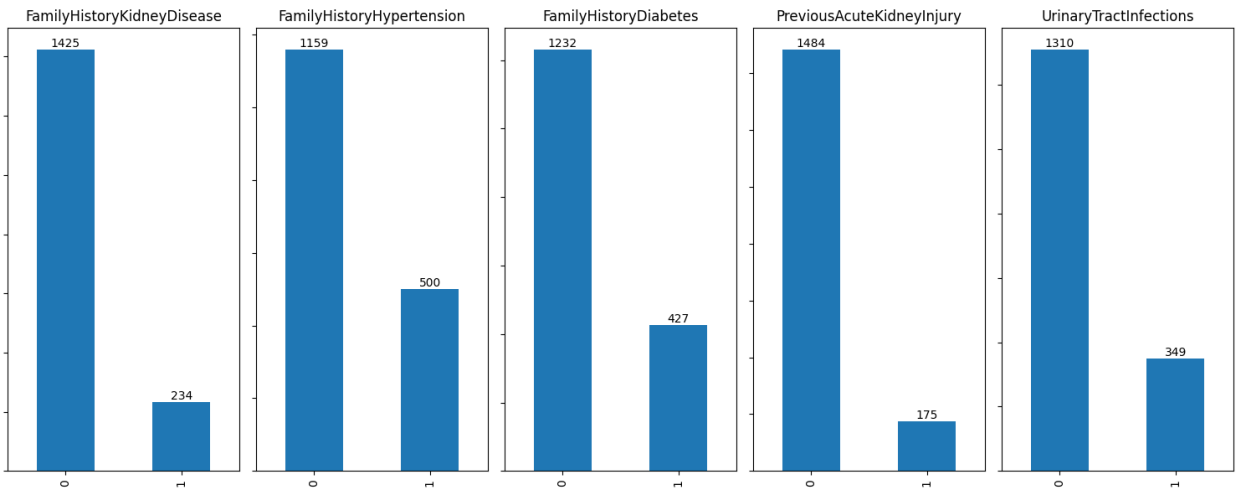




We can see that most of the data is not equally distributed among all age groups, genders, etc. On the correlation heatmap we can see that most of the data has almost no correlation with the exception of GFR because it is a good indicator of kidney function and in implication, chronic kidney disease.







3. Related work

There's quite a number of submissions on Kaggle for this data set and some of them came to the same conclusion that the data is imbalanced and that balancing is required in order to get non biased results. People suggested using some of the balancing tecniques like oversampling, undersampling or SMOTE(Synthetic Minority Oversampling Technique). There wasn't much need for preprocessing data. There was a wide variety of submissions attempting to solve the problem. They used a variety of classifiers and composition classifiers such as AdaBoost,

CatBoost, LightGBM, Perceptron, Ridge, Random Forest, Decision Tree and so on… all of them got a macro F1 score result around ~0.5-0.67

4. Methodology

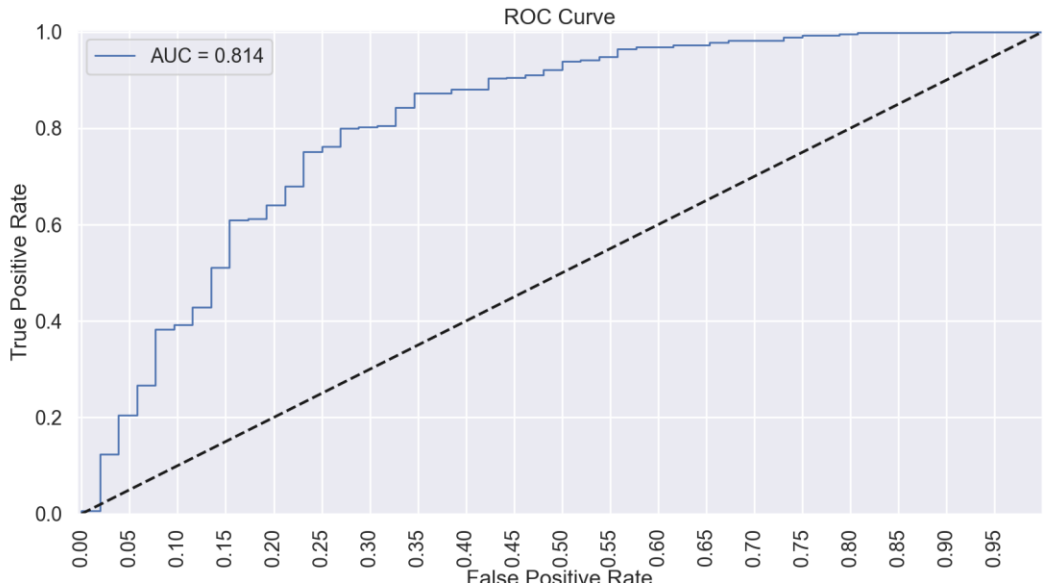
I haver approached the problem using logistical regression. A few unused columns like patientID and DoctorInCharge were dropped. I used a MinMaxScaler to transform all x values into values between 0 and 1. Because of time restrictions and already obtained results that were acceptable (and quite close to other people's score using logistical regression and SMOTE) i did not address imbalanced data in any way

5. Discussion

	precision	recall	f1-score	support
0	0.88	0.13	0.23	52
1	0.91	1.00	0.95	446
accuracy			0.91	498
macro avg	0.89	0.57	0.59	498
weighted avg	0.90	0.91	0.88	498

AUC: 0.8140738185581234

The data set has been split 70:30 into a train set and a test set. The biggest flaw to my approach was not oversampling data of patients without kidney disease or using any of the data balancing tecniques. We can see that the recall for healthy patients is quite low which is a consequence of imbalanced data. The achieved AUC is 0.814 which is not bad but considering that the data is imbalanced, this metric is also probably biased. My final thoughts for further improvement given more time would be to use data balancing techniques, I'm sure they would drastically improve the precision of the model, especially on real world data which would not contain 90% patients with Chronic kidney disease. With more preprocessing and a more balanced dataset,



this tecnique could be used for the intended application of this project

6. References

Data set: https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis/data

Code submissions: https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis/code