Hope Artificial Intelligence

Machine Learning - Classification Assignment

Problem Statement or Requirement:

A requirement from the Hospital, Management asked us to create a predictive model which will predict the chronic kidney disease (CKD) based on the several parameters. The Client has provided the dataset of the same.

1. Identify your problem statement:

The client wants to predict chronic kidney disease with several parameters.

- **Domin:** Machine learning (Numbers based dataset)
- **Learning method:** Supervised learning (Dataset have input and output values)
- **Output:** Classification (Outputs are in categorical value)
- **Prediction:** The patient has Chronical Kidney Disease or not?

2. Tell basic info about the dataset (Total number of rows, columns):

The dataset contains:

- **Total Rows: 400**
- **Total Columns: 25**

  These columns include both input variables and output variable.

- **Count of the output variable:**

```
classification_yes
1    249
0    150
```

  The data set is imbalanced.

- **Input variable details:**

BP	**Blood Pressure** - Specifically, high blood pressure, also known as hypertension, is a common chronic condition that can lead to serious health problems if left untreated. Blood pressure refers to the force of blood pushing against the walls of your arteries as the heart pumps blood.
SG	**Specific Gravity** - Specifically urinary specific gravity, which is a measurement used in urinalysis to assess kidney function and hydration status.
al	**Allostatic load** - It refers to the cumulative physiological burden on the body resulting from the chronic or repeated activation of the body's stress response systems. high allostatic load is associated with increased risk and severity.

su	**Substance Use Disorder** - which can be a significant factor contributing to various chronic conditions. It is also an abbreviation for Subarachnoid hemorrhage and Subconjunctival hemorrhage. Additionally, Chronic spontaneous urticaria (CSU) is a chronic skin condition with the abbreviation "CSU".
rbc	**Red Blood Cell** – It is also known as erythrocytes, are essential for carrying oxygen throughout the body. Chronic diseases can affect the production, function, or lifespan of RBCs, leading to various complications.
pc	**Primary Care** & **Palliative Care -** Primary care refers to the initial point of contact for healthcare, often provided by general practitioners or family doctors, and focuses on the overall health and well-being of a patient. Palliative care, on the other hand, focuses on improving the quality of life for patients with serious illnesses by managing symptoms and providing emotional and spiritual support.
pcc	**Patient Centered Care** - This approach emphasizes the importance of involving patients in their own care, considering their preferences, values, and goals when developing treatment plans and making healthcare decisions.
ba	**Bronchial Asthma** - It is a chronic inflammatory condition of the airways that causes variable and often reversible airflow limitations. Other possibilities, though less common, could include Biliary Atresia (BA), a disease affecting bile ducts in infants, or Bronchopulmonary Dysplasia (BPD), a chronic lung disease in newborns.
bgr	**Blood Glucose Regulator** - It is an Ayurvedic formulation developed in India as an over-the-counter medication for managing type 2 diabetes. The "34" in its name refers to the 34 key ingredients derived from medicinal plants that are used in its composition.
bu	**Buruli ulcer** - It is a chronic necrotizing skin disease caused by Mycobacterium ulcers. While BUN (Blood Urea Nitrogen) is also a common abbreviation in medical contexts, it refers to a blood test used to assess kidney function, not a chronic disease itself.
sc	**Sickle Cell disease** - or more specifically, Hemoglobin SC disease, which is a type of sickle cell disease. It's a genetic disorder affecting red blood cells, leading to various complications.
sod	**Sphincter of Oddi Dysfunction** - It refers to a condition where the sphincter muscle, located at the junction of the bile and pancreatic ducts with the small intestine, doesn't function properly. This can lead to various digestive issues and pain.
pot	**Postural Orthostatic Tachycardia Syndrome** - It's a condition where a change in posture, usually from lying down to standing, causes an abnormally large increase in heart rate. This can lead to symptoms like dizziness, lightheadedness, fatigue, and even fainting.
hrmo	**Highly Resistant Microorganism** – The bacteria that have developed resistance to multiple types of antibiotics, making them difficult to treat. This resistance poses a significant challenge in managing chronic illnesses, especially those that require long-term antibiotic therapy or are prone to secondary infections.
pcv	**Packed Cell Volume** – It is also known as hematocrit. It represents the proportion of red blood cells in the total blood volume, and is often used to assess oxygen-carrying capacity and overall blood health. Abnormal PCV levels can indicate various conditions, including anemia, dehydration, or polycythemia, and are often monitored in chronic disease management.
wc	**Waist Circumference** - It's a measurement used to assess abdominal fat and is a key indicator of metabolic risk and overall health. Elevated waist circumference is associated with an increased risk of developing chronic diseases like cardiovascular disease and diabetes.
rc	**radical cystectomy** - which is a surgical procedure involving the removal of the bladder and surrounding organs. It is often used in the treatment of muscle-invasive bladder cancer. RC can also refer to "resistance-compliance" time, which is a factor in assessing pulmonary hypertension. Additionally, RC can also be an abbreviation for "Research, Condition, and Disease Categorization", a system used by the National Institutes of Health to categorize research project.

htn	**Hypertension** - It is another term for high blood pressure. It's a condition where the force of blood against the artery walls is consistently too high, requiring the heart to work harder to pump blood. Hypertension is a significant risk factor for various other serious health problems like heart disease, stroke, and kidney disease.
dm	**Diabetes Mellitus** - It's a metabolic disorder characterized by elevated blood sugar levels. This can be due to the body not producing enough insulin or being unable to effectively use the insulin it does produce.
cad	**Coronary artery disease** - It's a common condition where the arteries supplying blood to the heart become narrowed or blocked due to plaque buildup, restricting blood flow to the heart muscle.
appet	**Amiodarone-induced Pulmonary Toxicity** - It is a serious adverse effect of the drug amiodarone, which is commonly used to treat certain heart rhythm problems. Additionally, APT can also stand for Antiplatelet Therapy when discussing coronary artery disease.
pe	**Pulmonary Embolism** - It refers to a blockage in one of the pulmonary arteries in the lungs, often caused by a blood clot that originates elsewhere in the body (usually a deep vein thrombosis in the legs). While often considered an acute condition, pulmonary embolism can also have chronic implications for some individuals.
ane	**Anemia of Chronic Disease** - It's a type of anemia (a deficiency in red blood cells or hemoglobin) that develops as a complication of various chronic illnesses, including autoimmune diseases, infections, and kidney disease.

3. **Mention the pre-processing method if you're doing any (like converting string to number – nominal data)**
   - **Encoding Categorical Variables**:

     Some inputs are in nominal data, so applying **One Hot Encoding** method to convert numerical value.
     - sg (a = 0, b = 1, c = 2, d = 3)
     - rbc (yes= 1, no= 0)
     - pc (normal = 1, abnormal = 0)
     - pcc (present = 1, not present= 0)
     - ba (present = 1, not present= 0)
     - htn (yes=1, no=0)
     - dm (yes= 1, no= 0)
     - cad (yes= 1, no= 0)
     - appet (yes= 1, no= 0)
     - pe (yes= 1, no= 0)
     - ane (yes= 1, no= 0)
     - classification (yes= 1, no= 0)

4. **Develop a good model with good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.**

Machine Learning Algorithms:

We can ensure which model have the highest **f1_score** (between **0 to 1**) that model can perform well.

1. Logistic Regression:

a) Confusion Matrix:

```
[[50  1]
[ 0 82]]
```

The model correctly identified 50-persons have chronic kidney disease (TP).

The model correctly identified 1-person don't have chronic kidney disease (TN).

The model incorrectly classified 0-person have disease (FP).

The model incorrectly classified 82-persons don't have disease (FN).

b) Parameters in Grid search:

```
param_grid = {'solver':['newton-cg', 'lbfgs', 'liblinear', 'sag',  'saga'],
              'multi_class' : ['auto', 'ovr', 'multinomial'],
             'penalty':['l1','l2','elasticnet', 'None']}
```

Best parameters:

'multi_class': 'multinomial', 'penalty': 'l2', 'solver': 'newton-cg'

These combinations of parameters that identify the set that yields the best performance of the model.

f1_macro value = 0.9924667654735397

c) Classification Report Explanation:

```
The report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        51
           1       0.99      1.00      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133
```

- **Chronic kidney disease positive:**
  - Precision: 0.99 (model predicted the disease 99% as positive were actually positive)
  - Recall: 1.00 (100% of actual disease in positive were correctly identified)
  - F1-score: 0.99 (a balance between precision and recall) the model catching 99% most actual positive and not mislabeling legitimate disease.
  - Support: 82 (82 persons have disease were labeled as 'yes' in the dataset)

- **Chronic kidney disease negative:**
  - Precision: 1.00 (model predicted the disease 100% as negative were actually negative)
  - Recall: 0.98 (98% of actual don't have disease were correctly identified)
  - F1-score: 0.99 (a balance between precision and recall) the model catching 99% most actual negative and not mislabeling legitimate disease.
  - Support: 51 (51 persons don't have disease were labeled as 'No' in the dataset)
- **Accuracy:**

  0.99 (99% of all predictions were correct)
- **Macro Avg:**

  Macro average: 0.99 (99% correct at identifying people in the "have kidney disease" group and do not have kidney disease" group)
- **Weighted Avg:**
  Weighted Avg: 0.99 (prediction was 99% correct overall, even when considering that there might be more healthy people than sick people in the patient group)

2. Support Vetor Machine:

a) Confusion Matrix:

```
[[24 27]
[11 71]]
```

The model correctly identified 24-persons have chronic kidney disease (TP).

The model correctly identified 27-person don't have chronic kidney disease (TN).

The model incorrectly classified 11-person have disease (FP).

The model incorrectly classified 71-persons don't have disease (FN).

b) Parameters in Grid search:

```
param_grid = {
    'C': [1.0, 10.0, 100.0],
    'kernel': ['rbf','sigmoid'],
    'degree': [3],
    'gamma': ['scale', 'auto'],
    'coef0': [0.0],
    'shrinking': [True],
    'probability': [True],
    'tol': [0.001],
    'cache_size': [200.0],
    'class_weight': [None, 'balanced'],
    'verbose': [False],
    'max_iter': [-1],
    'decision_function_shape': ['ovo', 'ovr'],
    'break_ties': [False],
    'random_state': [None]
}
```

Best parameters:
'C': 100.0, 'break_ties': False, 'cache_size': 200.0, 'class_weight': None, 'coef0': 0.0 , 'decision_function_shape': 'ovo', 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'max _iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'v erbose': False

These combinations of parameters that identify the set that yields the best performance of the model.

f1_macro value = 0.7004060538944259

c) Classification Report Explanation:

```
The report:
              precision    recall  f1-score   support

           0       0.69      0.47      0.56        51
           1       0.72      0.87      0.79        82

    accuracy                           0.71       133
   macro avg       0.71      0.67      0.67       133
weighted avg       0.71      0.71      0.70       133
```

- **Chronic kidney disease positive:**
  - Precision: 0.72 (model predicted the disease 72% as positive were actually positive)
  - Recall: 0.87 (87% of actual disease in positive were correctly identified)
  - F1-score: 0.91 (a balance between precision and recall) the model catching 99% most actual positive and not mislabeling legitimate disease.
  - Support: 82 (82 persons have disease were labeled as 'yes' in the dataset)

- **Chronic kidney disease negative:**
  - Precision: 69 (model predicted the disease 69% as negative were actually negative)
  - Recall: 0.47 (47% of actual don't have disease were correctly identified)
  - F1-score: 0.56 (a balance between precision and recall) the model catching 99% most actual negative and not mislabeling legitimate disease.
  - Support: 51 (51 persons don't have disease were labeled as 'No' in the dataset)
- **Accuracy:**

  0.71 (71% of all predictions were correct)
- **Macro Avg:**

  Macro average: 0.67 (67% correct at identifying people in the "have kidney disease" group and do not have kidney disease" group)
- **Weighted Avg:**

  Weighted Avg: 0.70 (prediction was 70% correct overall, even when considering that there might be more healthy people than sick people in the patient group)

3. Decision Tree:

a) Confusion Matrix:

```
[[51  0]
[ 9 73]]
```

The model correctly identified 51-persons have chronic kidney disease (TP).

The model correctly identified 0-person don't have chronic kidney disease (TN).

The model incorrectly classified 9-person have disease (FP).

The model incorrectly classified 73-persons don't have disease (FN).

b) Parameters in Grid search:

```
param_grid = {'criterion':['gini','entropy'],
              'max_features': ['auto','sqrt','log2'],
              'splitter':['best','random']}
```

Best parameters:

'criterion': 'entropy', 'max_features': 'log2', 'splitter': 'random'

These combinations of parameters that identify the set that yields the best performance of the model.

f1_macro value = 0.9331095830246934

c) Classification Report Explanation:

The report:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	51
1	1.00	0.89	0.94	82
accuracy			0.93	133
macro avg	0.93	0.95	0.93	133
weighted avg	0.94	0.93	0.93	133

- **Chronic kidney disease positive:**
  - Precision: 1.00 (model predicted the disease 100% as positive were actually positive)
  - Recall: 0.89 (89% of actual disease in positive were correctly identified)
  - F1-score: 0.94 (a balance between precision and recall) the model catching 99% most actual positive and not mislabeling legitimate disease.
  - Support: 82 (82 persons have disease were labeled as 'yes' in the dataset)

- **Chronic kidney disease negative:**
  - Precision: 0.85 (model predicted the disease 85% as negative were actually negative)
  - Recall: 1.00 (100% of actual don't have disease were correctly identified)
  - F1-score: 0.92 (a balance between precision and recall) the model catching 99% most actual negative and not mislabeling legitimate disease.
  - Support: 51 (51 persons don't have disease were labeled as 'No' in the dataset)
- **Accuracy:**

  0.93 (93% of all predictions were correct)
- **Macro Avg:**

  Macro average: 0.93 (93% correct at identifying people in the "have kidney disease" group and do not have kidney disease" group)
- **Weighted Avg:**

  Weighted Avg: 0.93 (prediction was 93% correct overall, even when considering that there might be more healthy people than sick people in the patient group)

4. Random Forest:

a) Confusion Matrix:

```
[[51  0]
[ 1 81]]
```

The model correctly identified 51-persons have chronic kidney disease (TP).

The model correctly identified 0-person don't have chronic kidney disease (TN).

The model incorrectly classified 1-person have disease (FP).

The model incorrectly classified 81-persons don't have disease (FN).

b) Parameters in Grid search:

```
param_grid = {'criterion':['gini','entropy'],
              'max_features': ['auto','sqrt','log2'],
              'n_estimators':[10,100]}
```

Best Parameters:

'criterion': 'entropy', 'max_features': 'log2', 'n_estimators': 100

These combinations of parameters that identify the set that yields the best performance of the model.

f1_macro value = 0.9924946382275899

c) Classification Report Explanation:

The report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	51
1	1.00	0.99	0.99	82
accuracy			0.99	133
macro avg	0.99	0.99	0.99	133
weighted avg	0.99	0.99	0.99	133

- **Chronic kidney disease positive:**
  - Precision: 1.00 (model predicted the disease 100% as positive were actually positive)
  - Recall: 0.99 (99% of actual disease in positive were correctly identified)
  - F1-score: 0.99 (a balance between precision and recall) the model catching 99% most actual positive and not mislabeling legitimate disease.
  - Support: 82 (82 persons have disease were labeled as 'yes' in the dataset)

- **Chronic kidney disease negative:**
  - Precision: 0.98 (model predicted the disease 98% as negative were actually negative)
  - Recall: 1.00 (100% of actual don't have disease were correctly identified)
  - F1-score: 0.99 (a balance between precision and recall) the model catching 99% most actual negative and not mislabeling legitimate disease.
  - Support: 51 (51 persons don't have disease were labeled as 'No' in the dataset)

- **Accuracy:**

  0.99 (99% of all predictions were correct)
- **Macro Avg:**

  Macro average: 0.99 (99% correct at identifying people in the "have kidney disease" group and do not have kidney disease" group)
- **Weighted Avg:**

  Weighted Avg: 0.99 (prediction was 99% correct overall, even when considering that there might be more healthy people than sick people in the patient group)

5. **All the research values of each algorithm should be documented. (You can make tabulation or screenshot of the results.)**

   Tabulation of f1_score value from different classification Models:

SI.NO	Classification Models	F1_score value
1	Logistic Regression	0.99246
2	Support Vector Machine	0.70040
3	Decision Tree	0.93310
4	Random Forest	0.99249

6. **Mention your final model, justify why u have chosen the same.**

   The Final Model: is **Random Forest**

   Justification for Choosing Random Forest Classifier:

   1. **Based on Performance:**

      - Comparatively the best performance of the Classification Model is Random Forest Classifier and it gives F1_score value = 0.99249

Our tool is really good at telling if someone has kidney disease or not **Random Forest Classifier**. you can trust it a lot when checking people, especially for those who are healthy, because it's very good at spotting them correctly.