Big Homework, LazyFCA

Nikita Kharzheev

December 2023

1 Description of first dataset

`https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data`

I have chosen healthcare dataset. About this dataset:

1. Name: This column represents the name of the patient associated with the healthcare record.
2. Age: The age of the patient at the time of admission, expressed in years.
3. Gender: Indicates the gender of the patient, either "Male"or "Female."
4. Blood Type: The patient's blood type, which can be one of the common blood types (e.g., "A+ "O- etc.).
5. Medical Condition: This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes,Hypertension, Asthma,"and more.
6. Date of Admission: The date on which the patient was admitted to the healthcare facility.

7. Doctor: The name of the doctor responsible for the patient's care during their admission.

8. Hospital: Identifies the healthcare facility or hospital where the patient was admitted.

9. Insurance Provider: This column indicates the patient's insurance provider, which can be one of several options, including "Aetna,Blue Cross,Cigna, UnitedHealthcare,"and "Medicare."

10. Billing Amount: The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.

11. Room Number: The room number where the patient was accommodated during their admission.

12. Admission Type: Specifies the type of admission, which can be "Emergency, Elective,"or "Urgent,"reflecting the circumstances of the admission.

13. Discharge Date: The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.

14. Medication: Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin,Ibuprofen,Penicillin,Paracetamol,"and "Lipitor."

15. Test Results: Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal,Abnormal,"or "Inconclusive,"indicating the outcome of the test.

2 Data Pre-Processing

The dataset is shown below

	Age	Gender	Medical Condition	Date of Admission	Insurance Provider	Billing Amount	Admission Type	Discharge Date	Medication	Test Results	days
0	80	Female	Cancer	2019-09-06	Medicare	24358.674062	Urgent	2019-09-17	Paracetamol	Abnormal	11
1	59	Male	Cancer	2022-03-27	Blue Cross	15450.181900	Emergency	2022-04-09	Ibuprofen	Abnormal	13
2	72	Female	Arthritis	2021-12-24	Blue Cross	42258.431826	Urgent	2022-01-18	Lipitor	Inconclusive	25
3	47	Male	Hypertension	2019-05-06	UnitedHealthcare	4397.847075	Elective	2019-05-28	Lipitor	Normal	22
4	35	Female	Diabetes	2022-10-23	Blue Cross	45960.250694	Urgent	2022-10-28	Ibuprofen	Normal	5
...	...	...	...	...	...	...	...	...	...	...	...
494	57	Female	Arthritis	2021-02-14	Medicare	21291.259735	Elective	2021-03-13	Ibuprofen	Abnormal	27
495	73	Male	Hypertension	2023-02-07	UnitedHealthcare	26156.213404	Urgent	2023-03-09	Lipitor	Inconclusive	30
496	42	Female	Cancer	2020-01-02	Medicare	26419.324813	Emergency	2020-01-30	Ibuprofen	Inconclusive	28
497	54	Female	Cancer	2021-12-17	Cigna	16479.896916	Elective	2021-12-27	Paracetamol	Normal	10
498	58	Female	Cancer	2021-03-17	UnitedHealthcare	36799.573154	Elective	2021-04-05	Ibuprofen	Inconclusive	19

Our data preparation process contains several steps:

1. Deliting extra colums
2. Deliting extra colums
3. One-hot encoding

   You can see detail in code. Result is below.

	Age18_40	Age40_62	Age62_85	Small_bill	Medium_bill	Large_bill	Small_days	Medium_days	Large_days	Gender_Female	...	Provi
0	False	True	False	True	False	False	True	False	False	True	...	
1	False	True	False	False	False	True	False	True	False	False	...	
2	False	True	False	True	False	False	False	True	False	True	...	
3	False	True	False	True	False	False	True	False	False	False	...	
4	True	False	False	True	False	False	False	True	False	True	...	

3 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. FCA
2. KNN
3. Logistic Regression
4. Decision Tree
5. Random Forest Classifier

	model	Accuracy
0	FCA	0.6533
1	KNN	0.6267
2	LogisticRegression	0.7000
3	DecisionTree	0.6267
4	RandomForest	0.6400

4 Description of first dataset

`https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`

I have chosen healthcare dataset. About this dataset:

1. gender: Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories in it male ,female and other.

2. age: Age is an important factor as diabetes is more commonly diagnosed in older adults.Age ranges from 0-80 in our dataset.

3. hypertension: Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values a 0 or 1 where 0 indicates they don't have hypertension and for 1 it means they have hypertension.

4. heart-disease: Heart disease is another medical condition that is associated with an increased risk of developing diabetes.It has values a 0 or 1 where 0 indicates they don't have heart disease and for 1 it means they have heart disease.

5. smoking-history: Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes.In our dataset we have 5 categories i.e not current,former,No Info,current,never and ever.

6. bmi: BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.

7. HbA1c-level: HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5 of HbA1c Level indicates diabetes.

8. blood-glucose-level: Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.

9. diabetes: Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

5 Data Pre-Processing

The dataset is shown below

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Male	5.0	0	0	No Info	12.73	3.5	126	0
1	Male	8.0	0	0	No Info	15.22	6.1	85	0
2	Female	80.0	0	0	never	27.99	6.2	80	0
3	Male	67.0	0	0	not current	32.72	5.7	140	0
4	Female	15.0	0	0	No Info	27.32	5.8	126	0
...	...	...	...	...	...	...	...	...	...
494	Female	24.0	0	0	never	34.65	3.5	159	0
495	Male	6.0	0	0	No Info	16.22	6.5	100	0
496	Female	77.0	0	0	never	27.32	4.8	85	0
497	Male	49.0	0	0	never	21.84	5.0	130	0
498	Female	17.0	0	0	never	24.34	3.5	100	0

Our data preparation process contains several steps:

1. Deliting extra colums
2. Deliting extra colums
3. One-hot encoding

   You can see detail in code. Result is below.

	bmi10_34	bmi34_58	bmi58_82	Age0_27	Age27_54	Age54_80	small_lvl	high_lvl	small_gluc	middle_gluc	...	HbA1c_level	blood_
0	True	False	False	False	True	False	True	False	True	False	...	True	
1	True	False	False	False	True	False	True	False	True	False	...	True	
2	False	True	False	False	True	False	True	False	True	False	...	True	
3	True	False	False	False	True	False	False	True	True	False	...	True	
4	True	False	False	False	False	True	True	False	True	False	...	True	

6 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. FCA
2. KNN
3. Logistic Regression
4. Decision Tree
5. Random Forest Classifier

	model	Accuracy
0	FCA	0.9533
1	KNN	0.9667
2	LogisticRegression	0.9667
3	DecisionTree	0.9667
4	RandomForest	0.9667

7 Description of first dataset

`https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset`

I have chosen employee dataset. About this dataset:

1. Education: The educational qualifications of employees.

2. JoiningYear: The year each employee joined the company.

3. City: The location or city where each employee.

4. PaymentTier: Categorization of employees into different salary tiers.

5. Age: The age of each employee.

6. Gender: Gender identity of employees.

7. EverBenched: Indicates if an employee has ever been temporarily without assigned work.

8. ExperienceInCurrentDomain: The number of years of experience employees have in their current field.

9. LeaveOrNot: target column.

8 Data Pre-Processing

The dataset is shown below

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2014	Bangalore	3	27	Female	No	5	0
1	Bachelors	2013	New Delhi	3	25	Female	Yes	3	1
2	Bachelors	2015	Pune	3	41	Female	No	3	1
3	Bachelors	2017	New Delhi	2	34	Female	No	4	0
4	Bachelors	2012	Bangalore	3	33	Female	No	1	0
...	...	...	...	...	...	...	...	...	...
494	PHD	2015	New Delhi	3	31	Male	No	3	0
495	Bachelors	2014	Bangalore	3	27	Female	No	5	0
496	Masters	2017	New Delhi	3	24	Male	No	2	0
497	Bachelors	2017	Bangalore	3	27	Female	No	5	0
498	Bachelors	2012	Bangalore	3	26	Male	No	4	0

Our data preparation process contains several steps:

1. Deliting extra colums
2. Deliting extra colums
3. One-hot encoding

   You can see detail in code. Result is below.

	Age22_28	Age28_35	Age35_41	small_lvl	high_lvl	late_year	not_late_year	small_exp	big_exp	JoiningYear	...	Education_Bachelors
0	False	True	False	True	False	True	False	False	True	True	...	True
1	True	False	False	True	False	True	False	False	True	True	...	True
2	True	False	False	True	False	True	False	True	False	True	...	True
3	True	False	False	True	False	True	False	True	False	True	...	True
4	True	False	False	True	False	True	False	False	True	True	...	True

9 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. FCA
2. KNN
3. Logistic Regression
4. Decision Tree
5. Random Forest Classifier

	model	Accuracy
0	FCA	0.8067
1	KNN	0.6867
2	LogisticRegression	0.7267
3	DecisionTree	0.7533
4	RandomForest	0.7933