Dr. Regis

DSC326

Comparison of Machine Learning Methods for Classification and Regression

Daniela Cruz-Perez

Emily Foreacre

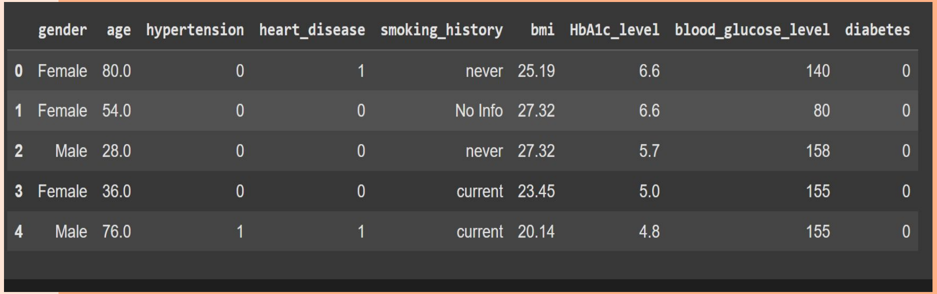
Classification

Dataset:

Diabetes was the 7th leading cause of death in the US in 2023. The Diabetes Prediction Database includes 100,000 patient medical records and their diabetic status. There are 8 features:

- Categorical: Age, Smoking History
- Binary: Hypertension (high blood pressure) and Heart Disease status
- Numerical: BMI Index, HbA1c levels (average blood sugar level over the past 2 to 3 months) and *Blood Glucose level.*

These features can help Medical Professionals to classify whether a patient is at risk of developing diabetes.



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Hyperparameter Tuning + Training

K-NN

K=13

MLP

Hidden Node= 30
Hidden Layer = 1
Activation Function = relu

RF

Trees = 100
Max. Depth=20
Min Leaves =5

SVC

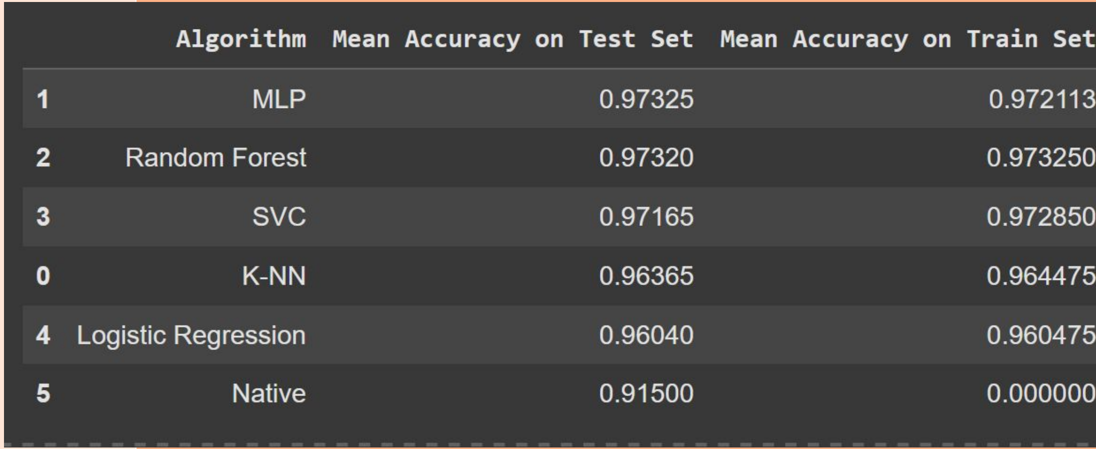
Kernel Function= rbf
C=100

Performance

MLP led performance with Random Forest just slightly under

Logistic Regression slightly Underperformed

Interestingly,
The other models performed substantially Better than the Naive Algorithm and Although RF had a high accuracy across different parameters, MLP had a higher accuracy



	Algorithm	Mean Accuracy on Test Set	Mean Accuracy on Train Set
1	MLP	0.97325	0.972113
2	Random Forest	0.97320	0.973250
3	SVC	0.97165	0.972850
0	K-NN	0.96365	0.964475
4	Logistic Regression	0.96040	0.960475
5	Native	0.91500	0.000000

Regression

Intro

Dataset

The LasVegasTripAdvisorReviews dataset consists of 504 hotel reviews on Tripadvisor, all collected between January and August of 2015.

Features used:

Nominal: 'Traveler type', 'User continent', 'Hotel name', 'User country', 'Period of stay'.

Ordinal: 'Hotel stars', 'Score'

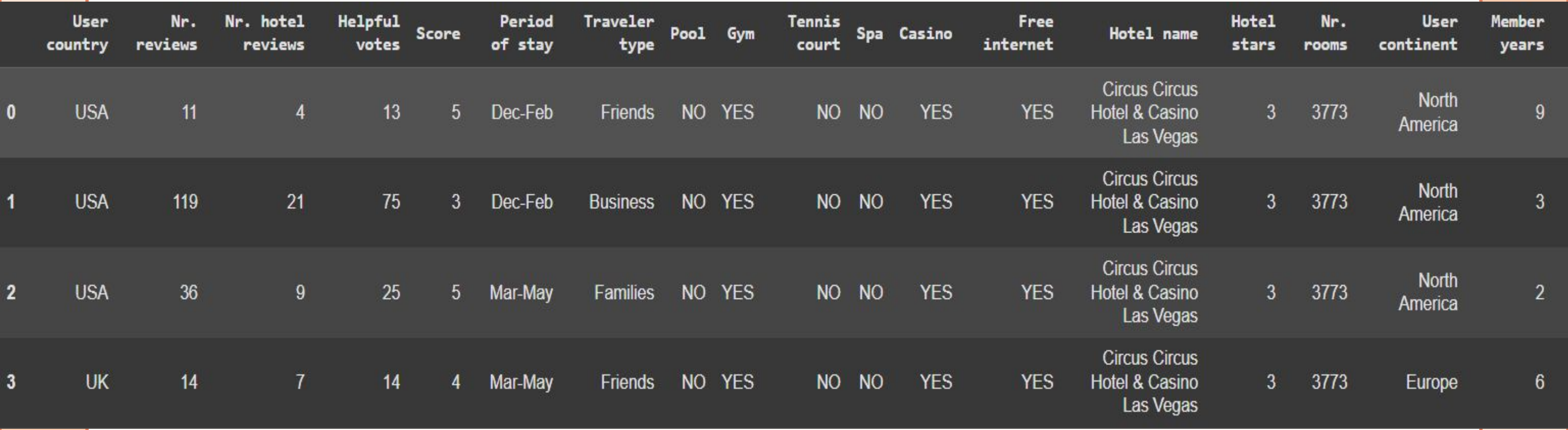
Numeric: 'Nr. rooms', 'Nr. reviews', 'Nr. hotel reviews', 'Helpful votes'

Binary: 'Pool', 'Gym', 'Tennis court', 'Spa', 'Casino', 'Free internet'

Target: ‘Member years’

Dropped: 'Review month', 'Review weekday'

The goal of the regression problem is to predict Member years, the number of years a user has been active on TripAdvisor, based on the features.



	User country	Nr. reviews	Nr. hotel reviews	Helpful votes	Score	Period of stay	Traveler type	Pool	Gym	Tennis court	Spa	Casino	Free internet	Hotel name	Hotel stars	Nr. rooms	User continent	Member years
0	USA	11	4	13	5	Dec-Feb	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	9
1	USA	119	21	75	3	Dec-Feb	Business	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	3
2	USA	36	9	25	5	Mar-May	Families	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	2
3	UK	14	7	14	4	Mar-May	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	Europe	6

Hyperparameter Tuning + Training

K-NN

Best K=41

MPR

Hidden Node= 10
Hidden Layer = 1
Activation Function = logistic

RF

Trees = 200
Max. Depth=20
Min Leaves =5

SVR

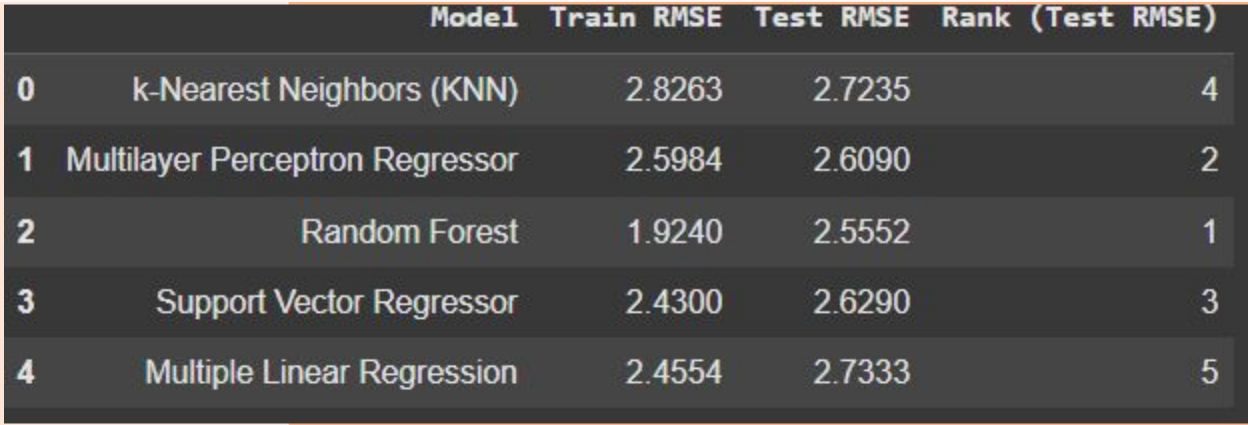
Kernel Function= rbf
C=10
Epsilon=0.5

Performance

Random Forest: had the lowest test RMSE, indicating it generalizes best.

MLP and SVR: also showed ok test performance, though with slightly higher RMSE.

KNN and Linear Regression: underperformed, likely due to limitations in handling complex patterns in the data.



	Model	Train RMSE	Test RMSE	Rank (Test RMSE)
0	k-Nearest Neighbors (KNN)	2.8263	2.7235	4
1	Multilayer Perceptron Regressor	2.5984	2.6090	2
2	Random Forest	1.9240	2.5552	1
3	Support Vector Regressor	2.4300	2.6290	3
4	Multiple Linear Regression	2.4554	2.7333	5

Principal Components Regression =
PCR Train RMSE: 2.6937
PCR Test RMSE: 2.6426
PCR Train R-squared: 0.1699
PCR Test R-squared: 0.1194

Second-Order Multiple Linear Regression=
Train RMSE: 0.0000
Test RMSE: 22.2491

Interesting Facts

Outliers

Support Vector Regressor

Ordinal treated as string

Correlation

Questions?