Student Number: 220681535

ST3189

Machine Learning

Coursework

Table of contents

Unsupervised Learning

For the task of Unsupervised Learning I decided to consider the problem of behavior of credit card holders. I defined the research problem of this task in the next way: identify customer behavioral groups and describe them. This information can be used by other people to build marketing and financial strategies to improve the system.

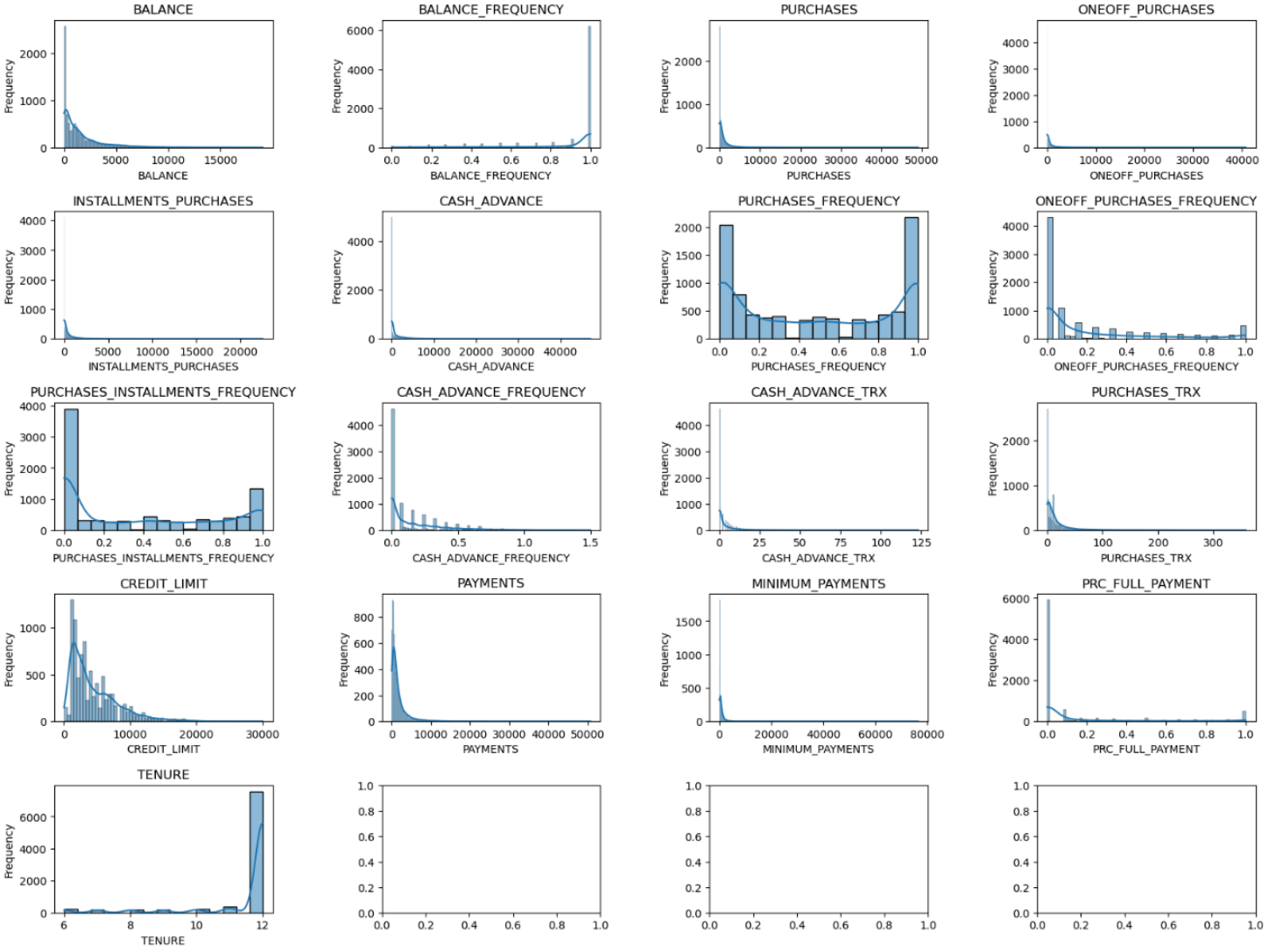
Working with this theme, we can consider one of the researches [1], where authors perform credit card users segmentation using K-means clusterization. In the result of this work they get 4 segments: Jane, Fashion Lovers, Executives and Limited Spenders. Unfortunately, researchers for confidentiality issues omit some details about their dataset, so I won't be able to work with it and try to find some new groups. Additionally we can consider the work of Sadrac Pierre [2], where he uses RFM segmentation to solve the same problem. In the result he gets next clusters: "Premium Customer", "Repeat Customer", "Top Spender", "At-risk Customer" and "Inactive Customer". As we can see, he gets completely different results in comparison with previous research. It shows that there is a big variety of how it is possible to segment clients. Instead, I will use another dataset from kaggle on this theme. In the end I will compare results, and probably will find some other patterns of behavior. My dataset looks in the next way: in the below table you can see variables and their statistical properties

	BALANCE	BALANCE FREQUENCY	PURCHASES	ONEOFF PURCHASES	INSTALLMENTS PURCHASES	CASH ADVANCE	PURCHASES FREQUENCY	ONE OFF PURCHASE FREQUENCY
mean	1564.47	0.88	1003.20	592.44	411.07	978.87	0.49	0.20
std	2081.53	0.24	2136.63	1659.89	904.34	2097.16	0.40	0.30
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	128.28	0.89	39.63	0.00	0.00	0.00	0.08	0.00
50%	873.39	1.00	361.28	38.00	89.00	0.00	0.50	0.08
75%	2054.14	1.00	1110.13	577.40	468.64	1113.82	0.92	0.30
max	19043.14	1.00	49039.57	40761.25	22500.00	47137.21	1.00	1.00

	PURCHASES INSTALLMENTS FREQUENCY	CASH ADVANCE FREQUENCY	CASH ADVANCE TRX	PURCHASES TRX	CREDIT LIMIT	PAYMENTS	MINIMUM PAYMENTS	PRC FULL PAYMENT
mean	0.36	0.14	3.25	14.71	4494.45	1733.14	864.21	0.15
std	0.40	0.20	6.82	24.86	3638.82	2895.06	2372.45	0.29
min	0.00	0.00	0.00	0.00	50.00	0.00	0.02	0.00
25%	0.00	0.00	0.00	1.00	1600.00	383.28	169.12	0.00

50%	0.17	0.00	0.00	7.00	3000.00	856.90	312.34	0.00
75%	0.75	0.22	4.00	17.00	6500.00	1901.13	825.49	0.14
max	1.00	1.50	123.00	358.00	30000.00	50721.48	76406.21	1.00
null	0.00	0.00	0.00	0.00	1.00	0.00	313.00	0.00

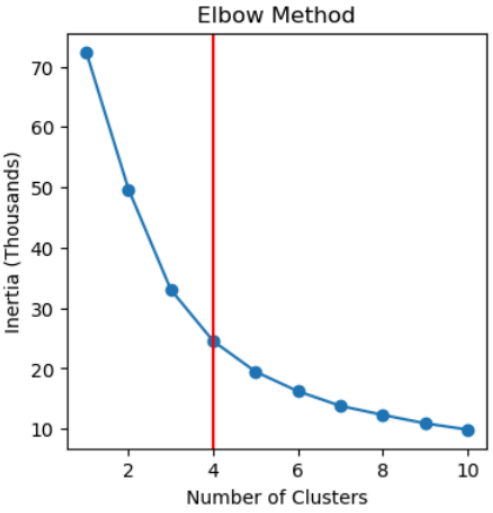
The dataset has a shape (8950, 17). As we can see from this table it has several empty values here - 1 in CREDIT_LIMIT and 313 in MINIMUM_PAYMENTS. Now let's investigate the distribution of our variables.

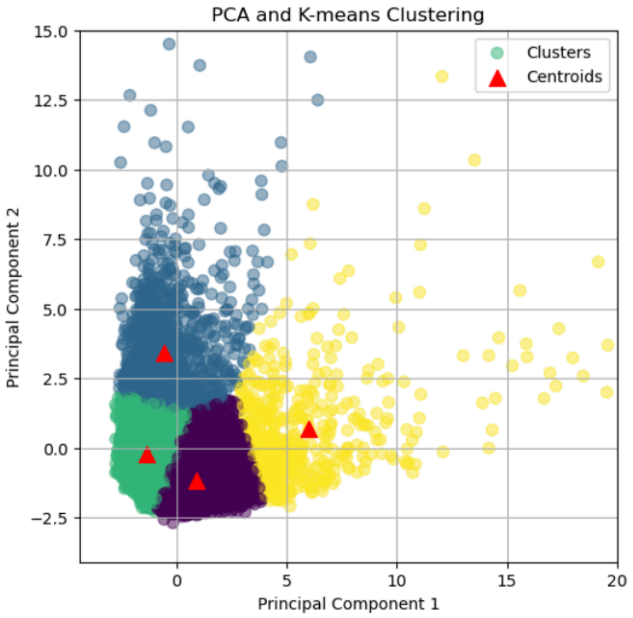


With these plots we can see the distribution of every variable in our dataset and detect some anomalies. After investigating this plot no problems were found.

Now let's make a cluster of our clients. At first I filled empty values with average, then I scale data and use PCA with 2 components and after that I use the Elbow method for finding optimal number of clusters. From the picture we can see the elbow point at 4 clusters. In the result I get 4 clusters with clear boundaries. Before interpretation of the model we need to assess it. For this I used Silhouettes Score and obtained a value of 0.4, which is quite good. After that I provided Kruskal

Wallis H test, which shows if there are statistically significant differences between the medians of groups(cluster) and it showed that clusters statistically differ.





After analyzing these clusters, I made characteristic for each of them:

1. This cluster has the smallest balance with moderate credit limit and high purchase frequency. They have equal of one-off and installment purchases and smallest cash using
2. Has the highest balance and cash advance, small purchase frequency. This shows that these people try to save money and try to avoid credits.
3. Moderate balance and the smallest purchase frequency. This cluster has smallest installments purchases, showing that they live within their means.
4. Cluster with the highest credit limit and purchase frequency. They have biggest installments, so we can call credit and shopping lovers

Comparing my results with the theoretical background from the beginning we can notice similarities and differences. For example, my 4 cluster is extremely similar in description to the Limited Spenders segment [1]. It is not possible to compare the other segments, since the study focuses more on the types of purchases, rather than on credit limits, the number of installments, etc.

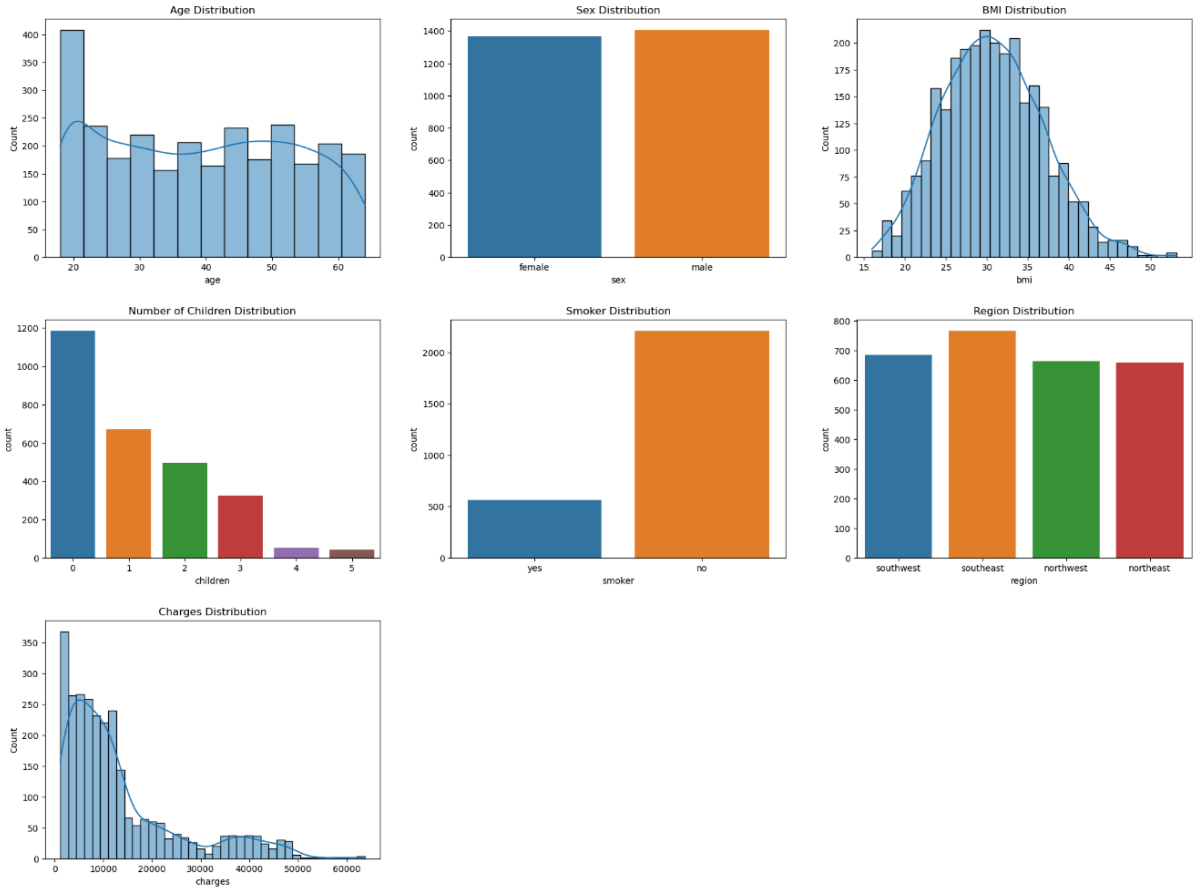
Talking about Sadrac Pierre study [2], we can spot similarities, because this research analyzes practically the same things as I, but with another method. So his “At-risk Customer” reminds me of 3 segments, “Repeat Customer” matches with my 1 cluster, “Premium Customer” and “Top Spender” coincides with my 2 and 4 segments.

Regression

For the task of regression I decided to research the cost of medical insurance. I defined a research problem as follows: to build a model to predict the cost of insurance and analyze what factors affect its cost.

In terms of literature review for this theme, we can discuss the work of Sazzad Hossen [3], where he researches health insurance costs using XGBoost regression. He gets R^2 of 0.8681, which is a good result, but the problem is that this model is not very interpretable, so we can't get any theoretical understanding of this problem, especially how different factors affect the price. Unfortunately, other researchers also don't provide any interpretations for this problem, so I will build hypotheses based on life experience. Most likely, health insurance cost depends on the age of the clients, his lifestyle and health. For example, if client consumes a lot of alcohol, has overweight, then risk for insurance case increases, so insurance itself will cost more.

I chose a dataset for this task from Kaggle. It has shape (2772, 7) with 7 variables: charges, age, sex, bmi, children, smoker and region. Dataset is clean and has no missing values. Let's investigate the distribution of our variables.

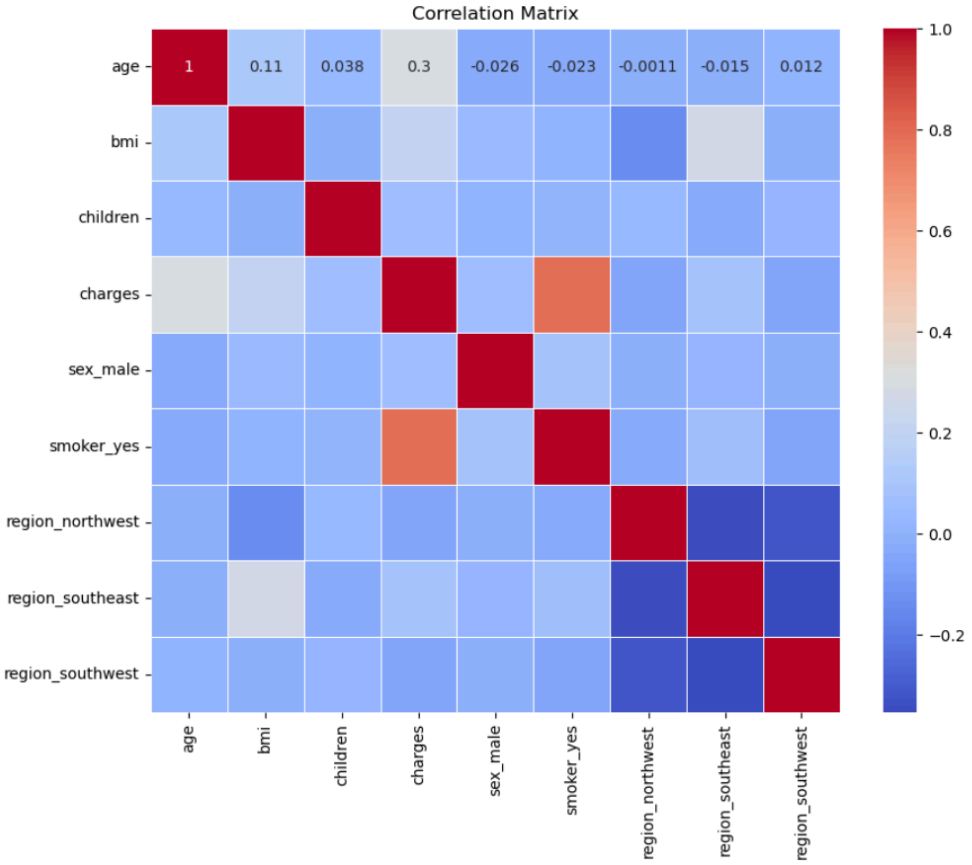


As we can see we don't have any anomalies, most of the variables are distributed normally or equally. Also I tried to find some functional dependencies between target and other variables, but there were no insights from these graphs,

so I won't put it here. Now let's investigate relations between our variables using correlation heatmap. As we can see from, our target variables strongly correlate only with smoker_yes, so we should take it into account.

Next was preparing data for modeling. I performed one-hot encoding, splitted data into train and test samples, scaled values and started building different models. Here's performance of all my models:

Linear Regression MSE: 39933194.54805147
Linear Regression R^2: 0.73981661775643
Trans Regression MSE: 39001973.55824823
Trans Regression R^2: 0.74588395670803
Random Forest Regressor MSE: 7566043.948864406
Random Forest Regressor R^2: 0.9507036958325416
Linear regression with partial poly MSE: 39215704.953494646
Linear regression with partial poly transformation R^2: 0.7444913970108604
Polynomial Regression MSE: 25157613.30703466
Polynomial Regression R^2: 0.836086418993506
Ridge Regression MSE: 39935330.427993745
Ridge Regression R^2: 0.7398027015027033
Lasso Regression MSE: 39933258.542003006
Lasso Regression R^2: 0.7398162008059972



As we can see from this output, Random Forest regressor has the best MSE and R^2 - 0.95 which is very high, so it will be our model for prediction problems.

Now let's interpret the results. It's very hard to analyze Random Forest results, but one of the things we can do is to calculate feature importance. So, the most important feature is smoker_yes with 0.6 value, next place is after bmi with 0.2 importance there is age on the third place with 0.12 value. Other variables have very little importance. I think this interpretation is not enough, so to analyze which factors affect the cost of insurance, I will use a model with a transformed variable which has R^2 equal 0.746 which is quite big. So here's its summary.

We can see that smoking(smoker_yes) is one of the key factors defining the price of insurance - it increases cost by 9714. we can also see that the cost of insurance increases
depending on age(age and age^2), and the rate of cost growth increases as a person gets older.

Another important factor is body mass index (bmi, bmi^2). We can see that this marker increases the price of the insurance, but the quadratic part shows that this influence decreases at high levels of bmi.

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.757
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	626.0
Date:	Wed, 03 Apr 2024	Prob (F-statistic):	0.00
Time:	15:52:51	Log-Likelihood:	-22415.
No. Observations:	2217	AIC:	4.485e+04
Df Residuals:	2205	BIC:	4.492e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.314e+04	235.766	55.718	0.000	1.27e+04	1.36e+04
age	3484.2734	128.366	27.143	0.000	3232.542	3736.005
bmi	2028.9736	135.272	14.999	0.000	1763.699	2294.248
children	984.6383	175.632	5.606	0.000	640.217	1329.059
sex_male	-29.9703	127.422	-0.235	0.814	-279.850	219.909
smoker_yes	9714.7217	127.897	75.957	0.000	9463.910	9965.533
region_northwest	-218.5099	157.057	-1.391	0.164	-526.506	89.486
region_southeast	-488.8636	163.650	-2.987	0.003	-809.789	-167.939
region_southwest	-472.6827	157.824	-2.995	0.003	-782.182	-163.183
age^2	770.0616	155.074	4.966	0.000	465.956	1074.167
bmi^2	-296.5355	97.381	-3.045	0.002	-487.504	-105.567
children^2	-238.8547	113.438	-2.106	0.035	-461.312	-16.398

Omnibus:	496.949	Durbin-Watson:	1.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1204.303
Skew:	1.221	Prob(JB):	3.08e-262
Kurtosis:	5.659	Cond. No.	4.78

Talking about other variables, we can see that they have quite big P values, so they are statistically insignificant so we can not interpret them.

In conclusion, we can compare the results with my hypothesis from the beginning: it was stated that the more unhealthy lifestyle a person leads, the more he will pay for insurance. In the result of the survey it was found out, that one of the

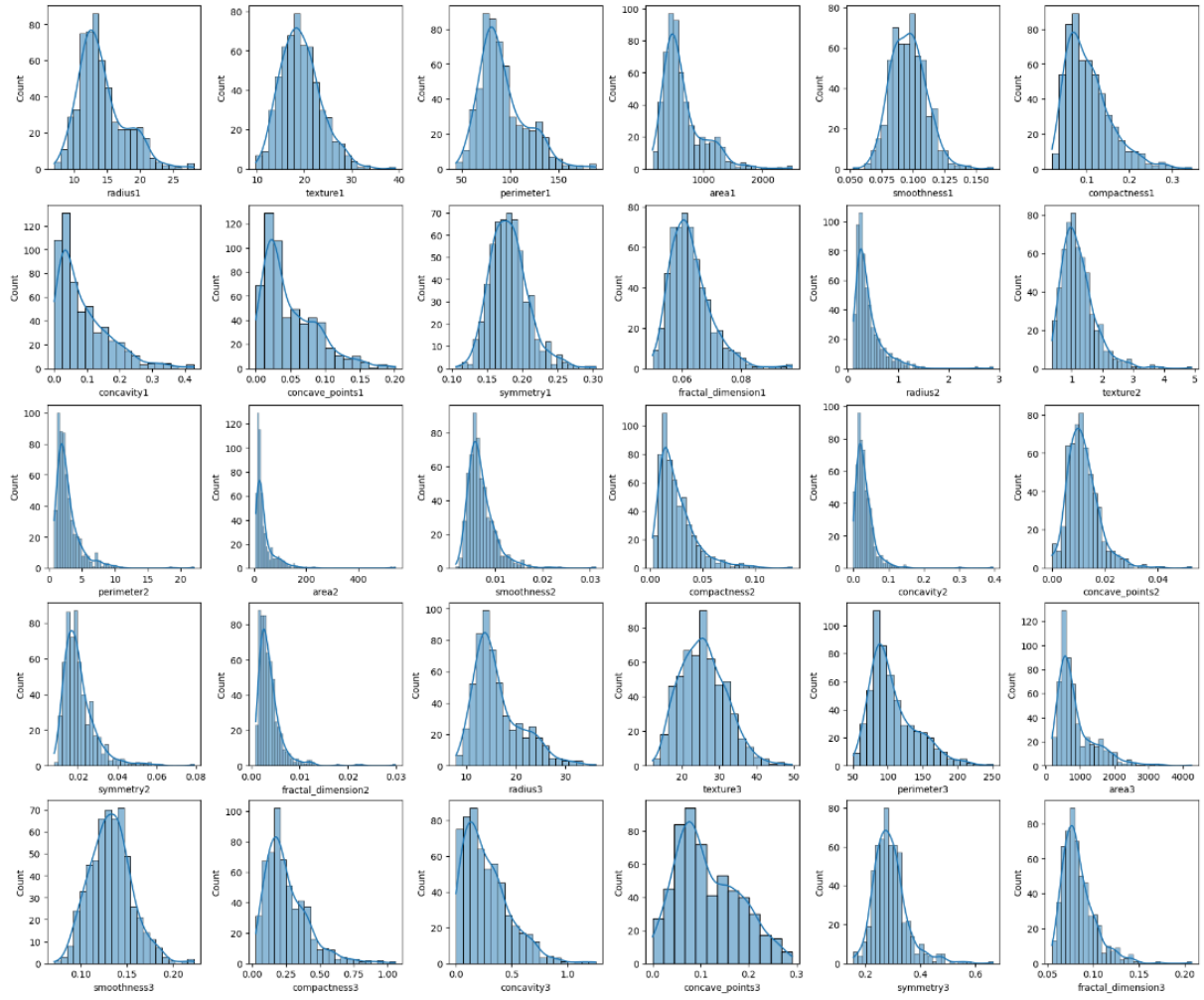
main factors of insurance is smoking and high BMI, so we can say that assumption from beginning confirmed.

Classification

I decided to devote the classification tasks section to biology, in particular the classification of benign and malignant breast tumors. Research problem is formulated in the following way: build and compare machine learning models which can with high accuracy classify the type of the tumors and find features that increase the odds of this event.

For this problem I used a dataset from UCI repository called “Breast Cancer Wisconsin (Diagnostic)”. It consists of 30 features that were obtained from digital images of the biopsy. Target variable can take 2 values - M (malignant) and B (benign).

Dataset don’t have missing values, distribution of practically all variables is unimodal, you can see it in the picture below.



Target variables don’t have a big imbalance in distribution: B - 357, M - 212.

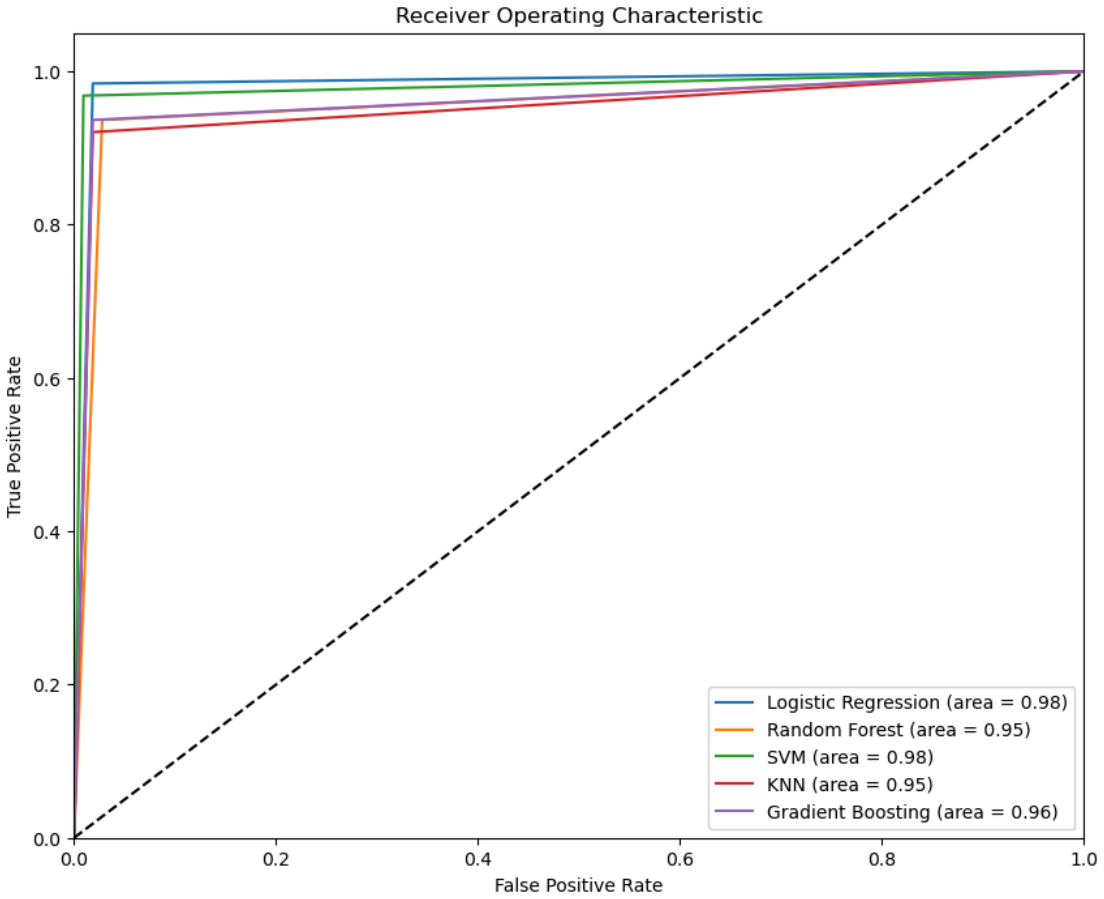
After exploring the data we can start to prepare it for our machine learning models. First of all I will code target variables, split data into train and test samples, and scale the data.

After preparing the data, I started building models. As a baseline I took the following models with default settings: Logistic Regression, Random Forest

Classifier, Support Vector Machine, K-Neighbors Classifier and Gradient Boosting Classifier. After that I applied Greed Search to every model to find optimal hyperparameters. This increased accuracy by 0.01 by average. Here's final results of every model:

```
Logistic regression: Accuracy = 0.9824561403508771, F1-score =
0.9763779527559054
Random Forest Classifier: Accuracy = 0.9590643274853801, F1-score = 0.944
Support Vector Machine: Accuracy = 0.9824561403508771, F1-score = 0.976
K-nearest Neighbors: Accuracy = 0.9590643274853801, F1-score =
0.943089430894309
Gradient Boosting: Accuracy = 0.9649122807017544, F1-score =
0.9516129032258064
```

As we can see, all models have very high accuracy, but Logistic regression has the best performance, which is very good, because it is an interpretable model and we can use it to solve 2 parts of my research problem. Now let's take a look at the ROC AUC graph for our models



From this graph we can see that all models have good quality, but Logistic Regression and SVM have the biggest area. What is more interesting, these models have equal accuracy scores, but different AUC scores. This fact shows that it is very important to use different metrics to assess your models to find the best one. Now let's take a look at the odds of our best model:

```
Odds for radius1: 1.4354955731168948      Odds for area1: 1.5201372569547507
Odds for texture1: 1.4418239501962937     Odds for smoothness1:
Odds for perimeter1:                      1.1997680816905911
1.3717690951300312
```

```
Odds for compactness1:
0.5342998748355042
Odds for concavity 1:
2.116727866479931
Odds for concave_points1:
3.0351530125221875
Odds for symmetry1: 0.8042626045543273
Odds for fractal dimension 1:
0.8712029879337999
Odds for radius2: 3.464412665506767
Odds for texture2: 0.8560557176821275
Odds for perimeter2:
1.8315942179456945
Odds for area2: 2.3902188545478955
Odds for smoothness2:
1.194085853569873
Odds for compactness 2:
0.5503034494922265
Odds for concavity2:
1.0844888315495387
Odds for concave_points2:
1.6079189223990002
Odds for symmetry2: 0.604526561266595
Odds for fractal_dimension2:
0.49738554520637335
Odds for radius3: 2.2565055565355774
Odds for texture3: 3.6252846172284006
Odds for perimeter3:
1.7057842434597679
Odds for area3: 2.185768564480711
Odds for smoothness3:
1.6559985014970922
Odds for compactness3:
0.8907436392593385
Odds for concavity3:
2.6563040294406055
Odds for concave_points3:
2.221891677757109
Odds for symmetry3: 3.3182876039532947
Odds for fractal_dimension3:
1.105985024436401
```

Interpretation of odds from the logit may be helpful to understand how every feature is connected with probability of breast cancer. We can see that with increase of variables concave_points1 (3.03), texture3 (3.625), concave_points3 (2.22) and symmetry 3 (3.318) increase the probability that it is malignant. Whereas increase of compactness1 (0.534) and symmetry2 (0.604) decrease probability that tumor will be malignant.

References

[1] UMUHOZA, Eric; NTIRUSHWAMABOKO, Dominique; AWUAH, Jane and BIRIR, Beatrice. Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa. SAIEE ARJ [online]. 2020, vol.111, n.3 [cited 2024-03-26], pp.95-101. Available from: <http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1991-16962020000300002&lng=en&nrm=iso>. ISSN 1991-1696.

[2] Sadrach Pierre, 6.07.2023, Mastering customer segmentation using credit card transaction data. Towards Data Science. https://towardsdatascience.com/mastering-customer-segmentation-using-credit-card-transaction-data-dc39a8465766

[3] Hossen, Sazzad. (2023). Medical Insurance Cost Prediction Using Machine Learning. 10.13140/RG.2.2.31456.25604.