Exercise 1: Classification

The four selected datasets are:

(1) Default of Credit Card Clients
(2) Abalone
(3) Loan
(4) Breast Cancer

	(1) Credit	(2) Abalone	(3) Loan	(4) Cancer
# Samples	large	small	large	small
# Attributes	high	low	high	low
Missing Value	no	no	no	no

Contents

1 Dataset Introduction

1.1 Dataset 1: Default of Credit Card Clients

Description: The dataset predicts the credit credibility of bank clients in terms of the estimated probability of default (not timely payed debt) depending on numerous factors of the social status and past payment behaviour (April to September 2005) of the person.

Characteristics: The dataset contains 30000 samples that record 23 attributes.

Attribute types: [X ... input, Y ... target, ID ... sample identifier]

Attribute	Description	Values	Range	Type
Y	default payment predict	(no, yes)	[0, 1]	nominal
ID	client identifier	unique integer	-	numeric
X1	given credit in dollars	positive integer	[10K, 1M]	numeric
X2	client gender	(male, female)	[1, 2]	nominal
X3	client education level	(graduate school, university, high school, others)	[0, 6]	nominal, arguably ordinal (scaling academic degree)
X4	client marital status	(married, single, others)	[0, 3]	nominal
X5	client age	positive integer	[21, 79]	numeric
X6-X11	past repayment delay	(duly paid, one month, ..., nine months+ delayed)	[-2, 8]	nominal, ratio (month as time unit)
X12-X17	bill statement in dollars (sum supposed to pay)	integer	[-339K, 1.67M]	numeric
X18-X23	past payment in dollars (sum actually payed)	positive integer	[0, 1.69M]	numeric

Source: `https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients`

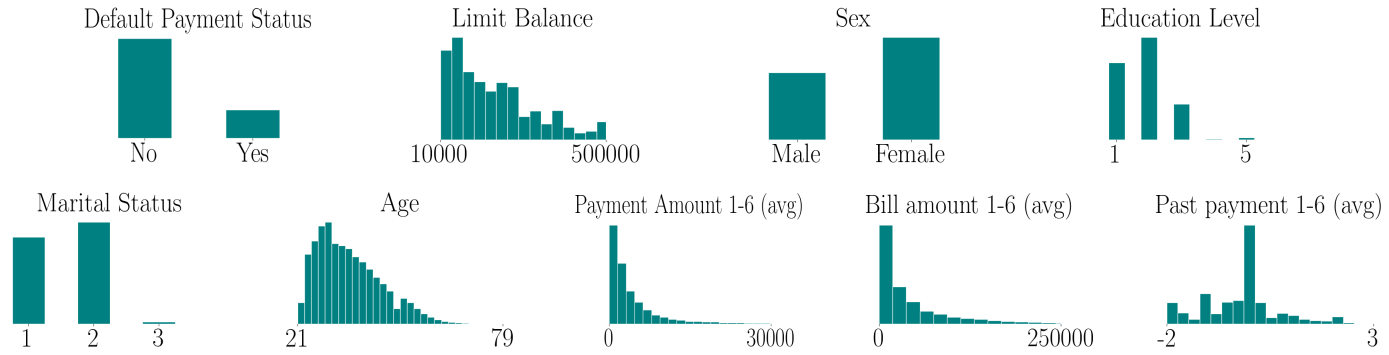

Figure 1: Graphical representations of the dataset **Default of Credit Card Clients**.

1.2 Dataset 2: Abalone

Description: The dataset predicts age of abalone using a variety of different measurements. The age is determined by counting the number of rings within the shell structure. Determining the age is a difficult process, for this reason this datasets aims to predict it, using some easier attainable characteristics, like measurements and weight.

Characteristics: The dataset contains 4177 samples that record 8 attributes.

Attribute types: [X . . . input, Y . . . target]

Attribute	Description	Values	Range	Type
Y	number of rings (used for age)	(no, yes)	[1,29]	numeric
Sex	abalone sex	(male, female, infant)	[M,F, I]	nominal
Length	longest shell measurement in mm	positive float	[0.075,0.815]	numeric
Diameter	diamater perpendicular to length in mm	positive float	[0.055,0.65]	numeric
Height	height with meat in shell in mm	positive float	[0.0,1.13]	numeric
Whole weight	weight of whole abalone in grams	positive float	[0.02,2.8255]	numeric
Shucked weight	weight of meat in grams	positive float	[0.001,1.488]	numeric
Viscera weight	weight of gut in grams	positive float	[0.0005,0.76]	numeric
Shell weight	weight of shell after being dried in grams	positive float	[0.0015,1.005]	numeric

Pre-processing:

Source: `https://archive.ics.uci.edu/dataset/1/abalone`

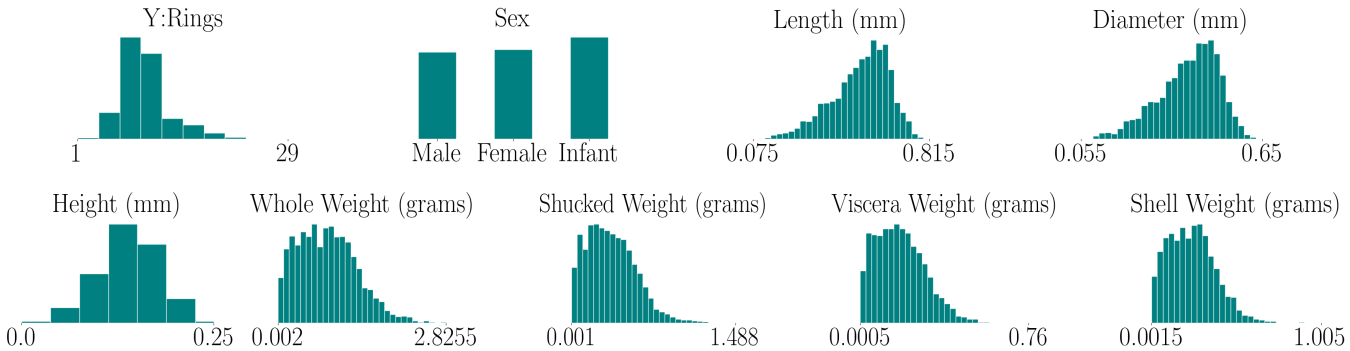

Figure 2: Graphical representations of the dataset **Abalone**.

1.3 Dataset 3: Loan

Description: This dataset aims to predict the score of a loan application. The Score is evaluated in the grades A to G and is predicted on the basis of a variety of features, including factors like the amount of the loan, the interest rate, the annual income and more.

Characteristics: The dataset contains 10000 samples that record 90 attributes.

Attribute types: [X ... input, Y ... target, ID ... sample identifier]

Attribute	Description	Values	Range	Type
Y	grade	(A, B, C, D, E, F, G)	[A,G]	nominal
ID	client identifier	unique integer	-	numeric
loan amnt	amount of given loan	positive float	[1K, 40K]	numeric
funded amnt	amount of given fund	positive float	[1K, 40K]	numeric
term	time of loan repayment	(36 months, 62 months)	[36 months, 72 months]	nominal
int rate	amount of interest per period	positive float	[5.31, 31]	numeric
installment	installment amount	positive float	[30.1, 1.72k]	numeric
total payment	total loan payment	positive float	[61, 62k]	numeric
annual inc	client yearly income	positive integer	[5k, 3.2M]	numeric
...	...	...	...	...

Pre-processing:

Source: `https://www.kaggle.com/competitions/184-702-tu-ml-ws-24-loan`

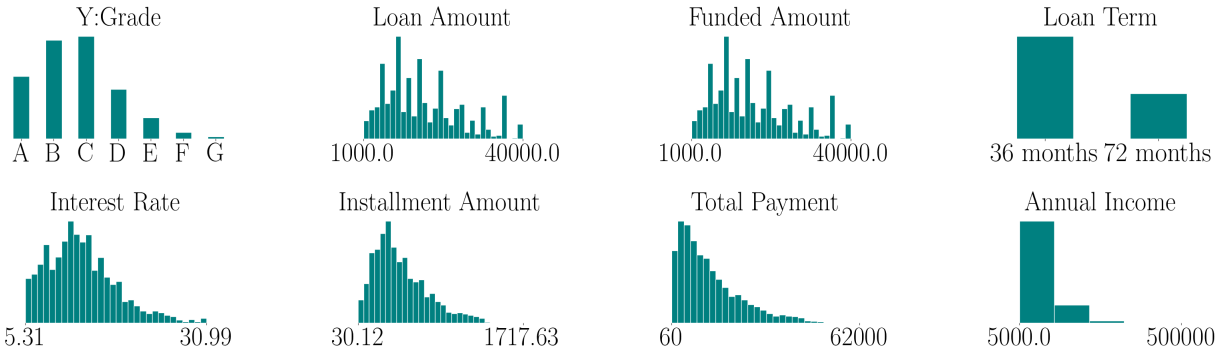

Figure 3: Graphical representations of the dataset **Loan**.

1.4 Dataset 4: Breast Cancer

Description: This datasets aims to predict if a patient has a breast cancer recurrence event or not. It tries to predict, by analyzing properties of the cancer cell, like radius, dimension, texture and many more.

Characteristics: The dataset contains ? samples that record ? attributes.

Attribute types: [X . . . input, Y . . . target]

Attribute	Description	Values	Range	Type
Y	class	(true, false)	[true,false]	bool
radiusMean	average radius of cell	positive float	[7.69, 25.7]	numeric
textureMean	average texture	positive float	[9.71,39.3]	numeric
perimeterMean	average perimeter of cell	positive float	[48,174]	numeric
areaMean	average area	positive float	[170,2.01k]	numeric
radiusWorst	worst radius	positive float	[8.68,33.1]	numeric
textureWorst	worst texture	positive float	[12,44.9]	numeric
perimeterWorst	worst perimeter	positive float	[54.5,229]	numeric
...	...	...	...	...

Pre-processing:

Source: `https://www.kaggle.com/competitions/184-702-tu-ml-ws-24-breast-cancer`

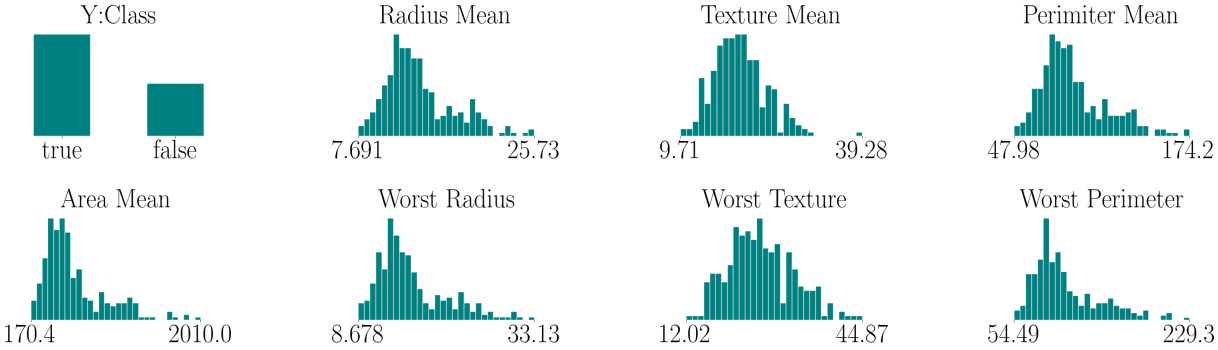

Figure 4: Graphical representations of the dataset **Loan**.

2 Dataset Characteristics and Preprocessing

Each of the chosen datasets, is different in a variety of factors. For this exact reason, the different datasets need to be analyzed and preprocessed before fitting them into a model.

For this step the values of each dataset were first seperated into numerical and categorical data. Afterwards each column was plotted as either a histogram (for numerical data) or a barplot. In this way unique dataset patterns were put to display. For example in the following histogram of the two features loan_amnt and funded_amnt it can clearly be seen that the distribution of data is exactly the same. This most likely means that the one of these features does not need to be included in the final classification model.

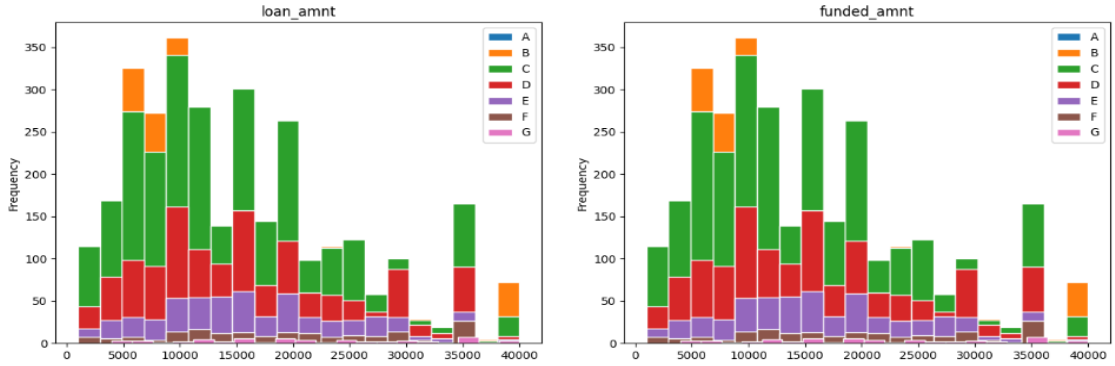

Figure 5: **loan_amnt** and **funded_amnt** in comparison

To make this step easier and see the releationship between different features, a correlation matrix was produced and analyzed. For this step, the categorical data was transformed in to a binary format. So by using a hot encoding method. The following graphic represents a correlation matrix of the abalone dataset. Intuitively it makes sense, that variables like whole_weight would be highly correlated to the other remaining weight variables. The same is the case for the variables length and diameter. A abalone with a longer length, will have a longer diamater perpendicular to it, as the data represents. This means one of the values can probably be dropped without reducing the success of the final model.

	length	diameter	height	whole_weight	shucked_weight	viscera_weight	shell_weight	sex_F	sex_I	sex_M	rings_young	rings_middle	rings_old
length													
diameter	0.99												
height	0.83	0.83											
whole_weight	0.93	0.93	0.82										
shucked_weight	0.90	0.89	0.77	0.97									
viscera_weight	0.90	0.90	0.80	0.97	0.93								
shell_weight	0.90	0.91	0.82	0.96	0.88	0.91							
sex_F	0.31	0.32	0.30	0.30	0.26	0.31	0.31						
sex_I	-0.55	-0.56	-0.52	-0.56	-0.52	-0.56	-0.55	-0.46					
sex_M	0.24	0.24	0.22	0.25	0.25	0.24	0.24	-0.51	-0.52				
rings_young	-0.61	-0.62	-0.56	-0.58	-0.52	-0.57	-0.60	-0.28	0.51	-0.23			
rings_middle	0.21	0.20	0.14	0.14	0.17	0.15	0.10	0.08	-0.18	0.09	-0.49		
rings_old	0.41	0.42	0.42	0.44	0.34	0.42	0.50	0.20	-0.33	0.13	-0.52	-0.50	

Figure 6: Correlation matrix of **abalone**

These steps were concluded for each dataset, resulting in the following preprocessing decisions.:

1. For default credit, the columns **BILL_AMT2** (Directly related to BILL_AMT1 and BILL_AMT3) and **BILL_AMT5** (Directly related to BILL_AMT4 and BILL_AMT6) were dropped.
2. For abalone as desribed earlier **diameter** and **whole_weight** were dropped.
3. Loan contains a variety of features including highly correleated ones. First of all, there are many features that have an inverted version of themselfes included in the set. This means they correlate with each other, therefore all the inverted variables were dropped. Furthemore the following features were dropped because they are correlated to the features which are mentioned in parantheses. **funded_amnt** (loan_amount, installment), **fico_range_high** (fico_range_low), **total_rec_prncp** (total_pymnt), **collection_recovery_fee** (recoveries), **num_rev_tl_bal_gt_0** (num_actv_rev_tl), **num_sats** (open_acc), **tot_hi_cred_lim** (tot_cur_bal). Furthermore **policy_code** was dropped, simply because it just contains an unrelated constant feature.

4. In cancer features that contain "Worst" were dropped, beacause they were directly related to features that contain "Mean". The same reasoning was made for features that contain "perimeter" or "area" because they are related to features that contain "radius", aswell as features that contain "concave" or "concavity" since they relate to features that contain "compactness".

Even though the preprocessing steps, that were concluded might be benifitial for the results of the classification models, in some cases the result might improve without acutally dropping any values. For this reason the models were tested with and without preprocessing.

3 Modeling Steps / Pipeline

To standardize the process of creating an appropriate model for the four datasets, a pipeline was constructed. The pipeline is displayed in the following diagram and consists of six main steps, which will be described in detail in this section.

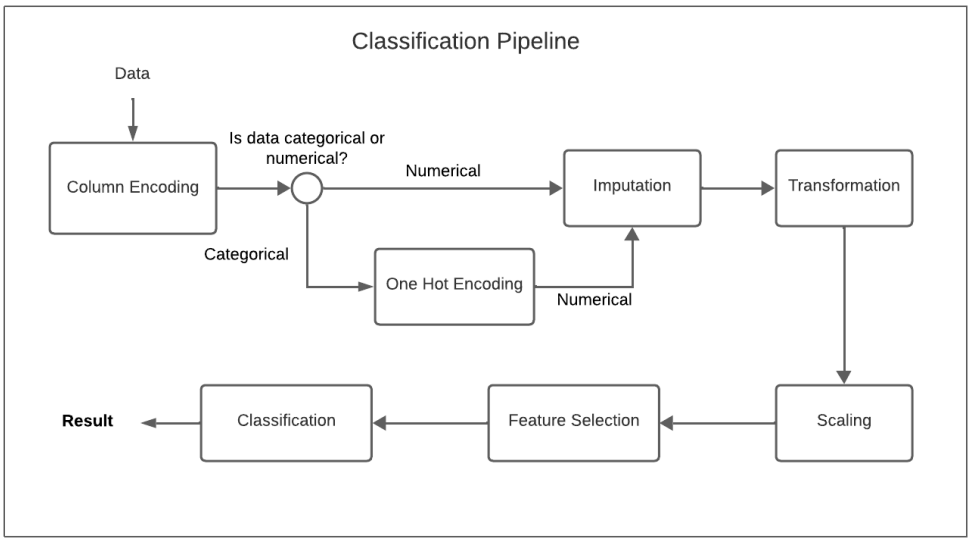

Figure 7: The classification pipeline

3.1 Column Encoding

The first step has to do with something, which was already explained in the preprocessing section. The transformation of categorical data. For the analysis preprocessed data with removed columns was used in the model, but data without preprocessing was also tested. For this reason the pipeline needed to handle categorical data, since this data was not tranformed before. For this task a ColumnTranformer is used, which checks wether a column is numerical and does nothing, if it is, or wether it is categorical in which case it is hot encoded, just like previously explained.

3.2 Imputation

The pipeline can use a SimpleImputer to handle missing data, deciding between strategies like "mean" if there are no outliers, or median if there are. In the case of the analyzed datasets, there are no missing values, so in this case there is no imputation.

3.3 Transformation

In the tranformation section the data is tranformed if it improves the result of the classifier, if not the data just stays the same. The Pipeline has the ability to use two different transformation functions.

1. Log Tranformation - This is used to compress large values and reduce skewness in data distribution. This works well with positive, right skewed data, which is common in financial data.
2. Yeo-Johnson Power Tranformation - Uses a power based transformation to make the data, more Gaussian like. For non-normally distributed data which needs to be made symmetric.

3.4 Scaling

To balance the impact of each feature, scaling is often necessary for machine learning models. For this reason the pipeline offers the option to scale the data using three different scaling types.

1. Standard-Scaling - Scales data to have a mean of 0 and a standard deviation of 1. Works better if the data is close to a normal distribution.
2. MinMax-Scaling - Scales the data to fit within a specified range for example [0,1]. Works well if data is not close to normal distribution. Preserves original distribution.
3. Robust-Scaling - Scales between the range o fthe 25th and 75th percentile. Useful when the data contains outliers, since its robust to them.

3.5 Feature Selection

Some datasets contain a high amount of features and for this is not suitable for every classifier. Thats why the pipeline includes some algorithms for selecting only specific features.

1. Principal Component Analysis (PCA) - Reduces the dimensionality of the dataset while retaining as much variance as possible, especially useful when features are correlated.
2. SelectKBest - Calculates feature importance and uses the top K amount of featuers in the classification.

3.6 Classification

The final part of the pipeline consists of the classifcation. The pipeline has the ability to choose 11 differnet classifiers.

1. Logistic Regression
2. Passive Aggressive
3. SGD
4. k-NN
5. Decision Tree
6. Random Forest
7. SVC
8. Gaussian NB
9. MLP
10. AdaBoost
11. QDA

It can also use Hyperparameter-Tuning to improve the performance of the classifiers even more.

3.7 Splitting Training/Test data

One additional modeling step which as taken, but is not necessarily a part of the pipeline is the splitting of the data. Since the datesets vary widely in their feature amount and distribution some variations for test data perperations are in order. For this reason we distinguish between small and big datasets. For small datasets the test size is equal to 1/3 of all data, while for the big datasets (datasets with more than 10000 instances) the test size is equal to 1/10. Furthermore since in some cases the target category might be uneven, stratified folds and stratified train/test splits are used so that the target distribution stays the same for training.

4 Experimental Results

4.1 Performace Measures

To appropiately interpretate the results the following performance measures, were used to determined wether a classfier performed well.

1. Accuracy - Measures the proportion of correct predictions out of all predictions.
2. Recall - Measures the ability to identify true positives. Very important in a dataset like cancer, where missing a positive case (false negative) can have severe consequences.

3. Precision - Measures how many instances predicted as positive asre actually positive
4. F1- Score - Combination of Recall and Precision. Our primary scoring metric. A high score indicates identifying positives correctly and minimizing false positives. For multi-class problems, weighted F1 is used by weighing each class according to its proportion.

4.2 Selected Classifiers

To increase efficiency and decrease the run-time, in this project the eleven above classifiers were used without any particular tuning. The result were compared and three classifiers were selected across all datasets. These three classifiers were than tested with many different pipeline configurations.

The following diagrams display the F1 Score of the eleven classifiers, for the cancer and loan dataset. The general performance without any adjustments seems to be better for the cancer dataset. This might be because loan is a much bigger and complexer dataset therefore needing more adjustments via the pipeline, in order to produce satisfying results. At this moment it contains far too many features. We can see however that the boxplots in loan are much more contained and more importantly, there is one classifier with a quite good performance, reaching a f1-score of 0.987. The Decision Tree classifier works excellent with large datasets, with many features. For cancer the logistic regression and Support Vector Classification works quite well this is because it does not have many irrelevant features and is relatively small.

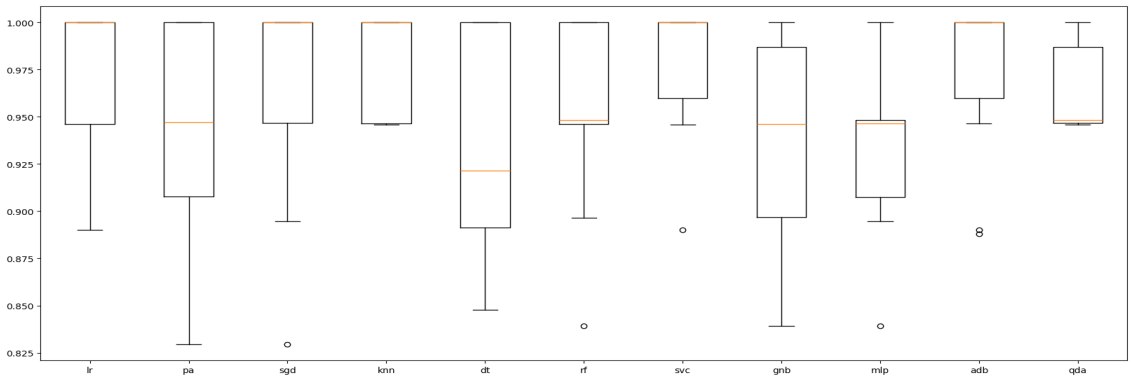

Figure 8: Results of 11 classifiers in dataset **cancer**

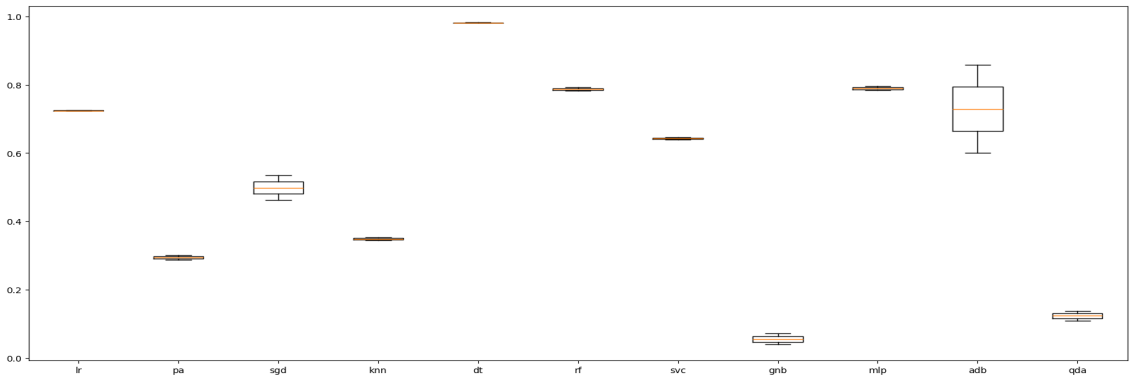

Figure 9: Results of 11 classifiers in dataset **loan**

Because of these performances, it was decided that these three classifiers should be used for further preperation of the datasets.:

1. LogisticRegression (LR) - Models relationship between input features and class probability using a logistic function. Its a simple regression alogorithm, that is good for datasets with independent features and few perfect corelleations. Source: `https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html`
2. C-Support Vector Classification (SVC) - This classifier works via a 2d space, it gathers the observation which are closest to each other while not belonging to the same class, to generate boundaries to gather all observation of each class within one region. This classifier is robust to overfitting and quite flexible, but it can be quite computationally expensive for large datasets.
   Source: `https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html`

3. DecisionTreeClassifier (DT) - Decision Tree splits the dataset into subsets, basesd on the values of the features. For this purpose it creates a trees tructure, where each node represent a decision. It works well with large datasets, and does not require scaling nor normalization. But it is prone to overfitting
   Source: `https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html`

4.3 Selected Classifier Results

4.3.1 Pipeline Settings

After selection of the three classifiers the following pipeline assignments were concluded. Note that these configurations do not necessarily lead to the best possible results. Since finding the optimal solution takes a lot of runtime, thats the reason why Randomized Cross Validation was used, which does not always lead to the best results.

In the following graphic the selection of the attributes of the LogicalRegression classifier is displayed. First all the models go through encoding, afterwards all of them passtrough the imputation step, since there are no missing values in the data. Interestingly the loan dataset is the only dataset, which does not do a column transformation. This may be because it consists of a variety of different values which may need to be processed individually. All of the datasets do scaling, which makes sense because LRM responds well to normal scaling. Finally only the k-best features are selected for each dataset aside from abalone. For cancer and loan they are each specified respectively with k=19 and k=43. For abalone choosing the best k might not improve performace since it has quite few features anyway. Finally the logistic regression is called with a variety of attributes chosen by hypertuning.

Figure 10: Pipeline of **LogisticRegression classifier**.

The models for the other classifiers were constructed in the same way, giving the following results. Imputation is not included in this table since there are no missing values anyway.

Since all datasets contain categorical data of some form, it makes sense that every pipeline needs encoding. SVC benefits from scaling so all the datasets use some sort of scaling. For the bigger sets there also needs to be a reduction of features,

whether through PCA or SelectKBest. The DTC pipeline needs the least amount of processing, this could be because it does not need any scaling. Interestingly the pipeline still includes feature selection for the datasets, even though it can handle big feature amounts. As further discussed in the results, this acutally might have some negative impacts, since the performance of DTC for example for the loan dataset, is actually worse using this pipeline, then the calculation without any processing, which was produced in the previous chapter.

Dataset	**Encoding**	**Transformation**	**Scaler**	**Feature Selector**	**Model**
Default Credit	OneHotEncoder	FunctionTransformer	StandardScaler	SelectKBest	SVC
Abalone	OneHotEncoder	FunctionTransformer	RobustScaler	None	SVC
Loan	OneHotEncoder	FunctionTransformer	RobustScaler	PCA (n_components = 13)	SVC
Breast Cancer	OneHotEncoder	None	StandardScaler	SelectKBest (k=25)	SVC

Table 1: SVC Pipelines example

Dataset	**Encoding**	**Transformation**	**Scaler**	**Feature Selector**	**Model**
Default Credit	OneHotEncoder	FunctionTransformer	StandardScaler	SelectKBest	DTC
Abalone	OneHotEncoder	FunctionTransformer	None	None	DTC
Loan	OneHotEncoder	None	None	SelectKBest (k=43)	DTC
Breast Cancer	OneHotEncoder	None	None	SelectKBest (k=19)	DTC

Table 2: DTC Pipelines

4.3.2 Results

Default Credit Dataset: Performance was moderate, with SVC achieving the highest accuracy (82.23%). The lack of critical feature information limited classification effectiveness, emphasizing the need for more comprehensive data.

Abalone Dataset: Performance was suboptimal for all classifiers, with SVC performing slightly better (62.96%) due to its ability to handle simpler patterns. Highly correlated features and limited data informativeness constrained classification potential. External factors like weather patterns and food availability might improve classification accuracy.

Loan Dataset: Decision Tree excelled, achieving the highest accuracy (96.90%), showcasing its robustness with large, complex datasets. Logistic Regression performed adequately (87.30%), while SVC failed to generalize effectively (37.00%).

Breast Cancer Dataset: SVC and Logistic Regression demonstrated excellent performance with near-identical accuracy and F1-scores (97.89%). Decision Tree lagged slightly (91.58%).

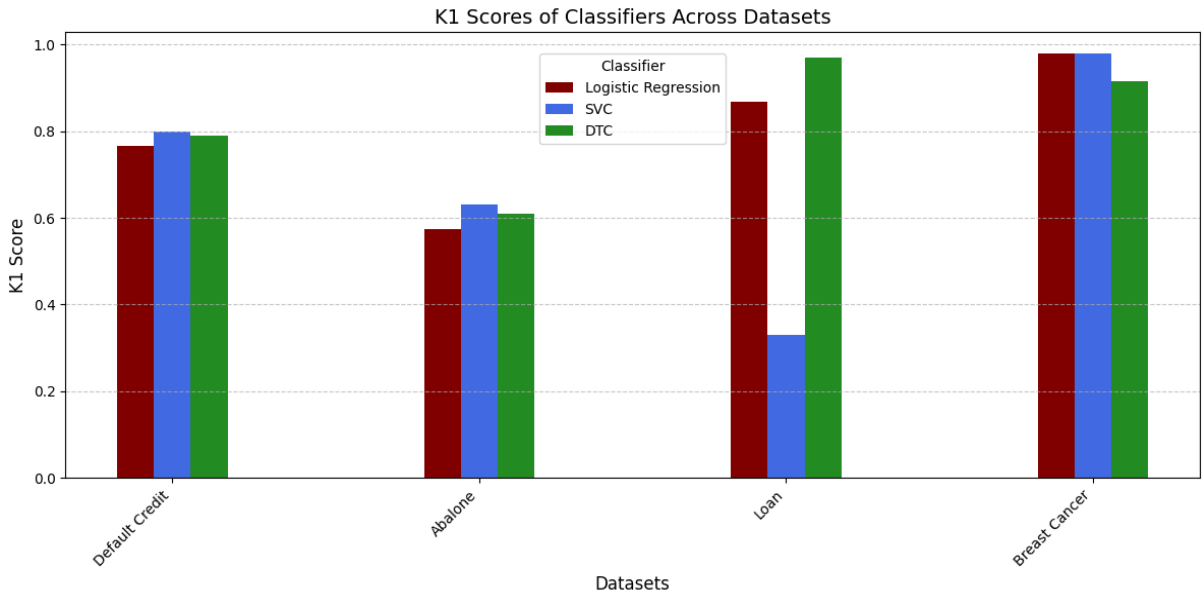

Figure 11: K1 scores of classifiers after modifying with pipeline.

Dataset	Classifier	Accuracy	F1-Score
Default Credit	Logistic Regression	0.8047	0.7650
	SVC	0.8223	0.7989
	Decision Tree	0.7870	0.7897
Abalone	Logistic Regression	0.5887	0.5751
	SVC	0.6296	0.6314
	Decision Tree	0.6116	0.6105
Loan	Logistic Regression	0.8730	0.8683
	SVC	0.3700	0.3303
	Decision Tree	0.9690	0.9690
Breast Cancer	Logistic Regression	0.9789	0.9789
	SVC	0.9789	0.9789
	Decision Tree	0.9158	0.9151

Table 3: Classifier Results Across Datasets (Accuracy and F1-Score)

5 Summary/Conclustion

In this project, we applied three classifiers—Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Tree Classifier (DTC)—to four datasets: Default Credit, Abalone, Loan, and Breast Cancer. These classifiers were selected after an initial assessment of multiple algorithms to ensure diversity in learning paradigms, best performance, and suitability for the given datasets.

We designed dataset-specific preprocessing pipelines to address challenges such as redundant features, scaling requirements, and categorical data encoding. Through systematic parameter tuning, we optimized the performance of each classifier and compared their effectiveness under consistent conditions across the datasets.

Our results showed that SVC performed best on clean and structured datasets like Breast Cancer, while Decision Tree achieved the highest performance on complex datasets like Loan. Logistic Regression demonstrated consistent and reliable performance across all datasets. The Abalone dataset posed significant challenges due to highly correlated features and weak target relationships, while Default Credit highlighted the impact of missing information on classification accuracy.