Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Predicting football matches with the use of weather conditions

Wessel van Putten

Supervisors:
dr. Yingjie Fan & dr. Arno Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

21/02/2023

Abstract

Football can earn both clubs and their fans a lot of money. The ability to make accurate predictions of upcoming matches can increase these earnings. Clubs can change their tactics when a bad result is about to happen and gamblers can make better predictions. To get accurate predictions, a model that takes both weather conditions and details about the two teams in account has been constructed. The used data consists of data about the games and the playing teams, weather data and player ratings. This data has been prepared and analysed with Random Forest Classification, classifying the games into three categories namely a home team win (w), a home team loss (l) and a draw (d). The predictions have an accuracy of more than 60%. This can help gamblers in getting more profitable bets, and teams in preparing better towards their next match.

Contents

1 Introduction

1.1 The situation

Football is the biggest sport in the world, with estimated around 4 billion followers worldwide [Ten22]. This might be due to its unpredictability. Every football fan knows the stories of the small teams unexpectedly beating the bigger team, but are these moments really unexpected or are there a lot of ways to know what will happen?

Field football is played outside, meaning that different weather conditions occur during a season or even during a match. Some stadiums neutralise the conditions inside the stadium by building a closable roof on the stadium [AT522] or installing airconditioning inside the stadium [VTB19], but this is an exception and most stadiums do not have these facilities. That means that weather can influence the game and maybe change the result of the game. For example, on a sunny day, the famous miss of Jurrie Koolhof would be very likely to be a goal, instead of the ball stranding in the mud [Hes19].

Since the result of a game may result in a club earning millions of euros, being prepared for what will happen is very important. If it is expected that a team will lose the upcoming match, the coach should choose to play a different tactic or use players that are more familiar with the expected weather conditions. On the other hand, gamblers can increase their wins when they are more certain of what the result of the game will be. Both clubs and supporters can have advantages of good predictions. Therefore, the research question of this thesis is:

How do weather conditions influence the result of a sportmatch and can we better predict the result of a match with the help of the weather conditions?

1.2 Thesis overview

This bachelor thesis is supervised by dr. Yingjie Fan and dr. Arno Knobbe and is written for Leiden Institute of Advanced Computer Science (LIACS).
This thesis contains six chapters. This chapter contains the introduction of the thesis. In Section 2, the related work can be found. Section 3 is all about the used data and how the data is processed. The Research method is described in section 4. The results of the experiments can be found in Section 5. Lastly, Section 6 contains the conclusion and further research.

2 Related Work

This section contains a short summary of previous work. It includes used data and results of what has been researched earlier about this topic.
In previous research, it has been shown that when the temperature rises by around 10 degrees Celsius, there are 0,48 more goals scored [Bra15]. This might be because players are more likely to be exhausted earlier in the match and exhausted players make more mistakes. It is also shown that national football teams from the Gulf Region, like Qatar or Bahrain, where there is a dessert climate, have a bigger home advantage than teams that play in other climates [BGFM15]. Ahead of the 2010 World Cup in South Africa, it was stated that teams should not lack any preparation on the expected weather conditions if they wanted to have higher winning chances [SD10]. Research has been done on environmental factors that influence physical activity in the German Bundesliga. Temperature was the only factor that affected the physical activity of players [CLA+21]. To reduce the effect of heat, FIFA rules tell the referee to apply a cooling brake when the temperature is 32 degrees Celcius or higher [Wil11]. Other research shows that adolescents are less active on rainy and snowy days [BGDO+09].
Besides direct effects, weather can indirectly influence performance as well. On a sunny day, there will probably be more fans, which increases the home advantage. More fans also gains a club more revenue, which can on the long term give them an advantage on teams who attract fewer fans due to poorer weather conditions [Tho77]. Figure 1 shows what effects weather conditions can have on both indoor and outdoor sports.

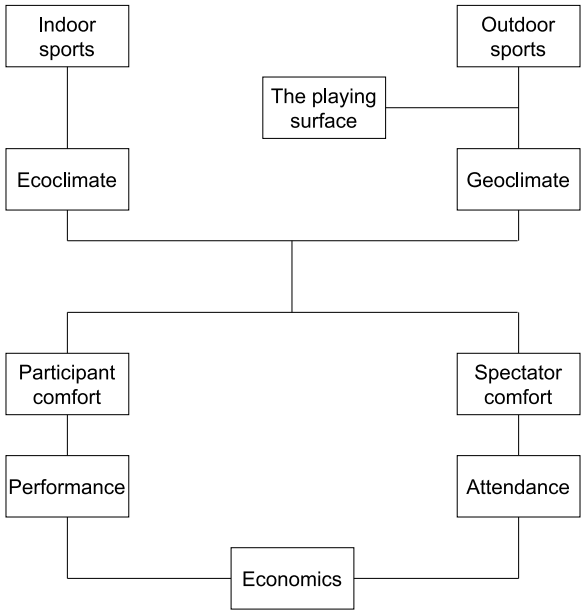

Figure 1: The effects of weather on sports [Tho77]

Some research has been done on predicting sports matches already, but most of them only focus on variables that describe the current form of the two teams that are playing and the recent history between those teams. It is stated that it is necessary to have a big amount of data about the players, the teams and match events to make an accurate prediction about the result [FMT19].

Some important features in earlier research are passing attempts, number of sprints and turnovers [LJ14]. Player attributes, goals, corners, managers' performance and recent form are said to be important features as well [FMT19]. Playing at home provides an advantage as well [LPLB11]. Research on the Spanish competition concluded that a combination of similar features and weather data can be useful to make predictions [IRPTO20].
The most used models for making predictions are Bayesian Networks, K-nearest neighbours, decision trees [JFN06], bagging regressor [KTS17] and Support Vector Machine Regression [IRPTO20] [KG21].

There is some prior research on predicting football matches, but they only focus on football-related attributes, like goals scored in previous matches, position of the teams etcetera. There are also some studies on the influence of weather on sports. Some of them focus on the result of a sportsgame, but most of those studies are related to the health of the player. Something that has not been studied before, is combining the weather conditions with features that are commonly used in predicting football. That combination is what is going to be researched in this study. This is relevant for gamblers, as they can increase their wins because they can be more certain that they are right. Sport clubs can profit from this study as well, since they can decide to change their tactics when it is likely that they will lose with their planned line-up, they can decide to not close the roof above the stadium if it turns out that they are more likely to win when it has rained. Clubs might also use this study when they are willing to buy players. If a players loses a lot in the climate of a club, they should not buy that player.

3 Data

3.1 Data Collection

For this thesis, the following data has been used:

- Football data from transfermarkt.com. Transfermarkt.com is an originally German website which keeps track of a lot of football statistics. The data is downloaded from kaggle.com[1]. The dataset contains six csv files. See Figure 2. These files contain data about the biggest football competions, matches played by the clubs in that competition, information about the clubs, information about the players in the competition, their transfervalue and the matches played by individual players. The model was built on the Dutch Eredivisie, so this data was filtered on this league.
- Player ratings from the game FIFA 23[2]. FIFA 23 is a football game that contains very detailed player stats. This data is downloaded from Kaggle as well.
- Weather data from the Royal Dutch Meteorological Institute (KNMI)[3]. Since the model is created based on the Dutch Eredivisie, the weather data is from The Netherlands. There are multiple weather stations located in The Netherlands, which are close to football stadiums. The data contains a lot of variables like, wind direction, wind speed, temperature, how sunny it was, rain, clearness of vision and humidity.

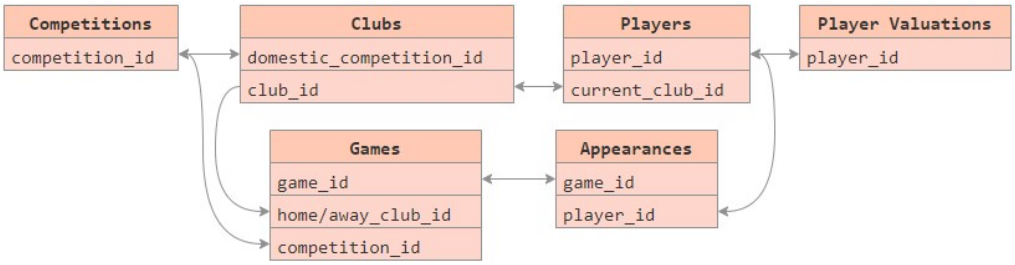

Figure 2: Layout of the transfermarket.com dataset

3.2 Data preparation

There is a lot of data about football matches, players and the weather. This means that it can be downloaded easily and it is not necessary to create it. The data that will be used can be found in Section 3.1.
The data will be handled through Python and will be loaded into a Pandas DataFrame. From there, the data will be prepared to be ready for a classification analysis by joining different CSV files and

[1] https://www.kaggle.com/datasets/davidcariboo/player-scores
[2] https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database
[3] https://www.knmi.nl/nederland-nu/klimatologie/daggegevens

changing categorical data to another format.

From the `games.csv` file, the columns containing the referee, stadium, managers, competition type and competition round were replaced by ID's because they were written as strings. The rows with ID's were then changed to columns containing booleans. This is called One Hot Encoding [Dat20]. Columns that contained strings but from where there was already an ID about that string like club names were deleted. This resulted in the `games.csv` file only containing integers.
The code below shows how the strings were changed to ID's

```
ref=games['referee'].to_frame()
ref=ref.drop_duplicates(subset=['referee'])
ref['ref_id']= ref.index+1
ref=ref.set_index('ref_id')
games['referee'] = games['referee'].map(ref.set_index('referee')['ref_id'])
```

In the code above, it is shown how the name of the referee is changed to an ID. This is done for the competition, stadium, managers, round and competition type. The date in the games file was made ordinal such that it can be used to make predictions. After the strings are changed to ID, the column containing the winner of the game is added, where 'w' means home win, 'l' means home loss and 'd' means draw. See the code below:

```
for index, row in games.iterrows():
    games['goal_diff'] = games['home_club_goals'] - games['away_club_goals']
games['toto_score'] = np.where((games['home_club_goals'] < games['
    away_club_goals']), 'l', 'w')
for index, row in games.iterrows():
    if games.loc[index]['goal_diff']==0:
        games.at[index,'toto_score'] = 'd'
return 'toto_score is added'
```

After these steps, the new game file is saved as a CSV file. The next step is to combine the players data. This is done by simply merging these two files on the last name of each player. Some last names appeared multiple times like for the twins Quinten Timber and Jurrien Timber. This means that both players appeared two times in the merged file. This issue was resolved by removing the wrong lines using Microsoft Excel. There were some playernames that were written differently by FIFA and Transfermarket, like 'Ould Chikh' (Transfermarket) and 'Ould-Chikh' (FIFA). This was manually resolved in Microsoft Excel as well. Adding the weather data to the games file was done by assigning each stadium to its nearest weather station, as described in Section 3.2. Then, the games file containing the nearest weather station for each stadium, was merged with the weather file on both the date and the station ID. The last step was adding the players, their FIFA rating and their nationality to the games file:

```
for index, row in games.iterrows():
    dwf=appearence.loc[appearence['game_id'].str.contains(str(row['game_id'
        ]), case=False)]
    home=row['home_club_id']
    away=row['away_club_id']
    dwf_home=dwf.loc[dwf['player_club_id'].str.contains(str(home), case=
        False)]
```

```
    dwf_away=dwf.loc[dwf['player_club_id'].str.contains(str(away), case=
        False)]
    i=1
    for dex, row in dwf_home.iterrows():
        games.at[index, 'home_player_{}'.format(i)]=int(row['player_id'])
        i+=1
    i=1
    for dex, row in dwf_away.iterrows():
        games.at[index, 'away_player_{}'.format(i)]=int(row['player_id'])
        i+=1
i=1
while i<17:
    games['rating_home_player_{}'.format(i)]= games['home_player_{}'.format(
        i)].map(players.set_index('player_id')['overall'])
    games['rating_away_player_{}'.format(i)]= games['away_player_{}'.format(
        i)].map(players.set_index('player_id')['overall'])
    games['country_home_player_{}'.format(i)]= games['home_player_{}'.format
        (i)].map(players.set_index('player_id')['country_of_citizenship'])
    games['country_away_player_{}'.format(i)]= games['away_player_{}'.format
        (i)].map(players.set_index('player_id')['country_of_citizenship'])
    i+=1
countries=players["country_of_citizenship"].values.tolist()
dichome={}
dicaway={}
for i in countries:
    dichome['home_{}'.format(i)]=0
    dicaway['away_{}'.format(i)]=0
games.join(pd.DataFrame(dichome,index=[0]), lsuffix='_left', rsuffix='_right
    ')
games.join(pd.DataFrame(dicaway,index=[0]), lsuffix='_left', rsuffix='_right
    ')
for index, row in games.iterrows():
    dichome={}
    dicaway={}
    home_players=[]
    away_players=[]
    i=1
    while i<=17:
        home_players.append(row['country_home_player_{}'.format(i)])
        away_players.append(games.at[index, 'country_away_player_{}'.format(
            i)])
        i+=1
    for i in home_players:
        dichome['home_{}'.format(i)]=home_players.count(i)
    for i in away_players:
        dicaway['away_{}'.format(i)]=away_players.count(i)
    for i in dichome:
        games.at[index, i]=int(dichome[i])
```

```
    for i in dicaway:
        games.at[index, i]=int(dicaway[i])
    i=1
while i<=17:
    games = games.drop(['country_home_player_{}'.format(i), '
        country_away_player_{}'.format(i)],axis=1)
    i+=1
```

Right before making predictions, some columns that contained ID's are changed into dummies with the use of the `pd.get_dummies()` function [Dat20]. In Section 3.3, the process of checking which features improved the model is shown. The columns of the features that did not improve the model were removed from the Dataframe. The Dataframe has been splitted into a test and training set with the line: `train, test = train_test_split(games, test_size=0.2, random_state=0)`. Then, both the train and test set were splitted into the 'x-part' containing the values that are used to predict and into the 'y-part' containing the value that is predicted, the toto_score in this case. Lastly, a classifier is used to classify the results. The quality of the model is evaluated by using the function `accuracy_score()`.
Each stadium has been linked to its nearest weather station as can be seen for some stadiums in Table 1. By making use of the date of the measured weather, the date of the game and the weather station that is close to the football stadium, the weather for every match has been added to `games.csv`.

stadium_id	Stadium	Weatherstation_id
3	Johan Cruijff ArenA	240
6	Sportpark Ter Specke	210
15	Goffertstadion	375
16	Stadion "Galgenwaard"	260
19	MAC^{3}PARK stadion	278
20	Stadion Feyenoord "De Kuip"	344
24	Euroborg	280

Table 1: Some stadiums and the id of the nearest weather station

3.3 Features

The features that could be used in the model will be explained below. The features are split in game features, player features and weather features. Section 3.3.1 will discuss the Game features, Section 3.3.2 will cover the features about the players and lastly, Section 3.3.3 will discover the features about the weather conditions.

3.3.1 Game Features

The features that are used and contain information about the game are displayed in Table 2.

3.3.2 Player features

There are multiple player-specific features. These are among other things the ID's of the players that appeared in the matches, their FIFA rating, the average team rating and the number of players

Feature name	Description
Competition id	Contains an ID about the league of the match, like Eredivisie, Premier League or KNVB Cup
Competition type	Contains an ID about the type of league, like cup or domestic league
Season	Contains the year of the season
Round	Contains an ID about the round of the match like semi-final, final or 12th matchday
Date	Contains the date of the match
Home club id	Contains the ID of the home club
Away club id	Contains the ID of the away club
Home club position	Contains the position of the home club ahead of the game
Away club position	Contains the position of the away club ahead of the game
Home club manager	Contains the ID of the home club manager
Away club manager	Contains the ID of the away club manager
Attendance	Contains the attendance of the match
Referee	Contains the ID of the referee
Stadium id	Contains the ID of the stadium the match is played in

Table 2: The used features about the match and their description

per nationality in that team. Other features that might be used are, length, weight, contract time, market value or salary.

3.3.3 Weather features

The data about the weather contains a lot of columns with information about the weather conditions. These have all been tested. The features that improved the quality of the model can bee seen in Table 3. Speeds are measured in 0.1 meters per second, temperatures are measures in 0.1 degrees Celsius and time is measured in 0.1 hours (6 minutes). Figure 3 shows some features and their possible impact on the result of the game.

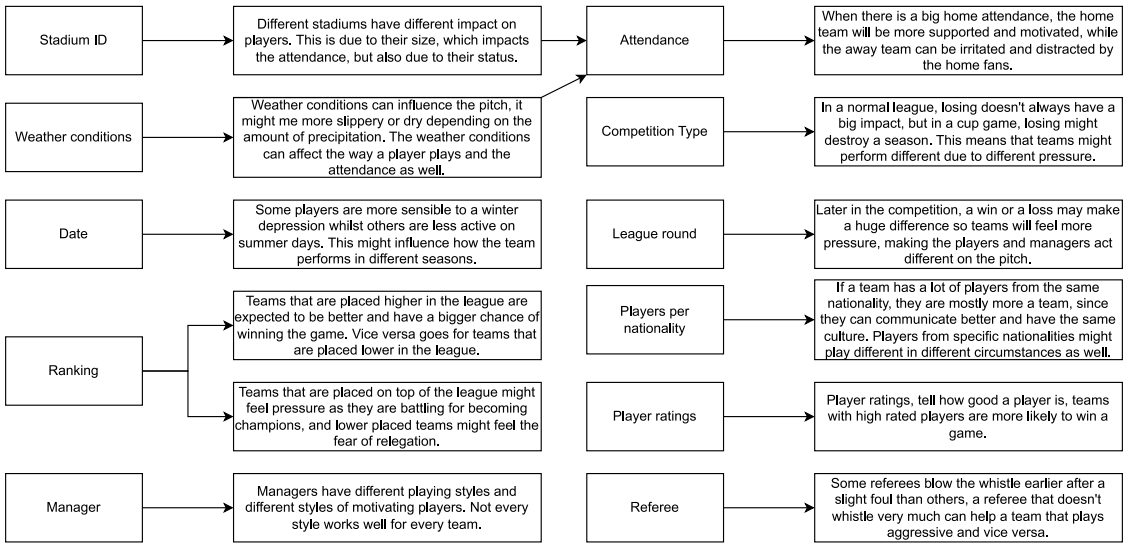

Figure 3: The possible effects of some features on the result of the game

Feature name	**Description**
FHVEC	Vector mean windspeed
FHX	Highest hourly mean windspeed
FXX	Fastest wind gust
TG	Daily mean temperature
TN	Lowest temperature
T10N	Lowest temperature 10 cm above ground level
SQ	Sunhours
Q	Global radiation
DR	Duration of precipation
RH	Amount of precipation in 0.1 mm
RHX	Maximum hourly amount of precipation in 0.1 mm
PX	Highest hourly sea level pressure
PN	Lowest hourly sea level pressure
VVN	Maximum visibility per 100 m
VVX	Minimum visibility per 100 m
NG	Mean daily cloud cover
UG	Mean daily humidity
UX	Maximum humidity
DDVEC	Winddirection in degrees
FG	Daily mean windspeed
FHN	Lowest hourly windspeed
TX	Highest temperature
SP	Percentage of highest potential sunshine duration
PG	Daily mean sea level pressure

Table 3: The used features about the weather and their description [KNM]

4 Research Method

4.1 Classification

To predict the winner of a football match, we make us of classification. Classification can be divided in three parts: Binary classification, Multi-class classification and multi-label classification. Multi-label classification assigns on or multiple lables to a problem. For example, a picture of someone can get labels like, brown hair, blue eyes, white shirt etc. Multi-label classification and binary classification both assign a problem to one class. Binary classification only choses between 2 classes. This is used to check whether an e-mail is spam or not. Multi-class classification can pick multiple classes like in this studie. A match can be assigned a win, a draw and a loss [Kei23].
There are multiple classification algorithms. The most popular algorithms will be described below [Gon22]

- Decision Trees: Decision trees can be seen as a lot of if-statements that split the input in subsets, based on the most important features.
- Logistic regression: Logistic regression estimates the probability of a label. The input is assigned to a class based on this probability.
- Support Vector Classifier: this is an algorithm that creates groups within borders. A new examples will be placed within a group to get a label assigned.
- K-nearest neighbours (KNN): KNN creates a space with all data that it has learned. When a new input is given. It places the input in the space again and assigns the class of the closest neighbours.
- Naive Bayes: Naive Bayes assumes that every feature is independent of each other. It predicts the probability of a class based on every feature and then assigns a class based on those probabilities
- Ensembles: this type of classifiers is a combination of other classifiers, like the classifiers above. Random forests, for example combines multiple decision trees into an other classifier. Random Forest classifier combines multiple decision trees into a more powerful classifier. All decision trees assign a class to the problem, the class that is assigned most of the time will be the class that is assigned by Random forest. All trees in the random forest need to diversify such that wrong predictions of some trees are covered by the other trees. Therefore, not all trees are trained on the same training set and make use of a random sample of features [Yiu].

The dataset will be split in a test set and a training set. After the model has been fit to the training set, the model will be validated. For this study, validation will be done with K-fold Cross Validation, with k=10. Cross validation splits the data into a train and test set where the size of the test set is $\frac{1}{k}$th of the dataset size. The model is fitted k times such that every bit of data is one time in the test set and k-1 times in the training set. This way, the model gets tested to new data every fold and the chance that a model performs good or bad by accident gets reduced, since the score of the model can be seen for every fold [Bro20].

4.2 Feature Testing

To create the best possible model, it is important to test which features have influence on the predictions. Features that have a positive influence on the model should be used to make predictions and features that affect the predictions negatively should be removed. A way to check the influence of a single feature is to compary the accuracy (classification) or the RMSE (regression) with and without the use of the specific feature. Once the removal of a feature improved the accuracy of the model, the feature will be removed for further analysis. If the removal of the feature weakened the model, the feature will not be removed. The predictions in this subsection are made with the use of Gradient Boosting Classifier and a random state of 0.

4.3 Testing Models

The variable `toto_score` can get three values, namely 'w', 'd' and 'l'. W stands for a win of the home team, d stands for a draw and l stands for a lose of the home team, so a win for the away team. Since there are multiple classes and a match can only get one class, we have decided to go for multiclass classification. To find the best classifier, the accuracy of multiple models has been compared.

4.4 Influence of Weather data

To compare whether we can say that using data about the weather conditions influence the results of our predictions, the algorithm is run on both data with and without weather. These results have been analysed using McNemar's test [Rif06] to compare whether the difference is big enough to say that it was influenced by the weather condition. Both results will be compared with a random guess classifier and a majority class classifier. A random guess classifier assigns a class randomly. A majority class classifier always assigns the class that occurs the most.
McNemar's test makes use of the contingency table shown in Table 4 and makes use of Formula 1.

		Model 1	
		Right	Wrong
Model 2	Right	a	b
	Wrong	c	d

Table 4: A contingency table that is used with a McNemar test

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{1}$$

The following hypotheses are made when doing McNemar's test:
H_0: Both models are the same
H_a: There is a difference between the models

If the p-value, retrieved by χ^2, is lower than the pre-determined α level, then the null hypothesis can be rejected.

4.5 Results

The aim of this research was to create a model that can predict the winner of football games. We wanted to know if using weather data improves the quality of our predictions. After testing different models and picking the best one, the model was run on both the data set with the weather data and the data set without the weather data. The results of running both data sets were compared with a chi-squared test, to test if there is a relation between predicting with and without weather data.
The Quality of the model is evaluated based on the accuracy of the predictions. The formula of the accuracy can be found in Equation 2. The accuracy is usually given as a proportion but can be given as a percentage when multiplied by 100. The results of the performed tests are presented in this section.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{2}$$

5 Results

5.1 Feature Testing

In Table 5, all features that have been tested are listed, together with the accuracy after that feature has been removed. Note that this is just the accuracy after the single feature has been dropped and that this does not mean that dropping all those features results in the best accuracy.

Removed feature	Accuracy	Removed feature	Accuracy	Removed feature	Accuracy
None	0.573	referee*	0.576	PN	0.581
competition_id	0.573	stadium_id	0.579	VVN	0.565
competition_id*	0.573	stadium_id*	0.584	VVX	0.581
competition_type	0.573	FHVEC	0.575	NG	0.586
competition_type*	0.573	FHX	0.571	UG	0.574
season	0.573	FXX	0.573	UX	0.576
round	0.573	TG	0.563	DDVEC	0.579
round*	0.573	TN	0.571	FG	0.582
date	0.571	T10N	0.570	FHN	0.584
home_club_position	0.563	SQ	0.571	TX	0.571
away_club_position	0.571	Q	0.576	SP	0.581
home_club_manager	0.584	DR	0.570	PG	0.573
away_club_manager	0.568	RH	0.573	Player ratings	0.576
attendance	0.578	RHX	0.573	Team ratings	0.571
referee	0.576	PX	0.573	Player countries	0.574

Table 5: Table with the accuracy after removing the single feature.
Feature is not removed but changed to dummies*

After testing combinations of removing features that are listed in Table 5, removing the following combination of features resulted in the highest accuracy:

- `DDVEC`
- `TX`
- `SP`
- `PG`
- `Player rating`
- `Team rating`
- `Salary`
- `Weight`
- `Length`
- `Market value`
- `Contract time`

After removing the features listed above and changing the features `competition_type`, `referee` and `stadium_id` to dummies, the accuracy was 0.588.

5.2 Testing Models

After testing multiple classification methods, which are shown in Table 6, RandomForestClassifier gave the best results.

Model	Parameters	Accuracy
KNeighborsClassifier	n_neighbors=2000	0.470
DecisionTreeClassifier	max_depth=5, min_samples_leaf=3	0.563
RandomForestClassifier	n_estimators=130,max_depth=14, random_state=0	0.596
AdaBoostClassifier	n_estimators=56	0.573
ExtraTreesClassifier	n_estimators=115, random_state=0	0.573
BaggingClassifier	n_estimators=115, random_state=0	0.555
GradientBoostingClassifier	n_estimators=300, max_depth=3, random_state=0	0.589

Table 6: Accuracy for some tested classifiers

Random Forest belongs to the ensemble classifiers. As discussed in Section 4.1, an ensemble is a combination of multiple classifiers. That section contains a description of Random forest classifier as well. After testing multiple values for hyperparameters, the following hyperparameters gave the best accuracy:
`n_estimators`=130, `max_depth`=15, `random_state`=0

5.3 Predictions with weather data

The data that has been processed as described in Section 3.2 has been analysed with multiple classification methods. Random Forest Classifier gave us the best result, as mentioned earlier. The predictions on the date with weather data had an accuracy of 0.6019. This means that 60.19% of the predictions are correct.

5.4 Predictions without weather data

For this test, the same steps has been done as for the section before except for merging the weather data with the `games.csv` file. That step has been skipped. Random Forest was used for classifying this problem as well. This gave us an accuracy of 0.5586, so 55.86% of the predictions had the right outcome.

5.5 Model comparison

To compare whether using weather data makes a difference, a McNemar test is performed. For this tests, we use an α level of 0.05. The McNemar test is described in Section 4.4.

5.5.1 Weather vs Random class classifier

A random guess assigns a class randomly. Since there are three classes, the accuracy of a random guess is 0.33 or 33%. Comparing the predicitions with weather information and a random guess gives the following contingency table shown in Table 7.

		Random Guess	
		Right	Wrong
With weather data	Right	119	256
	Wrong	87	161

Table 7: A contingency table to compare Predictions with random guess

After using Equation 1, we got a p-value of 1.18×10^{-19}, which is smaller than 0.05. This means that the null hypothesis can be rejected

5.5.2 Weather vs Majority class classifier

The class that occurs the most is the "w" class. This class is correct in 47% of the cases in the test set, so that is the accuracy of the majority class classifier. See Table 8 for the contingency table.

		Majority class classifier	
		Right	Wrong
With weather data	Right	255	120
	Wrong	38	210

Table 8: A contingency table to compare Predictions with Majority class classifier

This test gave us a p-value of 1.1×10^{-10}, which is smaller than 0.05. The null hypothesis can be rejected due to this p-value

5.5.3 No weather vs Random class classifier

Table 9 contains the contingency table of the comparison between the predictions without weather data and a random guess. We got a p-value smaller than 0.05, namely 9.45×10^{-14}. Therefore, the

		Random class classifier	
		Right	Wrong
Without weather data	Right	116	232
	Wrong	96	179

Table 9: A contingency table to compare predictions without weather and random class classifier

null hypothesis can be rejected.

5.5.4 No weather vs Majority class classifier

The predictions that did not use weather data have been compared with a majority class classifier as well in Table 10

		Majority class classifier	
		Right	Wrong
Without weather data	Right	247	101
	Wrong	46	229

Table 10: A contingency table to compare predictions without weather and majority class classifier

This test gave us a p-value of 8.43×10^{-6}. That p-value is smaller than 0.05 as well, which means that the null hypothesis can be rejected

5.5.5 No weather vs Weather

Table 11 shows the contingency table for the comparison of the dataset with and without weather.

		Without weather data	
		Right	Wrong
With weather data	Right	335	40
	Wrong	13	235

Table 11: A contingency table to compare predictions with weather and without weather

Comparing both ways of predicting the winner of a match gave us a p-value of 3.6×10^{-4}. This is smaller than 0.05, so the null hypothesis can be rejected.

6 Conclusions and Further Research

6.1 Conclusion

There are plenty ways of getting football and weather related data from internet. A lot of websites keep track of past matches. For this research, data from Transfermarket.com has been used because it contains data from around the past 10 seasons and from multiple leagues. Player ratings have been retrieved from FIFA 23 since that game has a good status on rating the quality of players. Weather data has been downloaded directly from the KNMI. They have multiple weather stations around the Netherlands and very thrustable data.
Our model was made with different features based on the game, the weather and the players that have participated in the game. We have chosen to classify the result of the game in three classes; a home team win, a draw and a home team loss. The classification was done with Random Forest Classifier.

The model was run on both data with the weather conditions and data without the weather conditions. The analysis with weather conditions got us an accuracy of 0.6019. The data without weather conditions got us an accuracy of 0.5586. That means that we have the highest accuracy when we take the weather conditions in account. Both predictions have been compared with each other, a random guess and a majority classifier by a McNemar test. We can reject H_0 if $p < 0.05$. Table 12 shows the p-value for each test.

Comparison	**P-value**
With weather vs Random	1.18×10^{-19}
With weather vs Majority classifier	1.1×10^{-10}
Without weather vs Random	9.45×10^{-14}
Without weather vs Majority classifier	8.43×10^{-6}
With weather vs Without weather	3.6×10^{-4}

Table 12: The different McNemer tests and their results

As you can see, all p-values are lower than 0.05. This means that both predictions are significantly better than both a random guesser and a majority classifier. Making use of weather data also give significantly better predictions than predictions without weather data. Therefore, we can conclude that making use of weather data affects the quality of predicting football matches in the Dutch Eredivisie positively.

6.2 Further Research

This thesis mainly focuses on the Dutch Eredivisie. For future research, other competitions could be analyzed. Making an analysis for other sports could also be a future research topic. Weather influence on indoor and outdoor sports can be analyzed as well. We have predicted which team will win the match. In further research, predicting the score of the game is an option.

References

[AT522] AT5. Hittemaatregelen in de arena: dak deels dicht en flesje water toegestaan. https://www.at5.nl/artikelen/216390/hittemaatregelen-in-de-arena-dak-deels-dicht-en-flesje-water-toegestaan, 2022.

[BGDO+09] Mathieu Bélanger, Katherine Gray-Donald, Jennifer O'loughlin, Gilles Paradis, and James Hanley. Influence of weather conditions and season on physical activity in adolescents. https://www.sciencedirect.com/science/article/pii/S1047279708003670, 2009.

[BGFM15] Franck Brocherie, Olivier Girard, Abdulaziz Farooq, and Grégoire P. Millet. Influence of weather, rank, and home advantage on football outcomes in the gulf region. https://journals.lww.com/acsm-msse/Fulltext/2015/02000/Influence_of_Weather,_Rank,_and_Home_Advantage_on.23.aspx, 2015.

[Bra15] Lotte Bransen. Doelpuntenfestijn bij zonneschijn. https://www.tussendelinies.nl/doelpuntenfestijn-bij-zonneschijn/, 2015.

[Bro20] Jason Brownlee. A gentle introduction to k-fold cross-validation, Aug 2020.

[CLA+21] P. Chmura, H. Liu, M. Andrzejewski, J. Chmura, E. Kowalczuk, A. Rokita, and M. Konefał. Is there meaningful influence from situational and environmental factors on the physical and technical activity of elite football players? evidence from the data of 5 consecutive seasons of the german bundesliga. https://pubmed.ncbi.nlm.nih.gov/33690609/, 2021.

[Dat20] Datacamp. Handling categorical data in python tutorial. https://www.datacamp.com/tutorial/categorical-data, 2020.

[FMT19] Gabriel Fialho, Aline Manhaes, and Joao Paulo Teixeira. Predicting sports results with artificial intelligence – a proposal framework for soccer games. https://www-sciencedirect-com.ezproxy.leidenuniv.nl/science/article/pii/S1877050919322033, 2019.

[Gon22] Destin Gong. Top 6 machine learning algorithms for classification. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501, Jul 2022.

[Hes19] David Hessing. Hoe de modder een goal van Jurrie Koolhof voorkwam. https://www.ad.nl/sport/hoe-de-modder-een-goal-van-jurrie-koolhof-voorkwam~aee179e1/, 2019.

[IRPTO20] Ditsuhi Iskandaryan, Francisco Ramos, Denny Palinggi, and Sergi Trilles Oliver. The effect of weather in soccer results: An approach using machine learning techniques. *Applied Sciences*, 10:6750, 09 2020.

[JFN06] A. Joseph, N.E. Fenton, and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[Kei23] Zoumana Keita. Classification in machine learning: An introduction. https://www.datacamp.com/blog/classification-machine-learning, 2023.

[KG21] Jan Kozak and Szymon Głowania. Heterogeneous ensembles of classifiers in predicting bundesliga football results. *Procedia Computer Science*, 192:1573–1582, 2021. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

[KNM] KNMI. Dagwaarden van weerstations. https://www.daggegevens.knmi.nl/klimatologie/daggegevens.

[KTS17] George Kyriakides, Kyriacos Talattinis, and George Stephanides. A hybrid approach to predicting sports results and an accurate rating system. *International Journal of Applied and Computational Mathematics*, 3(1):239–254, Mar 2017.

[LJ14] Carson K. Leung and Kyle W. Joseph. Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, 35:710–719, 2014. Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.

[LPLB11] Carlos Lago-Peñas and Joaquin Lago-Ballesteros. Game location and team quality effects on performance profiles in professional soccer. *J Sports Sci Med*, 10(3):465–471, September 2011.

[Rif06] Robert H Riffenburgh. Tests on categorical data. In *Statistics in Medicine*. Elsevier, 2 edition, 2006.

[SD10] MP Schwellnus and EW Derman. Jet lag and environmental conditions that may influence exercise performance during the 2010 FIFA World Cup in South Africa. *South African Family Practice*, 52(3):198–205, 2010.

[Ten22] Babu Tendu. Revealed! top 15 most popular sports in the world in 2022. https://sportsbrief.com/facts/top-listicles/16715-revealed-top-15-popular-sports-world-2022/, 2022.

[Tho77] John Thornes. The effect of weather on sport. *Weather*, 32:258–268, 07 1977.

[VTB19] VTBL. Airco's in wk-stadions blazen zo koele lucht op het veld. https://www.rtlnieuws.nl/sport/voetbal/artikel/4863251/aircos-wk-stadions-blazen-zo-koele-lucht-op-het-veld, 2019.

[Wil11] S. Willis. User's manual: Cooling break. https://www.ligue1.com/Articles/NEWS/2021/07/24/user-s-manual-cooling-break, 2011.

[Yiu] Tony Yiu. Understanding Random Forest — towardsdatascience.com. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.