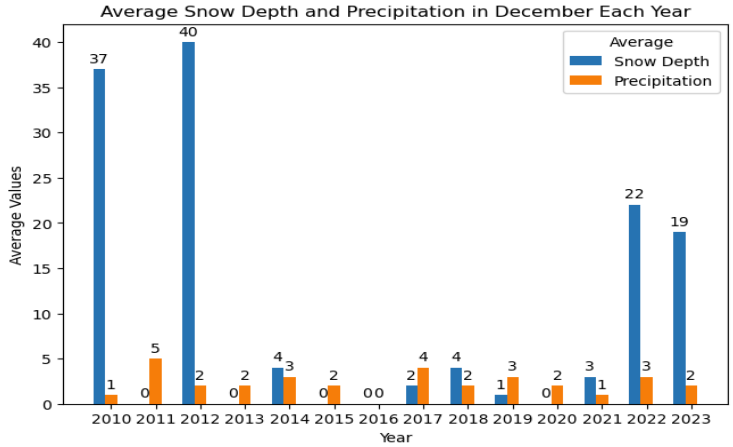
Predicting if there's snow on Christmas 2024

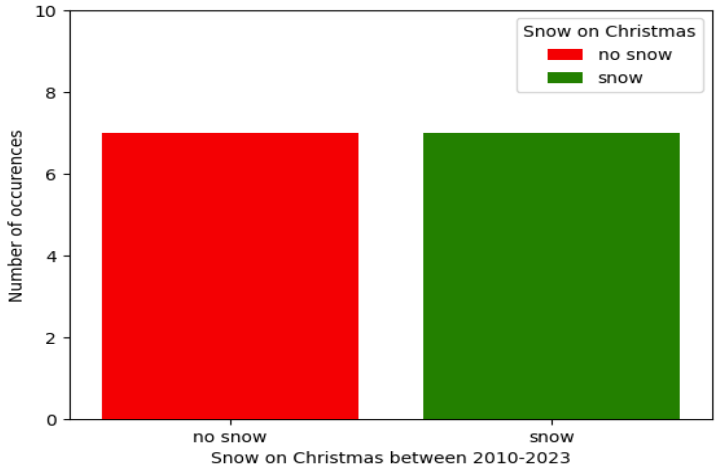
I) Introduction

Predicting whether there is snow on Christmas is a classical question in Finland at the beginning of December each year. This is of course not an easy task to predict considering it depends which part of Finland one lives in as well as changes in the weather patterns around the globe and the more recent impact of global warming. In this project, I decided to use machine learning to try and predict the outcome of this years result in Vantaa.

II) Problem Formulation

Description of the problem: Using the data gathered from Finnish Meteorological Institute[1] and more precisely Helsinki-Vantaa airport observation station from January 2010 to December 2023.
The metrics chosen for the dataset are daily observations of Average temperature, Precipitation amounts, Min temperature, and Snow depth.
Using these datapoints the goal of this project is to predict whether the snow depth on 24th December is above 1 to indicate there is snow on the ground, or 0 or -1 to indicate there's no snow on the ground.
The dataset was chosen to span 14 years of recent data throughout the year in order to find the weather patterns on different seasons and to have long enough period make more reliable prediction.
Had the dataset been longer, i.e. from the 1990, it might not notice the recent effects of global warming and result could be less reliable.
On the other hand, as the goal of this project is to predict whether there is snow on 24th December, As can be seen in the charts below, had I only used the precipitation and snow depth data from previous Decembers, the prediction would be less reliable as there are probably too little datapoints.





III) Methods

Preprocessing

The dataset consists of 5112 data points, which consists of the daily observations of Average temperature, Precipitation amounts, Min temperature, and Snow depth as well as datetime information. Existing datetime information needs to be partially replaced by Day_sin and Day_cos features to signify the cyclical nature of the years for the model. In addition to these "binary snow" feature is needed to further clear up, which days have snow or not. After these new features, the unnecessary ones are cleared out, so that only relevant information for the model remains.
The empty data points in the dataset are removed as there was only a few of them, so it doesn't affect the outcome.
Also one last thing the dataset needs to be normalized to even out the relative scale between the features and prevent bias in the model training. This is necessary, since there are values such as temperature that varies between (roughly) -25 and 30, and another like snow depth which can from -1(no snow) to even over 100.

Models

Logistic Regression

As this is a classification problem of "yes" or "no", one good choice for the ML model is Logistic regression. Logistic regression predicts the outcome as 1 as "yes" and 0 as "no", so in this case 1 would be "there is snow" and 0 "no snow" on a given day.

MLP Classifier

Multi-layer perceptron classifier is a supervised learning model that utilizes neural network for classification tasks. It uses rectified linear unit (ReLU) activation function in it's hidden layers to learn complex patterns. It is trained

by using Backpropagation and it optimizes the logistic loss between probabilities and true labels.

Loss function
In this kind of classification problem where logistic regression is used for predicting whether there is snow on a given day, the logistic loss is acting as a penalty for the model in order to optimize it's algorithm and learn more precise prediction. Both models in this case are utilizing logistic loss in their learning.

Validation
Dataset is split into training set that covers 70 % of the dataset, validation set is 15% and the remaining 15% is used for the testing set. This is to give big enough dataset to train the model as accurately as possible and also leaving enough data for validation and testing sets.
Accuracy score and confusion matrices are used to see how accurately the model is performing after training.

Results

Both models were trained and tested for their accuracy. Looking at their accuracy scores[8] and confusion matrices [9,10], the logistic regression seems to be performing better as it's accuracy for both test and validation sets are slightly higher and it produces more consistent number of true positives in both sets.
Therefore my choice for the actual prediction is logistic regression.
Test error for this model is 2,48%.

Conclusion
The biggest challenge in making the actual prediction for upcoming Christmas is that there is no data available yet for 24th December this year. Therefore I used Linear Regression to forecast the weather data for this years Christmas eve and use that as the data for predicting if there is snow or not. Whether that prediction is actually correct will remain a mystery still for a couple of months, but based on the models accuracy and the vast amount of data it was trained with, the model should perform quite well. So, if the this December's data proof to be close to what was estimated with LR, the prediction would be quite accurate.
In the end, the problem could have been formulated a bit differently and focus on training the linear model to predict possible weather conditions for Christmas and base the outcome on that or use logistic regression to predict the result based on the dataset provided by the linear regression model.

IV) References

[1] Weather dataset source
https://en.ilmatieteenlaitos.fi/download-observations

[2]Machine learning book, A. Jung.
https://github.com/alexjungaalto/MachineLearningTheBasics/blob/master/MLBasicsBook.pdf

[3]Pandas library for dataframe manipulation
https://pandas.pydata.org/docs/reference/frame.html

[4]Bar Chart visualizations:
https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html#matplotlib.pyplot.bar

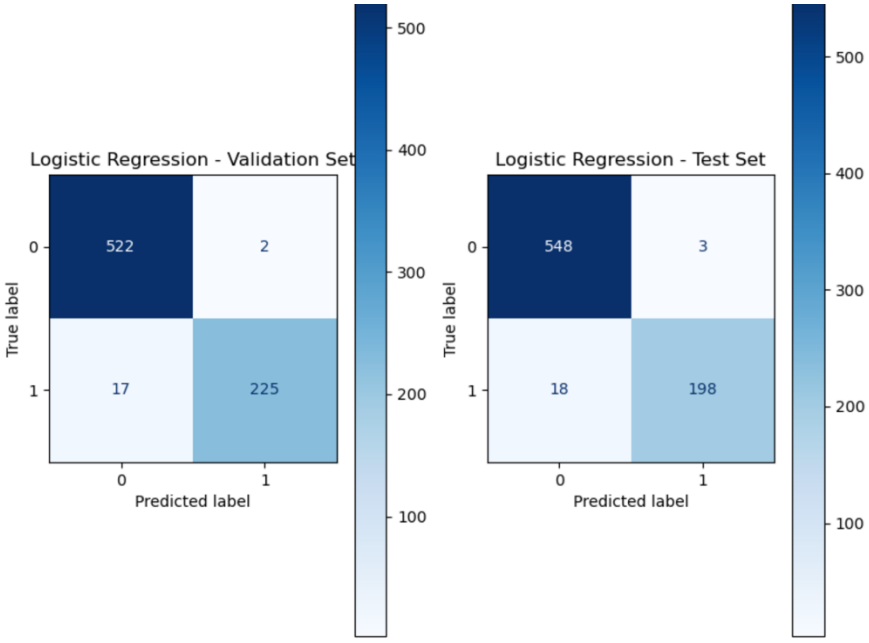
[5]Help with bar chart visualizations:
https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_colors.html

[6]StandardScaler in Scikitlearn
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler

[7]Train_test_split in Scikitlearn
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html#sklearn.model_selection.train_test_split

[8]Code and dataset:
https://github.com/PRoutamaa/ML_course_project2024

[9]Logistic regression confusion matrix



[10]MLPClassifier confusion matrix