**ML Project - Midterm Report Draft**

**Cheating Charon: Predicting Excess Mortality Due to Heat Waves in Vulnerable  
Groups in Germany**

**Abstract:**

*Predicting the consequences of our changing climate for public health is a crucial area where machine learning can assist policy makers. As the effects of man-made climate change are felt, excess deaths due to extreme temperatures and weather events are expected to rise.*

*For our project, we have chosen to test and compare two or three predictive machine learning models to anticipate the number of excess deaths in Germany due to heat waves within the next decade under two distinct emissions scenarios.*

*Within these two scenarios, we have chosen to focus on two groups within German society. Germany’s population, like many in Europe, is ageing rapidly and elderly people (those aged sixty-five and above) are particularly vulnerable to the health effects of heat waves. Additionally, young children under the age of five are also at risk. Consequently, we have chosen these two age demographics as our focus.*

**Keywords:** Heat-related mortality Excess deaths Climate change Projection Adaptation

**Proposed Method:**

*Since the fundamental goal of our project is to predict how climate change, in this case defined as increasing temperatures and heat waves, can affect the mortality rate in vulnerable populations within Germany, there are several important variables to consider for our possible machine learning models. In our view, the following list covers the basic necessary data for those models:*

* *Population growth rate*
* *Actual and predicted population under 5 years old*
* *Actual and predicted population 65 years and older*
* *Mortality rate and main causes of death for German population under 5 years old*
* *Mortality rate and main causes of death for German population 65 years or older*
* *Temperature forecasting for Germany for Germany in line with the IPCC high-emissions scenario Representative Concentration Pathway (RCP) also called RCP8.5*
* *Temperature forecasting for Germany in line with low-emissions scenario, also called RCP2.6*

*The project will consider two distinct, yet credible climate scenarios in which the degree to which temperatures increase will vary in its effect on the mortality rate of our target populations. These two scenarios are determined by net CO2 emissions and follow the forecast predictions in the Intergovernmental Panel on Climate Change (IPCC)’s Representative Concentration Pathways (RCP).*

*The pathways describe different climate futures, all of which are considered possible depending on the volume of greenhouse gases (GHG) emitted in the years to come. The two scenarios we have chosen are as follows:*

* *A low-emissions scenario, also called RCP 2.6*
* *A high-emissions scenario, also called RCP 8.5*

*Although there are other RCP trajectories, we have chosen RCP 2.6 as a realistic best-case scenario and RCP 8.5 as an absolute worst-case scenario for the climate. The considerable difference between these two foreseeable scenarios will allow for greater variation in the prediction variable in our models. Furthermore, the contrast between the two outcomes will – we hope – make the consequences of doing nothing to reduce CO2 emissions and mitigate climate change more strikingly obvious.*

*From both a technical and a non-technical standpoint, our project will be a success if we can:*

* *develop machine learning models that can predict the excess deaths in our target groups in line with a baseline projected by the WHO,*
* *we can visually represent the differences in terms of mortality between the two chosen climate scenarios,*
* *if we can compare & contrast the efficacy and interpretability of our chosen ML models*
* *reach high accuracy and precision scores (e.g., ranging from 0.6 to 0.8 in the relevant metric) in our chosen models.*

*Owing to the clearly determined pathways in either scenario, we will not need to predict temperature increases, but rather attempt to predict the outcomes of the temperature increases in terms of mortality rates. As we are considering the effects of two scenarios on two target groups using likely three different models, one key successful outcome of the project will be twelve comparable predictions.*

*We are currently considering using at least two particular machine learning models, namely logistic regression and random forest classifier. If at all possible, we would also like to use a time series model, but need first to investigate our data further before choosing a particular time series model. The models will be trained using the mortality rate data, demographic data, and the mean temperature predictions.*

*The key evaluation metrics for our logistic regression model will be Explained Variance, Mean Squared Error, and the R2 coefficient, whilst the metrics for the random forest classifier will very likely be Prediction, Accuracy, and Recall. However, these metrics may change as we become more accustomed with the requirements of the project.*

*We anticipate that the data processing requirements will very likely vary from model to model, increasing the overall time we spend on this phase of the workflow. However, as our pedagogical goal is to learn as much as possible about the differences between the various models, we believe that this effort will be worthwhile.*

*After the data wrangling process, we will separate all our data into training and testing data, using a 70\% and 30\% split, leaving aside the testing data and working solely with the training data. Our second step in the process will be feature engineering, such as scaling the data to normalize the range of independent variables and performing the necessary transformations.*

*We will also tune the parameters for each specific model referencing their particular guidelines to achieve the highest accuracy and precision scores: solver, penalty, C (regularization strength) for logistic regression, and min\\_samples\\_leaf, n\\_iter, cv for random forest classifier. Ideally, we aim to achieve a score ranging from 0.6 to 0.8 in the key accuracy metrics.*

*Once the parameter tuning testing is done, the remaining 30\% of the testing data will be introduced to the model so it can predict the mortality rate and temperature. This process will be repeated for both emissions scenarios. Our fourth step will be to import GridSearchCV to do cross-validation of the data.*

**Experiments:**

**Data:**

Temperature and meteorological data were obtained from NASA. Mortality data and population projections for Germany were obtained from the Federal Statistical Office of Germany. We also obtained datasets on the causes of death amongst our age groups from the Federal Statistical Office of Germany. This dataset was used to derive a ratio of the citizen who died due to environmental exposure in our historical data to the German population as a whole. The method uses a series of regression equations that quantify the current and historical relationships between mortality and a set of independent variables.

The datasets we used for our project so far include:

* Actual and predicted population of Germany
* Mortality rate and main causes of death for German population
* Temperature forecasting for Germany for Germany in line with the IPCC high-emissions scenario Representative Concentration Pathway (RCP) also called RCP8.5
* Temperature forecasting for Germany in line with low-emissions scenario, also called RCP2.5

We have also had to account for a class imbalance in our original datasets. Originally, we intended to analyze the effects of climate change on young children (under the age of five) and old people (over the age of sixty-five). However, in our data wrangling, we found that there more observations for older people, particularly in the population projection datasets, than young children. To address this imbalance, we sliced our populations into four distinct age groups that cover the entire German population, both actual and predicted. The groups are as follows:

* 0 – 25
* 26 – 50
* 51 – 75
* 76 – 100

Karon\_clean: 2003 – 2021

Karon\_ data: 2022 – 2050

**Datasets:** temperature, mortality, projections, population

* no ratio for the death

*For this project, we will draw on historical research into the effects of heat waves, including the heat wave in Paris in 2003\cite{poumadere2003HeatWave2005}. Additionally, we will use existing research models for a general methodology and baseline for evaluating our approach. We will use data from the Federal Statistical Office of Germany, World Health Organization (WHO), European mortality database, and Eurostat.*

*We will use an Agile model to manage our project work, assigning different tasks according to our respective skill sets and interests, and each team member will be responsible for a particular data collection relevant to the overall project.*

**Evaluation method:**

**Experimental details:**

Linear model

**Results:**

* Comment on quantitative results

**3. Future Work**

* Got a baseline model, now need to tweak it
* Can we identify hyperparameters that we want to tune?
* How do we intend to improve the fit, precision, and robustness of the proposed model have been improved from the previous projection

**4. References**

* **Honda et al,** Heat-related mortality risk model for climate change impact projection

\documentclass[10pt,twocolumn,letterpaper]{article}

\usepackage{statcourse}

\usepackage{times}

\usepackage{epsfig}

\usepackage{graphicx}

\usepackage{amsmath}

\usepackage{amssymb}

% Include other packages here, before hyperref.

% If you comment hyperref and then uncomment it, you should delete

% egpaper.aux before re-running latex. (Or just hit 'q' on the first latex

% run, let it finish, and you should be clear).

\usepackage[breaklinks=true,bookmarks=false]{hyperref}

\statcoursefinalcopy

\setcounter{page}{1}

\begin{document}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% DO NOT EDIT ANYTHING ABOVE THIS LINE

% EXCEPT IF YOU LIKE TO USE ADDITIONAL PACKAGES

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%% TITLE

\title{\LaTeX\ Template for ML Midterm Project Report}

\author{First Author\\

{\tt\small firstauthor@hertie-school.org}

\and

Second Author\\

{\tt\small secondauthor@hertie-school.org}

\and

Third Author\\

{\tt\small thirdauthor@hertie-school.org}

}

\maketitle

%\thispagestyle{empty}

% MAIN ARTICLE GOES BELOW

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%% ABSTRACT

\begin{abstract}

An abstract should concisely (200-250 words) motivate the problem, describe your aims, describe your contribution, and highlight your main finding(s). Given that your project is still a work-in-progress, it's OK if `your contribution' and `your findings' are things you're still working on.

\end{abstract}

%%%%%%%%% BODY TEXT

\begin{itemize}

{\color{blue}

\item Remember that you should \textbf{submit the report} via Moodle and \textbf{include in the report the link to accessible GitHub repository that contains the code}. Also, \textbf{only one member per team} needs to submit the project material. You must include a link to your GitHub repository for the project as the first footnote on the first page. \footnote{Here's a link to my GitHub account \url{https://github.com/sjankin}, Lynn's \url{https://github.com/LynnKaack} and Eric's \url{https://github.com/EricKolibacz}. Make sure that your repository is accessible to us!}

\item The midterm project report should be {\bf 4 pages long (not counting references), and a maximum 10 references}. The report should contain the sections that are already provided in this paper. It forms the basis of the final report with the same structure. Please check out the text in these sections for further information.

\item Your midterm milestone will be graded on the following criteria:

\begin{itemize}

\item Progress: Has the team made good progress on the project? You should have done approximately half of the work of your project.

\item As a minimum, your milestone should show that you have setup your data, baseline model code, and evaluation metric, and run experiments to obtain some results (assuming you are doing a typical model-building project). Other than this, `good progress' depends on various factors (e.g., whether your model is implemented from scratch or based on an existing codebase).

\item Understanding: Does the milestone show a strong understanding of its problem, tasks, methods, metrics, and research context?

\item Writing quality: Does the milestone clearly communicate what you've done and why, providing the requested information, to an appropriate level of detail (given the page limit)?

\item You will receive some brief feedback on your milestone. Feedback may contain helpful suggestions for your project (e.g., try a particular method, read a particular paper) and/or warnings about your project plan (e.g., if your plans are too ambitious or not ambitious enough), and how you could improve your technical writing (e.g., adjustments to clarity, level of detail, formatting, use of references).

\end{itemize}

\item Technical writing is an important skill in this class, in research, and beyond. It's well worth spending time developing your ability to communicate technical concepts clearly. Here are some resources which might help you improve your technical writing:

\begin{itemize}

\item Tips for Writing Technical Papers, Jennifer Widom (\url{https://cs.stanford.edu/people/widom/paper-writing.html}).

\item Write the Paper First, Jason Eisner (\url{https://www.cs.jhu.edu/~jason/advice/write-the-paper-first.html}).

\end{itemize}

\item Here are some other things you can do to improve your technical writing:

\begin{itemize}

\item Look carefully at several ML papers to understand their typical structure, writing style, and the usual content of the different sections. Model your writing on these examples.

\item Think about the ML papers you've read (for example, the one you summarised for your proposal). Which parts did you find easy to understand and why? Which parts did you find difficult to understand and why? Can you identify any good writing practices that you could use in your technical writing?

\item Ask a friend to read through your writing and tell you if is clear. This can be useful even if the friend does not have the relevant technical knowledge.

\end{itemize}

}

\end{itemize}

\section{Proposed Method}

This section details your approach(es) to the problem. For example, this is where you describe the architecture of your model, and any other key methods or algorithms.

You should be specific when describing your main approaches - you may want to include equations and figures (though it's fine if you want to defer creating time-consuming figures until you write your final report).

This is an example of a mathematical equation:

$$f(\mathbf{x}; \mathbf{w}) = \sum\_{i=1}^{n} w\_ix\_i.$$

This is a mathematical expression, $h(\mathbf{x}) = \hat{y}$ formatted in text.

You should also describe your baseline(s). Depending on space constraints, and how standard your baseline is, you might do this in detail, or simply refer the reader to some other paper for the details.

If any part of your approach is original, make it clear (so we can give you credit!). For models and techniques that aren't yours, provide references.

If you're using any code that you didn't write yourself, make it clear and provide a reference or link. When describing something you coded yourself, make it clear (so we can give you credit!).

\section{Experiments}

This section contains the following.

\paragraph{Data:} Describe the dataset(s) you are using (provide references). If it's not already clear, make sure the associated task is clearly described.

\paragraph{Evaluation method:} Describe the evaluation metric(s) you used, plus any other details necessary to understand your evaluation.

\paragraph{Experimental details:} How you ran your experiments (e.g. model configurations, learning rate, training time, etc.)

\paragraph{Results:} Report the quantitative results that you have found so far. Use a table or plot to compare multiple results and compare against baselines.

Table \ref{tab:some\_table} shows an example for formatting a table.

\begin{table}

\centering

\begin{tabular}{|l|c|}

\hline

Method & Accuracy \\

\hline\hline

Method 1 & $70 \pm 3$ \% \\

Method 2 & $76 \pm 3$ \% \\

\hline

\end{tabular}

\caption{This is an example of a table.\label{tab:some\_table}}

\end{table}

\paragraph{Comment on your quantitative results.} Are they what you expected? Better than you expected? Worse than you expected? Why do you think that is? What does this tell you about what you should do next? Including training curves might be useful to discuss whether things are training effectively.

You don't need to report any qualitative results (`analysis') in the milestone, though you can if you like.

\section{Future work}

Describe what you plan to do for the rest of the project, and why. You can include stretch goals if you like.

{\small

\bibliographystyle{ieee}

\bibliography{bibliography.bib}

}

\end{document}