Abstract—O trabalho prático foi dividido em duas partes: A primeira, referente à formulação de possíveis questões e respostas que se poderiam formular a partir de um conjunto de dados e, consequentemente, ao pré-processamento desse mesmo conjunto de dados para viabilizar respostas a essas questões. Daqui resultou a formulação do problema objetivo de proceder à correta classificação de um município como industrial ou residencial com base na informação do seu consumo energético, assim como, através da predição utilizando um regressor, prever o consumo energético em Lisboa face às condições meteorológicas. A segunda parte completa o trabalho prático, tendo por objetivo a aplicação e compreensão de métricas de performance sobre os modelos de aprendizagem automática para classificação e regressão, e a influência da aplicação de técnicas de amostragem dos dados sobre os resultados obtidos, visando as questões do problema objetivo.

Neste documento são apresentados os melhores resultados obtidos através de *oversampling* para o classificador e o melhor modelo sem qualquer amostragem para o regressor.

keywords—Megadados, Amostragem, Aprendizagem Automática, Classificação, Regressão

Numa era tecnológica onde predomina a abundância de dados, a utilização de modelos de aprendizagem automática viu um rápido crescimento, permitindo extrair informações que outrora escapariam a abordagens analíticas. Desta forma, tornou-se necessário obter e preparar os dados para fornecê-los a algoritmos de aprendizagem automática, para que estes possam tomar decisões e obter soluções para problemas definidos em requisitos funcionais.

Com vista a compreender a modelação de dados e descobrir problemas que possam ser respondidos pelos mesmos, na primeira fase do trabalho o grupo formulou duas questões:

1) "É possível inferir se um município é predominantemente industrial ou residencial com base na evolução do consumo energético ao longo do dia?"

2) "Conseguimos prever o consumo energético atendendo às condições meteorológicas previstas para uma determinada data?"

Com as duas questões em mente, foram formulados cinco conjuntos de dados através de diferentes métodos de *Feature Selection* e *Feature Reduction* para aplicar a um algoritmo de aprendizagem automática *Random Forest* utilizando diversos tipos de amostragem.

Na aprendizagem automática, a escolha do modelo certo para um determinado conjunto de dados é complexa, uma vez que não depende apenas do tipo de algoritmo, mas também da escolha do modelo que equilibra o desempenho em termos de objetivo final, eficiência computacional e a sua interpretação. A abordagem mais típica na comparação de modelos envolve começar por modelos mais simples, chamados de *baseline models*, e compará-los com modelos mais complexos, procurando sempre manter o modelo mais simples e com o melhor compromisso entre desempenho e eficiência [1].

O desempenho dos modelos pode ser medido com um conjunto de métricas, que permitem monitorizar e avaliar a performance dos modelos durante a fase de treino e de teste. Atendendo que os problemas de *Machine Learning* podem ser reduzidos a classificação ou regressão, existem diferentes métricas associadas a cada um. Para a regressão, uma vez que o output é contínuo, as métricas associadas são calculadas com base na distância entre o valor previsto (predicted) e o considerado verdadeiro (ground truth). Para classificação, compara-se o desempenho num problema discreto, com métricas tipicamente associadas ao número de previsões corretas (accuracy), o rácio entre verdadeiros casos positivos e os classificados (precision), entre outras, a serem mostradas [2].

Na segunda fase do trabalho prático, pretende-se treinar modelos de classificação e regressão que respondam às questões formuladas, utilizando técnicas de amostragem na seleção de instâncias, como *oversampling* e *undersampling*. Serão comparadas criticamente as diferentes métricas de avaliação de performance dos modelos. Além disso, objetiva-se aproveitar a infraestrutura computacional da ferramenta SPARK para minimizar o overhead introduzido durante a fase de treino dos modelos.

Fornecidos dados de consumos energéticos para múltiplos códigos postais, em determinadas horas do dia, bem como a evolução das condições meteorológicas ao longo do dia em Lisboa, o grupo definiu as duas perguntas formuladas anteriormente no capítulo da Introdução.

Na primeira fase do trabalho, foram aplicadas técnicas de redução de dimensionalidade com o objetivo de evitar problemas causados pela *Curse Of Dimensionality,* onde a aprendizagem dos modelos é afetada pelo elevado conjunto de características dos dados e pela sua natureza qualitativa ou quantitativa discreta dos atributos.

1. Classificação de municípios

Do problema de classificação de municípios, foram extraídas, do conjunto de dados original, instâncias para diversos códigos postais, classificados manualmente e, por isso, suscetíveis a erro humano, como sendo predominantemente "Industriais" ou "Residenciais".

Após a redução de dimensionalidade, verificou-se um cenário de classificação pouco complexo, ou seja, o fator determinante para distinguir o tipo de município pode ser analisado quase sem auxílio de algoritmos de aprendizagem automática, comparando diretamente o atributo referente ao consumo energético.

Não obstante à fácil classificação, e com foco em compreender as diversas técnicas de amostragem e comparação de modelos de aprendizagem automática para os diversos conjuntos de dados, desenvolveu-se e compararam-se as diferentes técnicas de amostragem: *oversampling* e *undersampling*, bem como os resultados obtidos do treino de um modelo classificador *Random Forest*.

1. Previsão de consumo energético

Para abordar a previsão do consumo de energia com base nas condições meteorológicas previstas para um determinado dia, foram criados três conjuntos de dados distintos utilizando diferentes técnicas de seleção de características. O objetivo foi treinar e comparar os resultados obtidos na construção de modelos regressores.

Para a geração do modelo de aprendizagem, é necessário determinar o número de instâncias de cada classe nos conjuntos de dados preparados para ambas as questões. Isto permite compreender quais técnicas de amostragem devem ser aplicadas ou quais modelos são adequados para responder às questões colocadas com base nos conjuntos de dados disponíveis.

Na classificação de municípios, o objetivo é resolver um problema de classificação binária, onde o número de instâncias manualmente categorizadas como "Industrial" ou "Residencial" difere no conjunto de treino, numa proporção aproximada de 40/60. Assim, o uso de técnicas de *oversampling* e *undersampling* pode ser aplicado sem prejudicar a performance do algoritmo *Random Forest*.

Em contrapartida, na previsão do consumo energético, considerando o atributo das condições meteorológicas como classe, surge um novo problema, onde algumas condições meteorológicas são muito mais frequentes do que outras. Nesse sentido, uma solução a adotar seria comparar o desempenho entre um algoritmo regressor *Random Forest* e uma abordagem *One-vs-All* (ou *One-vs-Rest*). A segunda opção é geralmente mais indicada para problemas em que o número de instâncias de cada classe é discrepante, permitindo obter melhores métricas de performance baseadas numa medida de erro calculada pela distância.

Para a partição do conjunto de dados, o grupo seguiu a abordagem típica de utilizar 2/3 para o conjunto de treino, que será fornecido ao algoritmo de aprendizagem automática, e 1/3 para validação de resultados e avaliação das métricas de performance.